Delft University of Technology
Faculty Electical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics
together with Erasmus Medical Center Rotterdam

Treatment Prediction in PDAC Patients: A Predictive Model for FOLFIRINOX Chemotherapy Response using Random Forest and Integrative Analysis of Blood and Tumor Markers

Master Thesis

Delft Institute of Applied Mathematics
in coorporation with Erasmus Medical Center Rotterdam
as partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE
in
Applied Mathematics

by

Jenny Lu

Delft, the Netherlands, 2023-08-23

MASTER THESIS APPLIED MATHEMATICS

Treatment Prediction in PDAC Patients: A Predictive Model for FOLFIRINOX Chemotherapy Response using Random Forest and Integrative Analysis of Blood and Tumor Markers

JENNY LU

Delft University of Technology
Erasmus Medical Center Rotterdam

Supervisor (TU Delft)

Prof. Dr.ir. G. Jongbloed

Supervisors (Erasmus MC)

Prof.dr. C.H.J. van Eijck
Drs. G.J. Strijk

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is a devastating disease with a high mortality rate, poor prognosis, and a mere 7.7% 5-year survival rate [1] compared to 65% for all cancer types [2]. Approximately 80% of patients are diagnosed at the advanced stage [3], for which only palliative chemo(radio)therapy remains as treatment option. However, the efficacy of chemotherapy varies among patients, e.g. FOLFIRINOX has a response rate of only 25-30% in metastatic patients and a disease control rate of 70.95% [4]. Therefore, stratifying patients is crucial for individual benefit and for addressing the socioeconomic challenge of rising healthcare costs and increasing cancer incidence. This master thesis aims to investigate the relationships between tumor markers CA19-9 and CEA as well as blood marker data, both before and after one cycle of FOLFIRINOX and their correlation to the chemotherapy response. The goal is to subsequently develop a robust classification model to improve patient stratification and facilitate personalized treatment approaches.

The analyzed cohort comprises 247 PDAC patients of which 55% are male and 45% are female participants. Among them, 152 had (borderline) resectable, 54 locally advanced and 41 metastatic PDAC. All patients received FOLFIRINOX treatment, and tumor responses were categorized using RECIST 1.1 [5]. First a thorough data analysis, including outlier and principal component analysis (PCA) is conducted to identify patterns and relationships between and within variables. Subsequently, a robust classification model was built using random forest modeling, accounting for dataset imbalance. Three optimal models are proposed based on (a) only pre-chemotherapy values, (b) values before and after the first FOLFIRINOX cycle and (c) only the top 10 identified variables in (b). For each of the most important variables, partial dependence and accumulated local effect plots are generated to gain further insights into their marginal effect on the classification outcomes.

The initial data analysis revealed the prognostic and predictive significance of the tumor marker CA19-9, both before and after one cycle of treatment, and the difference in its levels. Additionally, various blood markers, including Hemoglobin, Thrombocytes and γ-Glutamyl Transferase, showed associations with treatment outcomes. The assessment of variable importance further confirmed these relationships between tumor and blood markers and their impact on treatment response. However, PCA did not identify significant patterns or relationships within or between groups of blood markers. Moreover, the developed random forest classification models exhibited promising balanced accuracy, with values of 0.97, 0.98, and 0.90 for models (a), (b), and (c), respectively, in stratifying PDAC patients into the two distinct response groups (disease control and progressive disease), facilitating treatment decision-making.

In conclusion, this master thesis emphasizes the crucial role of comprehensive and rigorous data analysis in PDAC research, particularly when employing machine learning for predicting treatment outcomes. Integrating information from measured tumor and blood markers into the random forest models enables the prediction to FOLFIRINOX therapy both before and after one cycle. The implications of these findings are significant, as they can lead to improved patient management, efficient allocation of resources, personalized approaches, and contribution to the research and development efforts in PDAC. To validate and expand upon the presented results, further studies are required, ultimately advancing the field of personalized medicine in pancreatic cancer.

Keywords: *pancreatic cancer, PDAC, prediction model, random forest, data analysis, outlier analysis, principal component analysis.*

Preface

Dear reader,

The completion of this master's thesis marks the end of my academic journey as an Applied Mathematics student at Delft University of Technology. I developed a mathematical model using blood sample data to categorize pancreatic cancer patients based on their response to chemotherapy and therefore combined my passion for mathematics with a lifelong aspiration to become a doctor. The absence of such a model in current medical practice troubled me deeply, and when approached by the Erasmus Medical Centre Rotterdam to undertake this research, I readily accepted. Over the past ten months, I immersed myself in understanding the biological mechanisms of pancreatic cancer, conducting extensive literature reviews, and analyzing the provided dataset to construct a classification model. My goal was to provide medical professionals and patients with an evidence-based tool for informed treatment decisions. Additionally, I aimed to contribute to the identification of significant biomarkers and enhance the understanding of treatment responses through detailed data analysis. I hope that the outcomes of this research will have a positive impact, empowering medical professionals to make personalized treatment choices and improving the lives of individuals affected by pancreatic cancer.

One of the driving forces behind my research journey, and a constant source of motivation throughout, has been the personal experience of my aunt's battle with pancreatic adenocarcinoma (PDAC). Despite her healthy lifestyle and access to advanced treatments, her fight against the disease was unfortunately cut short after 18 months. This sad example highlights the aggressiveness of the disease and ongoing knowledge gap in the management of pancreatic cancer, underscoring the need for comprehensive data analysis that integrates clinical and biomarker information. Moreover, I am deeply grateful for the guidance and support of my supervisors, whose expertise and mentorship have been invaluable. Geurt, thank you for providing me with the opportunity to collaborate with the Erasmus MC. Your mentorship, openness to my ideas, and insightful inquiries have enriched this research and kept me focused on the primary objectives. Gaby, I appreciate your support and supervision throughout the development of my thesis. Your prompt responses and clinical experience have been immensely helpful. Thank you for your constant encouragement and for helping me navigate the world of PDAC. Casper, I am grateful for giving me this opportunity and I hope to have met your expectations. I would also like to extend my gratitude to the PDAC department at Erasmus MC for making me feel welcomed as part of your community.

Furthermore, I am also grateful for the support I have received from my parents and friends. Mom and Dad, thank you for reminding me of the importance of taking breaks and working hard in equal measure. Thank you for your endless love, support, and care. Lastly, I would like thank everyone I encountered during my time at TU Delft. It is through your contributions that these years have been truly unforgettable.

I wish you a pleasant reading.

Jenny Lu
Delft, August 2023

Nomenclature

Abbreviation	**Meaning**
AF	Alkaline Phosphatase
ALAT	Alanine Aminotransferase
ALE	Accumulated Local Effects
AUC	Area Under the Curve
ASAT	Aspartate Aminotransferase
BR	Bilirubin
CA19-9	Carbohydrate Antigen 19-9
CEA	Carcinoembryonic antigen
CII	Complete Class Inclusion
CR	Complete Response
CRP	C-Reactive Protein
DC	Disease Control
GFR	Glomular Filtration Rate
GGT	Gamma-Glutamyl Transferase
HB	Hemoglobin
IQR	Inter-Quartile Range
INR	International Normalized Ratio
K	Potassium
LAPC	Locally Advanced Pancreatic Cancer
LC	Lymphocytes
LK	Leukocytes
ML	Machine Learning
M-plot	Marginal Effect Plot
NLR	Neutrophil-to-Lymphocyte Ratio
NP	Neutrophils
OOB	Out-Of-Bag
OS	Overall Survival
OVB	Omitted Variable Bias
PCA	Principal Component Analysis
PC	Principal Component
PD	Progressive Disease
PDAC	Pancreatic Ductal Adenocarcinoma
PDP	Partial Dependence Plot
PFS	Progression Free Survival
PLR	Platelet-to-Lymphocyte Ratio
PR	Partial Response
RECIST	Response Evaluation Criteria in Solid Tumors
ROC	Receiver Operating Characteristic
ROSE	Random Over-Sampling Examples
SII	Systemic Inflammation Index
SD	Stable Disease
TB	Thrombocytes
WBC	White Blood Cells

Table 1: Table of abbreviations and their full meaning used in this thesis.

Variable name	**Meaning**	**Response Label**	**Response value**
Study Subject ID	Study Subject ID		
Protocol ID	Protocol ID		
INFORMED_CONSENT_E1_C1	Informed consent given	No, Yes	0,1
DATE_IC_E1_C1	Date of Informed consent		
DATE_BIRTH_E1_C2	Date of Birth		
GENDER_E1_C2	Gender	Male, Female	0,1
LENGTH_E1_C2	Length (cm)		
WEIGTH_E1_C2	Weight (kg)		
STAGE_DISEASE_E1_C2	Stage of disease	(Borderline) resectable, Locally advanced, Metastatic disease	0, 1, 2
CA199_DIAGN_E1_C2	CA-19.9 (kU/L) at diagnosis		
CEA_DIAGN_E1_C2	CEA (µg/L) at diagnosis		
FAMILY_HISTORY_E1_C2	Positive family history for pancreatic cancer	No, Yes, Unknown	0, 1, 2
SMOKING_E1_C2	Smoking	Never, Former, Current, Unknown	0, 1, 2, 3
ALCOHOL_E1_C2	Alcohol use	No, Yes, Stopped, Unknown	0, 1, 2, 3
DM_E1_C2	Diabetes mellitus	No, Yes	0, 1
HISTORY_PANCREAT_E1_C2	History of pancreatitis	No, Yes	0, 1
HISTORY_MALIGN_E1_C2	History of malignancy	No, Yes	0, 1
HISTORY_MALIGN_SPEC_E1_C2	Previous malignancy specified		
HISTORY_CHEMO_E1_C2	History of chemotherapy	No, Yes	0, 1

Table 2: Patient characteristic variable names and their corresponding definitions.

Variable Name	**Meaning**	**Section Label**	**Unit**
CA199_CHEMO_E2_C3	CA19-9 (tumor marker)	LAB_CHEMO	kU/L
CEA_CHEMO_E2_C3	CEA (tumor marker)	LAB_CHEMO	µg/L
HB_CHEMO_E2_C3	Hemoglobin	LAB_CHEMO	mmol/L
TROMBOCYTES_CHEMO_E2_C3	Thrombocytes	LAB_CHEMO	$10^9/L$
LEUKOCYTES_CHEMO_E2_C3	Leukocytes	LAB_CHEMO	$10^9/L$
NEUTROPHILS_CHEMO_E2_C3	Neutrophils	LAB_CHEMO	$10^9/L$
LYMPHOCYTES_CHEMO_E2_C3	Lymphocytes	LAB_CHEMO	$10^9/L$
CREATININ_CHEMO_E2_C3	Creatinin	LAB_CHEMO	µmol/L
GFR_CHEMO_E2_C3	Glomular Filtration Rate	LAB_CHEMO	mL/min
SODIUM_CHEMO_E2_C3	Sodium	LAB_CHEMO	mmol/L
POTASSIUM_CHEMO_E2_C3	Potassium	LAB_CHEMO	mmol/L
ASAT_CHEMO_E2_C3	Aspartate Aminotransferase	LAB_CHEMO	U/L
ALAT_CHEMO_E2_C3	Alanine Aminotransferase	LAB_CHEMO	U/L
AF_CHEMO_E2_C3	Alkaline Phosphatase	LAB_CHEMO	U/L
GGT_CHEMO_E2_C3	γ-Glutamyl Transferase	LAB_CHEMO	U/L
BR_CHEMO_E2_C3	Bilirubin	LAB_CHEMO	$\mu mol/L$
INR_CHEMO_E2_C3	International Normalized Ratio	LAB_CHEMO	U/L
ALB_CHEMO_E2_C3	Albumin	LAB_CHEMO	g/L
CRP_CHEMO_E2_C3	C-Reactive Protein	LAB_CHEMO	mg/L

Table 3: Table of the measured variable names and their respective meaning in the dataset.

Variable Name	**Meaning**	**Response label**	**Response Values**
SURVIVAL_ STATUS_E6_C10	Survival status	Alive, Dead, Lost in follow-up	0, 1, 2
DATE_DEATH _E6_C10	Date of death		
CAUSE_DEATH _E6_C10	Cause of death	Progression of disease, Complications of treatment chemotherapy or surgery), Unknown, Other	0, 1, 2 , 3
DATE_LAST_ ALIVE_E6_C10	Date last seen alive		
OTHER_TREATMENT _E6_C10	Other treatment after chemotherapy	No, Yes, Unknown	0, 1, 2
TREATMENT_ SPEC_E6_C10	Type of therapy	Radiotherapy, Surgery, Immuno therapy, Other	0, 1, 2, 3
PROGRESSION _E6_C10	Progression	No, Yes, Unknown	0, 1, 2
DATE_ PROGRESSION _E6_C10	Date of progression		
CYCLES_ FOLFIRINOX _E6_C11	Total amount of cycles FOLFIRINOX received		
FINAL_ RESPONSE_ OUTCOME_ E6_C11	Final response outcome	Complete response, Partial response, Progressive disease, Stable disease, Unknown	0, 1, 2, 3, 4
FINAL_R ESPONSE_ DICHOTOMIZED _E6_C11	Disease control (= stable disease, partial response or complete response) or progressive disease after last cycle of FOLFIRINOX	Disease control, Progressive disease, Unknown	0, 1, 2
G_CSF_ YES_NO _E6_C11	Granulocyte Colony Stimulating Factor received	No, Yes	0, 1
G_CSF_ PROPHYLAXIS _E6_C11	Granulocyte Colony Stimulating Factor prophylaxis	No, Yes	0, 1
G_CSF_ AFTER_ CYCLE_E6_C11	Granulocyte Colony Stimulating Factor start		

Table 4: Table of variable names and their respective meaning in the dataset.

Contents

1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) ranks as the 12^{th} most common cancer worldwide according to the World Cancer Research Fund International [6]. The mortality rates for PDAC are alarmingly high, almost equal to the incidence rates, with a mere 7.7% 5-year survival rate [1]. Its incidence is rising [7], and by 2030, PDAC is predicted to be the second leading cause of cancer-related deaths, surpassing colorectal and breast cancer [8]. PDAC presents significant challenges in terms of diagnosis and treatment, particularly due to the majority of patients (80%) being diagnosed at an advanced stage [3], rendering them ineligible for surgical resection [9]. Currently, chemotherapy serves as the standard treatment for these patients, while the search for more effective therapies continues.

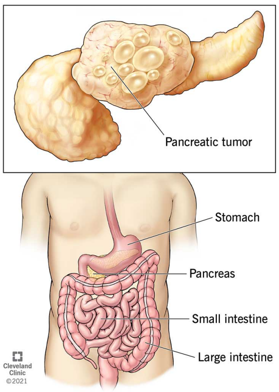


Figure 1.1: Illustration of the anatomy of the most important organs surrounding the pancreas [10].

Class	Criteria
Complete Response (CR)	*Disappearance of all target lesions*
Partial Response (PR)	*Decrease of* $\geq$ *30% in the sum of diameters of target lesions*
Progressive Disease (PD)	*Increase of* $\geq$ *20% in the sum of diameters of target lesions* *or the appearance of a new cancer lesion.*
Stable Disease (SD)	*Neither sufficient shrinkage nor increase to be classified as PR or PD.*

Table 1.1: Classification of PDAC patients based on the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 criteria [5].

For a long period of time, upfront surgery followed by adjuvant gemcitabine has been the standard treatment for (borderline) resectable PDAC [11]. However, neoadjuvant therapy has emerged as a promising strategy for localized PDAC, offering potential advantages over upfront surgery [3]. FOLFIRINOX, a combination of fluorouracil, leucovorin, irinotecan, and oxaliplatin, has shown promise in the neoadjuvant setting, although further randomized studies are needed to establish its efficacy. One such study is the Dutch PREOPANC-2 trial, a multicenter randomized phase III trial conducted by the Dutch Pancreatic Cancer Group [12]. The trial demonstrated a significant improvement in overall survival (OS) with neoadjuvant therapy compared to upfront surgery [12]. Therefore, pre-operative chemotherapy may become the new standard in the future. For locally advanced pancreatic cancer (LAPC) and metastatic disease, FOLFIRINOX is considered the preferred treatment option. In a phase III multicenter randomized study, FOLFIRINOX demonstrated superior outcomes compared to gemcitabine monotherapy, with significantly longer median progression-free survival (PFS) (6.4 vs. 3.3 months) and OS (11.1 vs. 6.8 months) [13]. The survival benefits were consistent across patient subgroups, and additional evidence from pooled phase II trials and off-trial series supported the observed OS of approximately 10-11 months with FOLFIRINOX [13]. A recent meta-analysis showed a response rate [1] of only 25-30% in metastatic patients treated with FOLFIRINOX, indicating that a substantial proportion of patients may not derive significant clinical benefit from this

[1] The Overall Response Rate refers to the PR and CR groups together. However, the disease control rate consists of PR, CR and SD. Consequently the percentages are higher.

treatment [4]. When considering the disease control rates (DCR), they found that overall the DCR was 70.95% (95% CI 58.0-83.9%) [4]. However, other studies like [14] show a FOLFIRINOX DCR of 77.2% in LAPC and 51.4% in metastatic patients, respectively. An overview of the current way of deciding between treatment options based on the different stages of PDAC is provided in Figure 1.2. Further information on clinical trials and their phases can be found in Appendix E. Besides, treatment with FOLFIRINOX based chemotherapy in particular, is also associated with significant side effects. Reports have indicated high rates of ≥ grade 3 toxicities, reaching 60-70% [15] [16] [17]. Adverse effects commonly encountered with FOLFIRINOX include leukopenia, diarrhea, polyneuropathy, and infectious complications. Leukopenia, infections, and diarrhea are most likely to occur during the initial cycles of chemotherapy, while polyneuropathy occures later. Strategies such as dose modifications and the use of growth factor support, including Granulocyte-Colony-Stimulating Factor (G-CSF) and erythropoietin, have been suggested to improve tolerability and manage these adverse effects. These growth factors stimulate the production of white and red blood cells, helping to prevent or manage complications such as infections and anemia during chemotherapy.

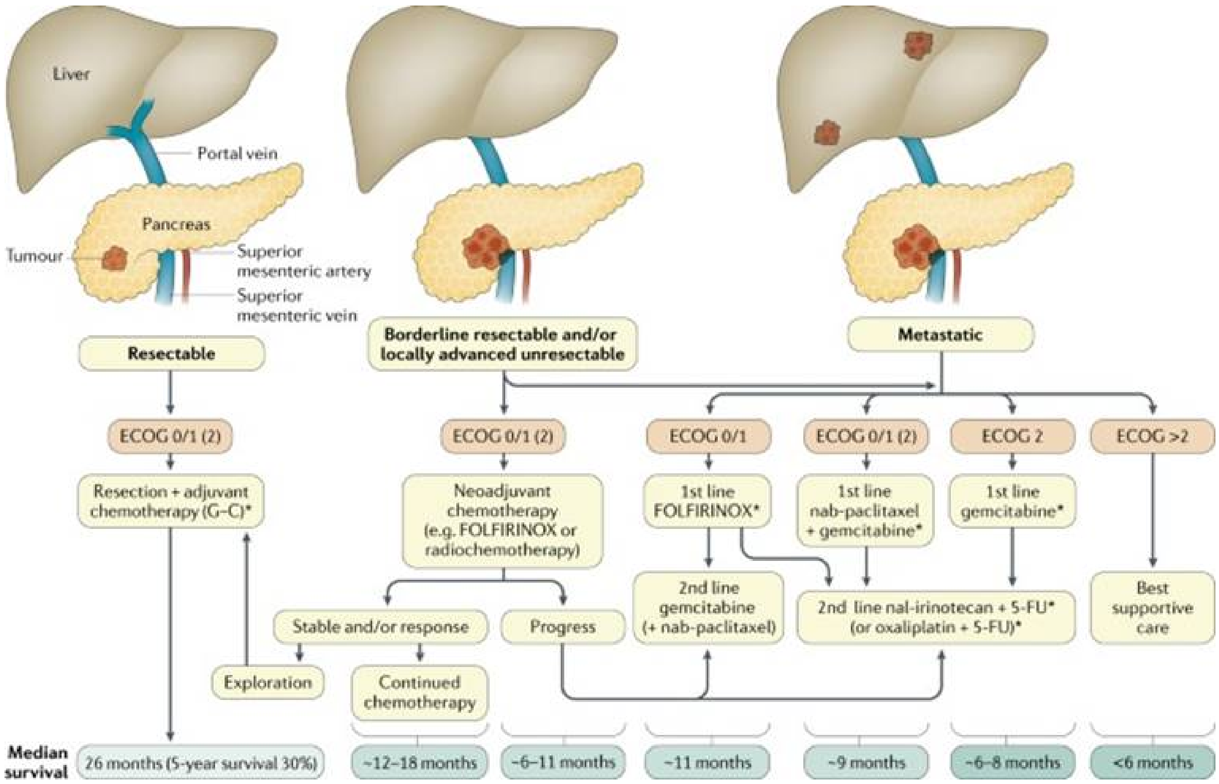

Liver
Portal vein
Pancreas
Tumour
Superior mesenteric artery
Superior mesenteric vein
Resectable
Borderline resectable and/or locally advanced unresectable
Metastatic
ECOG 0/1 (2)
ECOG 0/1 (2)
ECOG 0/1
ECOG 0/1 (2)
ECOG 2
ECOG >2
Resection + adjuvant chemotherapy (G-C)*
Neoadjuvant chemotherapy (e.g. FOLFIRINOX or radiochemotherapy)
1st line FOLFIRINOX*
1st line nab-paclitaxel + gemcitabine*
1st line gemcitabine*
2nd line gemcitabine (+ nab-paclitaxel)
2nd line nal-irinotecan + 5-FU* (or oxaliplatin + 5-FU)*
Best supportive care
Stable and/or response
Progress
Exploration
Continued chemotherapy
Median survival
26 months (5-year survival 30%)
~12–18 months
~6–11 months
~11 months
~9 months
~6–8 months
<6 months

Figure 1.2: Current treatment guideline for different disease stages in PDAC [18].

Moreover, certain biomarkers have already been linked to influence OS in PDAC patients. One established conventional tumor marker, Carbohydrate Antigen 19-9 (CA19-9), is widely studied as a diagnostic, prognostic and predictive biomarker in PDAC [19]. Nevertheless, up to 20% of the pancreatic cancer patients is Lewis negative and cannot synthesize CA19-9 [20] [21]. Azzariti et al investigated CA19-9 levels in metastatic patients treated with FOLFIRINOX. Twenty-seven patients were studied of which twenty-one received FOLFIRINOX and six gemcitabine based treatment. CA19-9 levels pre-treatment and post-treatment were compared within the PR, SD and PD patient groups. Decreasing CA19-9 levels after treatment were found in patients with PR or SD, while an increasing trend was found in patients with PD [22]. Another conducted by Schlick et al. [23] identified two other accessible factors that independently influence OS in patients receiving FOLFIRINOX therapy, namely elevated Carcinoembryonic Antigen (CEA) levels > 4 and a body mass index (BMI) > 25 prior to the initiation of palliative chemotherapy. These factors were identified as negative prognostic indicators for OS. Additionally, the study revealed that thrombocytosis and low BMI were predictors of early treatment-related toxicity [23]. Furthermore, in recent years, there has been an increasing interest in the identification of ciculating biomarkers to predict the response to chemotherapy. Various biomarkers, including genetic and immunological markers, have shown potential prognostic value. Van der Sijde et al. [24] discovered a panel of circulating biomarkers with promising predictive capabilities. However, comparing findings across studies is challenging due to variations in biomarkers, cut-off values, and chemotherapy regimens. Due to limited validation and small patient populations in many studies, these biomarkers can only be considered potential predictors for now.

Despite advancements in pancreatic cancer research, a knowledge deficit remains in analyzing complex datasets for disease management. This master's thesis aims to develop a robust classification model for PDAC patients treated with FOLFIRINOX using comprehensive blood results from two multi-center trials in the Netherlands (PREOPANC-2 NL7094, and iKnowIT NL7522). The dataset includes patients with all stages of PDAC, and measurements were performed before and after the first chemotherapy cycle. Data analysis techniques such as principal component analysis, outlier analysis, and random forest modeling are used to classify patients into disease control (DC) and progressive disease (PD) groups, classified according to the RECIST 1.1 criteria [5]. The analysis also identifies key variables contributing to the model's performance, providing insights into significant blood and tumor markers for PDAC patient outcomes.

The master's thesis is structured into several sections aimed at analyzing the provided dataset comprehensively. Section 2.1 offers background information on PDAC and the measured variables. This section serves as a foundation for understanding the subsequent analyses. Following the literature review, the dataset undergoes a thorough data analysis in section 3.1. This analysis aims to explore the dataset and gain insights into the blood markers associated with PDAC, as well as identify potential relationships and differences across the final response groups. A detailed account of the data analysis is provided in Appendix B.1. To ensure the accuracy and reliability of subsequent analyses, outlier analysis is performed in section 3.2, with an in-depth examination presented in Appendix B.2. This process identifies and addresses any anomalous data points that could impact the validity of the results. To reduce the dataset's dimensionality while preserving its most informative features, principal component analysis (PCA) is used. This technique is applied after grouping the blood markers based on their characteristics, providing a comprehensive understanding of the underlying patterns within each group. A summary of the most significant findings from the PCA is outlined in section 3.3, with the complete analysis based on the correlation matrix provided in Appendix B.3, the covariance matrix in Appendix B.5 and without outliers in Appendix B.6. Subsequently, three random forest classification models are developed in section 5, with additional mathematical background provided in section 4.1. These models use the blood results as predictors to classify PDAC patients into the DC or PD categories. The three models include a random forest model incorporating all variables after removal of outliers, a model featuring only the top 10 most important variables identified in the previous model, and a model solely using variables measured pre-chemotherapy. Performance assessment of the random forest classifier includes metrics such as balanced accuracy, precision, and f1-score, and strategies to address the class imbalance between the DC and PD group.

Additionally, a variable importance analysis is conducted based on the optimized random forest models to identify the most important blood markers contributing to the classification model's predictive performance. In order to enhance the understanding of the relationship between the identified key variables and the final response, partial dependence plots and accumulated local effect plots, are presented in section 5.2 of which the mathematical background for these visualizations is explained in section 4.2. These plots provide insights into the marginal effects of the most important variables on the final response. Medical interpretation of the important variables is provided in section 5.3 with clinical implications given in section 5.4. Finally, section 6 addresses the study's limitations and presents recommendations for future research and section 7 offers a conclusive summary of the thesis. Overall, using data analysis and machine learning, this study intents to contribute to clinical decision-making and patient stratification. Ultimately, this might lead to improved patient outcomes and efficient resource utilization in the management of pancreatic cancer care.

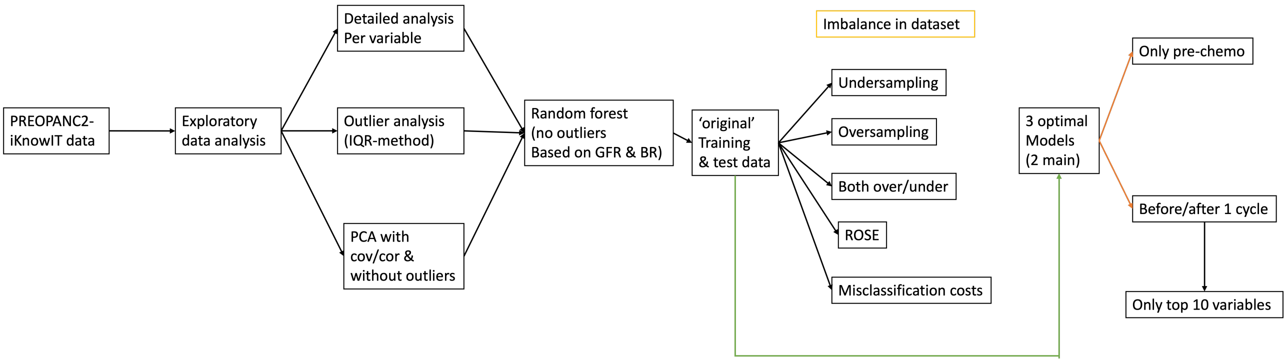


Figure 1.3: Steps taken in this master thesis from data analysis to final model

2 | Literature Review

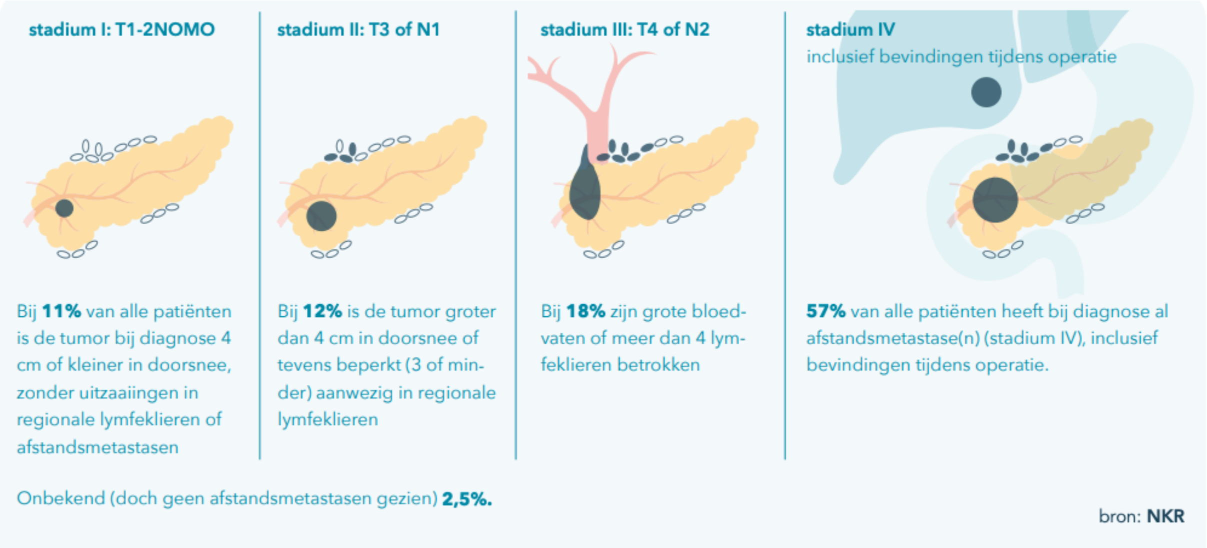
2.1. Background information

This section aims to provide background information on relevant papers that were reviewed to gain a comprehensive understanding of PDAC, FOLFIRINOX treatment, biomarkers, and other important concepts used throughout the thesis. Key concepts from previous studies as well as relevant clinical background information will be summarized. Additionally, a concise explanation of the measured tumor and blood markers will be provided.

2.1.1 Pancreatic Ductal Adenocarcinoma

Pancreatic cancer has exhibited a significant increase in incidence over the years, with new cases rising from approximately 1400 in 1995 to approximately 2922 in 2021 in the Netherlands [25]. Within the realm of pancreatic cancer, Pancreatic Ductal Adenocarcinoma (PDAC) emerges as the most predominant subtype, accounting for about 93% of diagnosed cases. Other subtypes, such as Neuro-Endocrine Tumors (NET), are less prevalent [25]. Moreover, the survival rate for PDAC remains significantly lower compared to other cancer types. While the collective 5-year survival rate for all cancers hovers around 65%, the corresponding rate for pancreatic cancer (encompassing all subtypes) is merely 7%. Notably, PDAC patients experience an even more dismal 5-year survival rate of only 3.8% compared to those with other subtypes.

In the context of the Netherlands, PDAC cases are evenly distributed between men and women, with an average age of approximately 71 years. A substantial proportion, around 71%, of diagnosed patients exhibit a good performance status based on the ECOG performance status score (ECOG 0-1) [15] [25]. Moreover, more than half of pancreatic cancer patients have reported comorbidities, with 28% diabetes, 12% lung problems, 8% history of other cancers diagnosis, and 7% having a history of a heart attack [25]. The diagnosis of pancreatic cancer primarily relies on Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans, and other tissue research methods. Among the diagnosed cases, more than half of the tumors are located in the head of the pancreas. Due to the manifestation of symptoms at an advanced stage, a majority (57%) of diagnosed patients already present with metastatic cancer. The most prevalent form of distant metastasis in these cases is observed in the liver, affecting approximately 75% of patients with metastatic disease, with 43% having isolated metastases in the liver [25].



Bij **11%** van alle patiënten is de tumor bij diagnose 4 cm of kleiner in doorsnee, zonder uitzaaiingen in regionale lymfeklieren of afstandsmetastasen

Bij **12%** is de tumor groter dan 4 cm in doorsnee of tevens beperkt (3 of minder) aanwezig in regionale lymfeklieren

Bij **18%** zijn grote bloedvaten of meer dan 4 lymfeklieren betrokken

57% van alle patiënten heeft bij diagnose al afstandsmetastase(n) (stadium IV), inclusief bevindingen tijdens operatie.

Onbekend (doch geen afstandsmetastasen gezien) **2,5%.**

bron: **NKR**

Figure 2.1: Stadium at diagnosis of pancreatic cancer [25].

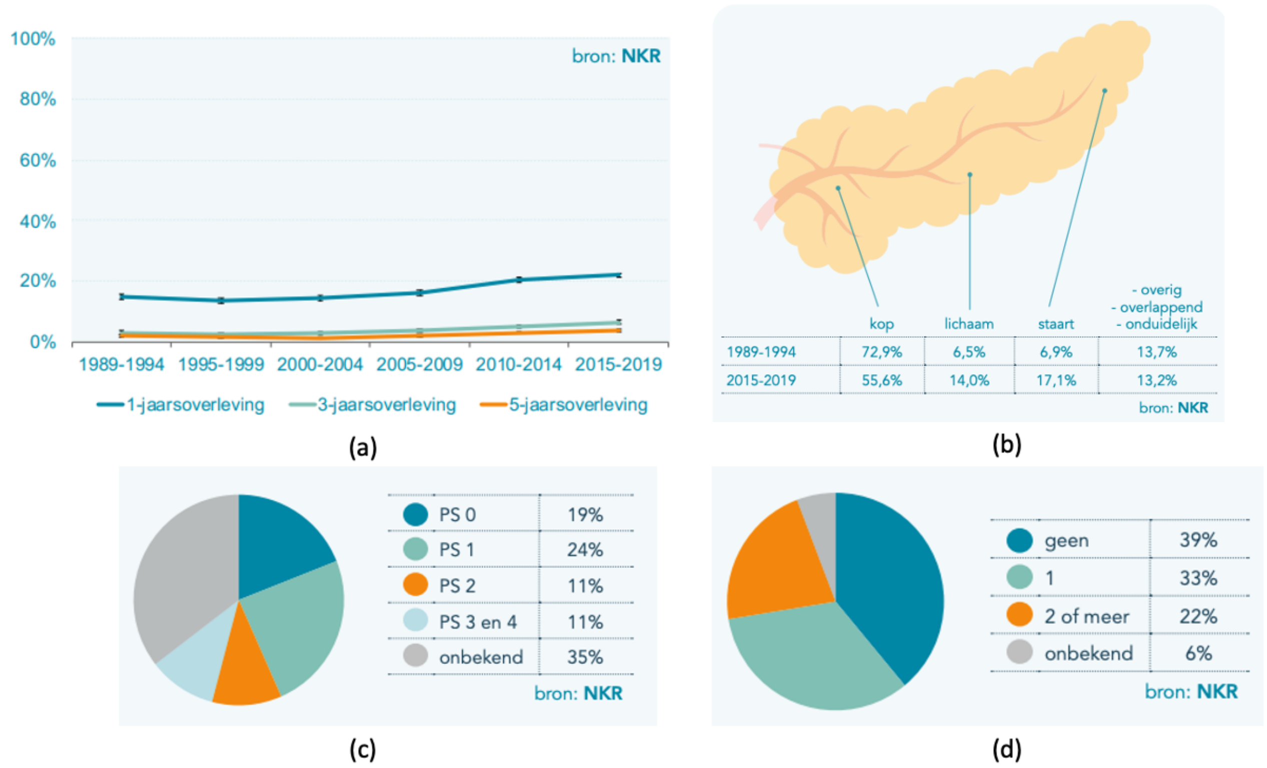


Figure 2.2: Pancreatic Cancer data in the Netherlands. (a) Relative survival rate of PDAC patients from 1989 to 2019 (dark blue = 1 year, light blue = 3 years, orange = 5 years survival rate) (b) Location of pancreatic cancer (c) Performance status at diagnosis of pancreatic cancer in 2019 (d) Number of comorbidities at diagnosis of pancreatic cancer in 2019[25].

2.1.2 Biomarkers

Biomarkers are defined as measurable indicators of biological processes, disease states, or responses to treatments and play a crucial role in providing information about normal physiological processes, pathological conditions, or the effects of interventions. They encompass a wide range of molecules, substances, or characteristics that can be objectively measured and evaluated. Examples include proteins, genetic markers, imaging markers, and physiological measurements [26]. Ideally, biomarkers should be cost-effective, easy to analyze, and non-invasive for the patient. The current PREOPANC2-iKnowIT dataset only consists of the tumor markers CA19-9 and CEA, as well as conventional blood markers. Therefore, the data will be divided into two categories: tumor markers and blood markers. The tumor markers will encompass the CA19-9 and CEA, while the blood markers will be grouped based on their properties. Blood samples have garnered increasing attention in PDAC biomarker research due to their distinctive attributes, such as accessibility and potential for non-invasive monitoring of treatment response. Additional reasons that underscore the usefulness of blood samples in PDAC biomarker research are as follows:

- **Representative tumor biopsies:** Blood samples offer an alternative to invasive biopsies, allowing for accessible and non-invasive investigation of molecular characteristics of cancer cells or the immune component of the disease.
- **Intratumoral heterogeneity:** PDAC exhibits significant intratumoral heterogeneity [1], which cannot be adequately captured by a single biopsy. However, blood samples have the potential to provide a more comprehensive molecular profile of both the primary tumor and its distant metastases.
- **Longitudinal monitoring:** Peripheral blood can be sampled repeatedly with minimal safety risks and discomfort, allowing for longitudinal monitoring of PDAC patients and assessment of treatment response over time.

[1] Intratumoral heterogeneity refers to the presence of diverse genetic and molecular characteristics within the same tumor. This variation can lead to differences in how tumor cells behave and respond to treatments, making it challenging to predict and target the tumor's behavior uniformly. Essentially, it means that different parts of the same tumor can exhibit distinct genetic profiles, potentially influencing disease progression and treatment outcomes.

2.1.3 Tumor Markers

2.1.3.1 Carbohydrate Antigen 19-9 (CA19-9)

Carbohydrate Antigen 19-9 (CA19-9) is a sialylated Lewis blood group antigen associated with various cancers, including PDAC [19]. It is widely investigated as a diagnostic, prognostic, and predictive biomarker in PDAC [27], [28]. The recommended upper limit of CA19-9 as a biomarker for pancreatic cancer, as suggested by medical experts, is 37 U/mL [29]. However, approximately 20% of patients with pancreatic cancer exhibit no or low secretion of CA19-9. Specifically, the Lewis (-) individuals, comprising about 5%-10% of the global population, have been identified as lacking or minimally expressing CA19-9 [30], [31], [20], [21]. The Lewis genotypes, determined by certain variants in the Lewis gene, such as T59G, T202C, C314T, G508A, and T1067A, play a role in the absence of CA19-9 expression [32]. Prognostically, Lewis (-) patients with advanced pancreatic cancer exhibit worse outcomes compared to Lewis (+) patients [33]. Hence, despite the promise of CA19-9 as a tumor marker, it cannot be used for Lewis (-) individuals.

Furthermore, several studies have demonstrated that Lewis (+) pancreatic cancer patients with normal CA19-9 levels (< 37 U/mL) exhibit better prognoses in comparison to those with elevated CA19-9 levels (> 37 U/mL) [34], [35], [36]. These findings suggest that pancreatic cancer patients with low CA19-9 secretion and Lewis (+) genotype may have long-term survival prospects. Additionally, a collective analysis of multiple studies has indicated that baseline as well as post-treatment CA19-9 levels are generally higher in non-responders than in responders. Particularly, the post-treatment CA19-9 levels have shown more significant differences between responders and non-responders. In general, the overall consensus is that CA19-9 levels decrease over time in responders, while they either remain unchanged or increase in non-responders [24].

2.1.3.2 CarcinoEmbryonic Antigen (CEA)

CarcinoEmbryonic Antigen (CEA) is another promising tumor marker that is known to be over-expressed in various cancers, including breast, lung, and thyroid cancers [37], [38], [39]. Elevated levels of CEA have been observed in patients with PDAC and have been suggested as an independent predictor of poor survival rates in these patients [40], [41], [42]. In a study conducted by Hiroshi et al. [40], a cohort of 433 patients with metastatic disease was examined, and their CEA and CA19-9 levels were analyzed. Among these patients, 36 had normal levels of both CEA and CA19-9, 149 had elevated levels of both markers, 30 had high CEA levels alone, and 218 had high CA19-9 levels alone. The study revealed that patients with high CEA levels had a significantly higher prevalence of liver metastasis, while patients with normal CEA levels had a significantly higher prevalence of lung metastasis. Furthermore, the median overall survival (OS) was significantly shorter for patients with high CEA levels compared to those with normal CEA levels (6.8 months vs. 10.3 months, respectively). These findings suggest that patients with elevated CEA exhibit distinct biological behavior compared to those with normal CEA levels. It is important to note that the patients included in the study received different forms of therapy, and there were no statistically significant differences in outcomes between patients treated with single-agent chemotherapy or combination chemotherapy [40]. Additionally, CEA, along with CA125, has been identified as important biomarkers for Lewis (-) individuals [33].

2.1.4 Blood Markers

A brief explanation of each of the measured blood markers provided in the PREOPANC2-iKnowIT dataset and their functionality in the human body is provided in this subsection. A summary of the data analysis is given in section 3.1 with the full analysis in Appendix B.1. The blood markers are categorized into groups based on their type and function to explain their functionality as in Table 2.1. When conducting principle component analysis (PCA) in section 3.3, with the full analysis presented in Appendix B.3, these groupings will be used.

Group	Blood Markers
Blood Cells	Hemoglobin (HB), Thrombocytes (TB), Leukocytes (LK), Neutrophils (NP), Lymphocytes (LC)
White Blood Cells	Leukocytes (LK), Neutrophils (NP), Lymphocytes (LC)
Kidney Function	Sodium (Na), Potassium (K), Creatinin (CR), Glomular Filtration Rate (GFR)
Liver Function	Aspartate Aminotransferase (ASAT), Alanine Aminotransferase (ALAT), Alkaline Phosphatase (AF), Gamma-Glutamyl Transferase (GGT), Bilirubin (BR), International Normalized Ratio (INR)
Nutritional Status	Albumin (Alb), Sodium (Na), Potassium (K)
Inflammation	Systemic Inflammation Index (SII), Neutrophil-to-Lymphocyte Ratio (NLR), Platelet-to-Lymphocyte Ratio (PLR), C-reactive protein (CRP), Leukocytes (LK)

Table 2.1: Blood marker grouping based on their type and functionality.

2.1.4.1 Blood cells: Hemoglobin (HB), Thrombocytes (TB), Leukocytes (LK), Neutrophils (NP), Lymphocytes (LC)

The group of blood cell measurements designated for this study comprises of hemoglobin (HB), thrombocytes (TB), leukocytes (LK), neutrophils (NP), and lymphocytes (LC). Hemoglobin is essential for the proper functioning of the human body as it plays a crucial role in transporting oxygen from the lungs to the organs and tissues. This protein is present in red blood cells (erythrocytes) and is composed of four protein molecules, called globulin chains, along with four heme molecules that contain iron. Moreover, thrombocytes, also referred to as platelets, are small cell fragments present in the blood that are responsible for blood clotting. These platelets rush to the sites of injury, where they aggregate to form a clot, which helps to stop bleeding and facilitates the healing process. Additionally, leukocytes, also known as white blood cells, play a crucial role in the body's immune system by fighting off infections and diseases. Among leukocytes, neutrophils are a specific type that primarily work by engulfing and destroying invading microorganisms. Similar to neutrophils, lymphocytes are also a type of leukocyte and their main responsibility is the identification and neutralization of foreign invaders, such as viruses and bacteria.

The bone marrow plays a crucial role in the production of erythrocytes, thrombocytes, and leukocytes. These precursor blood cells are classified as rapidly dividing cells, similar to tumor cells. Therefore, chemotherapy, which primarily targets rapidly dividing cells, with the aim of tumor cells, will as a result also unintentionally affects healthy rapidly dividing cells leading to a decline in bone marrow function. This phenomenon is known as “bone marrow depression" and hinders the production of new blood cells. Consequently, the efficacy of chemotherapy is impacted, as not only the tumor cells are killed but also many healthy cells. Especially neutrophils are affected, leading to a condition known as neutropenia, characterized by a low count of neutrophils [43]. This condition can increase the patient's risk of developing infections. Additionally, chemotherapy can lead to a decrease in the number of lymphocytes, leading to a condition known as lymphopenia. Therefore, medication is often provided to simulate white blood cell production, such as granulocyte-colony stimulating factor (G-CSF). Given the impact of chemotherapy on white blood cell counts, it becomes crucial to monitor their levels throughout the treatment process. Analyzing the fluctuations in these cells' levels can provide valuable insights into the patient's response to chemotherapy.

2.1.4.2 White Blood Cells: Leukocytes (LK), Neutrophils (NP), Lymphocytes (LC)

The White Blood Cell (WBC) group consists of the three measured white blood cell variables: Leukocytes (LK), Neutrophils (NP) and Lymphocytes (LC). Leukocytes are actually another name for all the white blood cells together, with NP and LC being two subtypes of LK. However, since we have measurements of all these three variables, we will classify them all under the name white blood cell (WBC). To be specific, there exist two main categories of leukocytes, these are granulocytes and agranulocytes of which the latter is split in lymphocytes, and monocytes. Granulocytes are split in neutrophils, eosinophils and basophil cells. The combined function of leukocytes is to defend the body against pathogenic microorganisms and infections. The ‘hierarchy’ of leukocytes can be described as follows:

- **Leukocytes**: Leukocytes, commonly known as white blood cells, are a type of cell that play a vital role in the immune system. They are responsible for defending the body against foreign invaders, such as bacteria, viruses and other pathogens. All white blood cells are produced in the bone marrow and can be found throughout the body. They distinguish themselves from red blood cells and platelets by having nuclei. Leukocytes can be split into two broad categories: granulocytes and agranulocytes.
- **Granulocytes**: Granulocytes, are a category of leukocytes that have granules in their cytoplasm. These are characterized by the presence of specific granules that contain various enzymes and proteins. The three main types of granulocytes are: neutrophils, eosinophils and basophils.
  - *Neutrophils*: Neutrophils are the most abundant type of granulocytes. These are involved in the innate immune response and are responsible for phagocytosis, targeting and destroying pathogens.
  - *Eosinophils*: Eosinophils are primarily involved in allergic reactions and combating parasitic infections.
  - *Basophils*: Basophils are the least common type of granulocyte. They play a role in allergic reactions and inflammation by releasing histamine and other mediators.
- **Agranulocytes**: Agranulocytes are a category of leukocytes that do not possess granules in their cytoplasm. The two main types of agranulocytes are: lymphocytes and monocytes
  - *Lymphocytes*: Lymphocytes are involved in adaptive immunity. They can be further classified into B-cells, T-cells and natural killer (NK) cells. B-cells produce antibodies, T-cells participate in cell-mediated immunity and NK cells are responsible for recognizing and eliminating infected or abnormal cells.
  - *Monocytes*: Monocytes are the largest type of leukocytes. They circulate in the blood stream and can differentiate into macrophages or dendritic cells when they migrate into tissues. Macrophages are involved in phagocytosis and antigen presentation, while dendritic cells are responsible for capturing and presenting antigens to activate the immune response.

A schematic overview of the ‘hierarchy’ in white blood cells is presented in Figure 5.17. More detailed information about B-cells, T-cells, Granulocytes, Monocytes, Macrophages and Dendritic cells can be found in Appendix E.

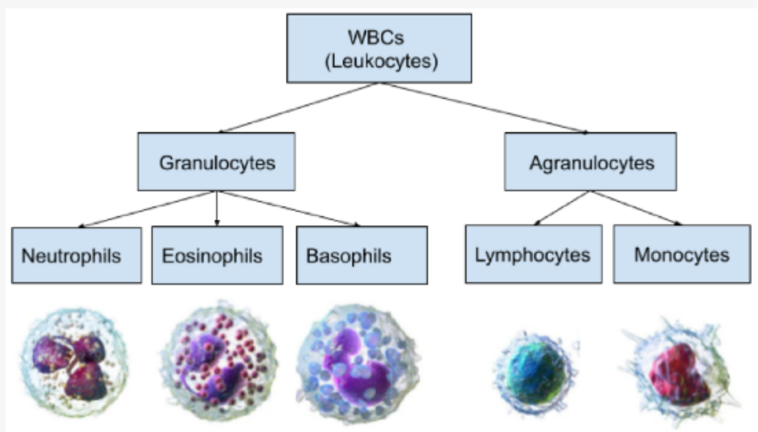


Figure 2.3: Overview of the different types of white blood cells [44].

In the context of pancreatic cancer, elevated leukocyte levels can be attributed to the immune response of the body against the tumor. Specifically, increased counts of neutrophils [43] , [45], monocytes [46], [47], and regulatory T-cells [48] have been identified as poor prognostic factors for survival, whereas elevated levels of total lymphocytes have shown associations with more favorable outcomes [49], [50], [51]. However, the functions of immune cells are often complex, and cell counts alone do not always indicate strictly positive or negative outcomes. The negative impact of elevated neutrophil counts on survival appears contradictory at first. An increased count could suggest an activated immune system that is actively combating cancer cells. However, in PDAC, neutrophilia is associated with poorer overall survival, whereas chemotherapy-induced neutropenia in patients with advanced PDAC is correlated with improved overall survival [45] [43]. This inverse correlation may be explained by an altered balance of pro-inflammatory and anti-inflammatory cytokines induced by tumors.

2.1.4.3 Kidney function: Sodium (Na), Potassium (K), Creatinin (CR), Glomular Filtration Rate (GFR)

The variables Sodium (Na), Potassium (K), Creatinin (CR) and Glomerular Filtration Rate (GFR) provide important information about the kidney functionality. Specifically, GFR is an indicator of the amount of blood (in milliliters) that is filtered by the kidneys per minute. Higher GFR values are generally indicative of better kidney function. Creatinin, on the other hand, is a waste product that results from the breakdown of proteins, for example from muscles. Its levels can increase as a result of cancer and/or chemotherapy. When GFR decreases, the kidneys filter creatinin less effectively, leading to an increase in creatinin concentration in the blood. However, individuals with lower muscle mass tend to have lower levels of creatinin overall, meaning that even if kidney function decreases, creatinin levels may remain low. Hence, elderly individuals and women tend to have lower creatinin levels compared to men.

Moreover, sodium is an important electrolyte that helps regulate fluid balance in the body. High levels of sodium can lead to fluid retention and swelling, while low sodium levels can cause muscle cramps, weakness and confusion. Potassium, another key electrolyte, plays a critical role in regulating fluid balance and muscle function. High levels of potassium can cause muscle weakness, heart palpitations and even cardiac arrest, while low levels can lead to muscle cramps, constipation and weakness. Cancer treatment can have side-effects, such as diarrhea or vomiting, that result in dehydration and alterations in sodium and potassium levels. Because both sodium and potassium are excreted by the kidneys, their levels also reflect the functioning of these organs.

2.1.4.4 Liver function: Aspartate Aminotransferase (ASAT), Alanine Aminotransferase (ALAT), Alkaline Phosphatase (AF), Gamma-Glutamyl Transferase (GGT), Bilirubin (BR), International Normalized Ratio (INR)

The liver function of a patient can be characterized by levels of Aspartate Aminotransferase (ASAT), Alanine Aminotransferase (ALAT), Alkaline Phosphatase (AF), Gamma-Glutamyl Transferase (GGT), Bilirubin (BR) and International Normalized Ratio (INR). ASAT, ALAT, AF and GGT are enzymes that are synthesized in the liver and released into the bloodstream when liver cells are damaged. In cases of pancreatic cancer, increased levels of these markers can indicate the spread of cancer and may serve as indicators for monitoring liver functionality. Bilirubin is a waste product resulting from the breakdown of red blood cells in the liver. It typically increases when bile flow is obstructed, which may occur when a tumor is pressing against bile ducts. This obstruction can also result in elevated levels of ASAT, ALAT, AF and GGT. To prevent or resolve narrowing of the bile ducts from the tumor, a stent may be inserted to maintain their patency. Additionally, chemotherapy may cause toxicity affecting the liver, leading to elevated levels of ASAT,ALAT, AF and GGT. Conversely, when the tumor shrinks, the pressure on the bile ducts reduces and consequently bilirubin levels may decrease.

In addition, the international normalized ratio (INR) is a measure of blood clotting time that is frequently used to monitor liver function, as the liver synthesizes clotting factors. Therefore, increased INR values may indicate diminished liver functionality, as the liver is unable to produce adequate clotting factors. However, due to the lack of observations (n=19), INR will not be taken into account in further analyses.

2.1.4.5 Nutritional Status: Albumin (Alb), Sodium (Na), Potassium (K)

The nutritional status of a patient can be determined using the Albumin (Alb), Sodium (Na) and Potassium (K) levels. Albumin is an important protein synthesized in the liver that plays a crucial role in maintaining fluid balance in the body. Chemotherapy is known to cause nausea in patients, which often results in poor appetite and subsequent weight loss. This effect is compounded by the high energy demands of the tumor as well as the chemotherapy treatment itself. Reduced levels of albumin in the bloodstream can serve as an indicator of malnutrition, liver disease or other underlying conditions. Conversely, increased levels of albumin might be a response to dehydration, which may result from vomiting or diarrhea. Furthermore, sodium and potassium are electrolytes that can also be impacted by vomiting and/or diarrhea associated with side effects of chemotherapy.

Low levels of sodium and potassium in the bloodstream may also signal malnutrition or dehydration, among other possible conditions. Taken together, these factors can provide insights into the overall health status of the patient.

2.1.4.6 Inflammation: C-reactive protein (CRP), Leukocytes (LK), Systemic Inflammation Index (SII), Neutrophil-to-Lymphocyte Ratio (NLR), Platelet-to-Lymphocyte Ratio (PLR)

The variables categorized under inflammation include C-reactive protein (CRP), Leukocytes (LK), the Systemic Inflammation Index (SII), Neutrophil-to-Lymphocyte Ratio (NLR) and Platelet-to-Lymphocyte Ratio (PLR). Leukocytes play a crucial role in fighting inflammation and infection in the body (as explained earlier in the (white) blood cell group). Besides that, elevated leukocyte levels may also signify inflammation caused by pancreatic cancer or by drugs such as G-CSF given to prevent leukopenia. On the other hand, CRP, a protein produced by the liver in response to inflammation, can also increase due to the cancer or a chemotherapy-induced weakened immune system. In order to not only look at individual inflammation values, we also consider the ratios between various white blood cells, namely the SII, NLR and PLR to be specific. SII is calculated as follows,

$$SII = \frac{N \times P}{L} \tag{2.1}$$

where N is the absolute neutrophil count, P the platelet (thrombocyte) count and L the absolute lymphocyte count. Elevated levels of SII can indicate a state of chronic inflammation and immune dysfunction. NLR and PLR compare the levels of neutrophils and platelets to lymphocytes, respectively. Similar to SII, these can indicate chronic inflammation and immune dysfunction. High levels of these markers are associated with worse outcomes in pancreatic cancer patients [52] [53]. Finally, chemotherapy often weakens the patient's immune system, making them more susceptible to infections.

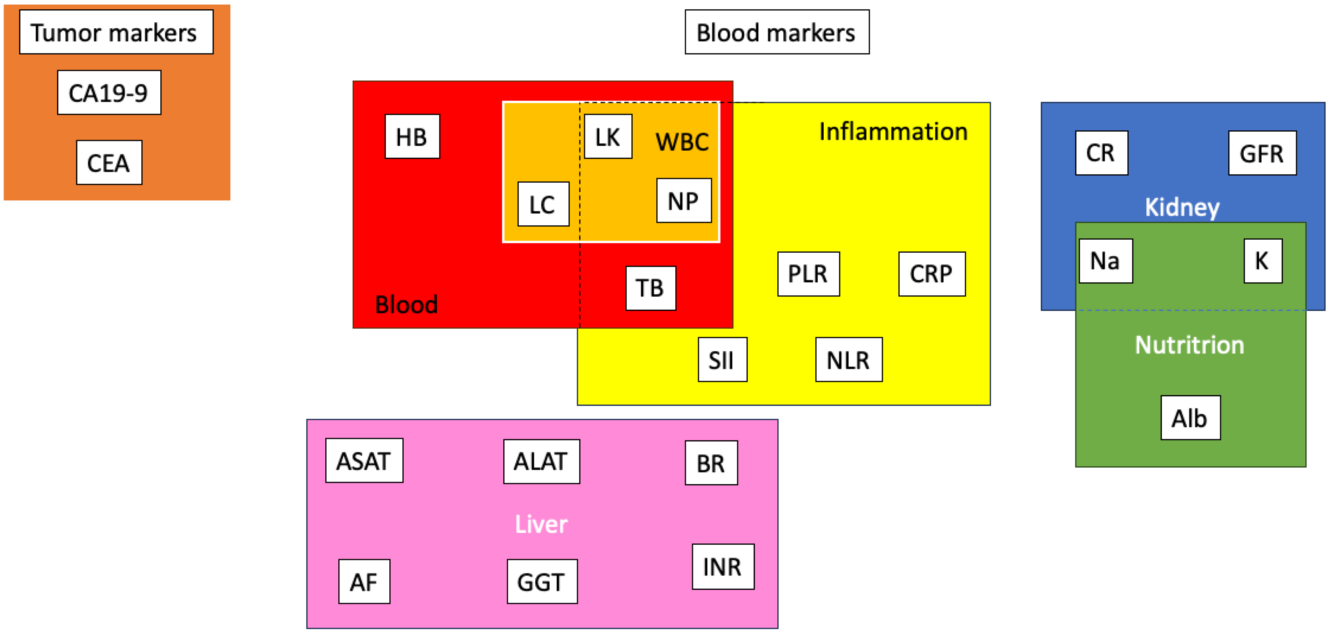


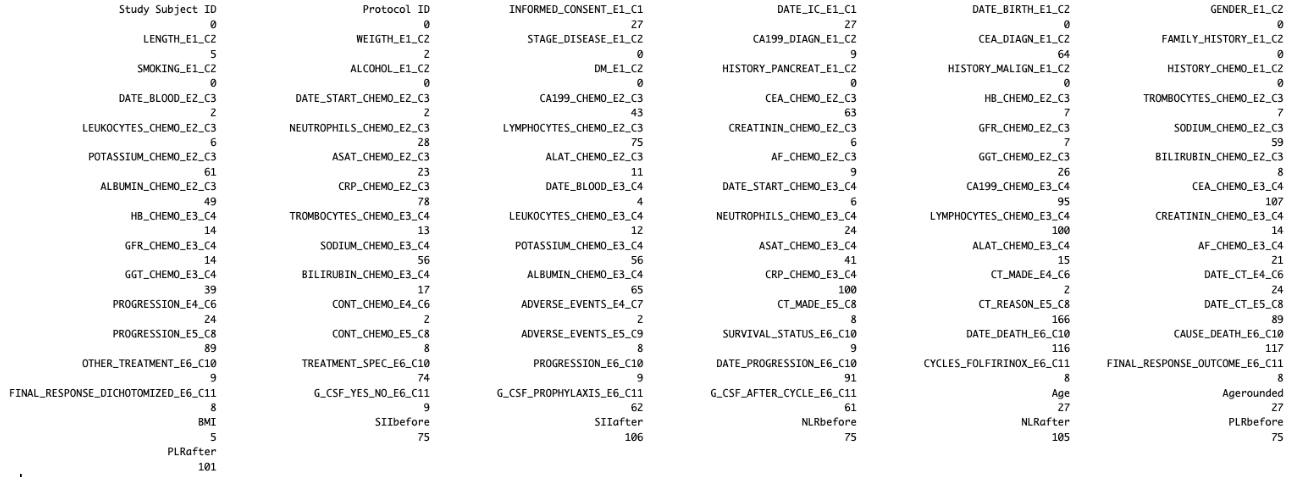
Figure 2.4: Overview of the groupings of the variables considered in the analyses.

3 | Exploratory Data Analysis

3.1. Data Analysis

The patient data used in this study comprises individuals selected from two multicenter, prospective trials conducted in the Netherlands. All four disease stages are present in the data. The first trial, registered as PREOPANC-2 in the Dutch trial register (NL7094), employed a randomized clinical design comparing neoadjuvant FOLFIRINOX with neoadjuvant gemcitabine-based chemoradiotherapy in resectable and borderline resectable patients. Blood was obtained at several moments for biomarker investigation in both groups. However, only the blood measurements of the patients who have undergone the FOLFIRINOX treatment have been used in this study. The second trial, known as iKnowIT in the Dutch trial register (NL7522), focused on the predictive value of circulating biomarkers and included patients with locally advanced or metastatic PDAC. From all the measured blood samples, the measurements were taken before and after the first FOLFIRINOX cycle. Additionally, tumor marker measurements of CA19-9 and CEA are also present in the data. The aforementioned trials received ethical approval from the respective ethics committees of all participating hospitals contributing patients to this thesis. These hospitals include: Erasmus Medical Center Rotterdam (MEC-2018-087 and MEC-2018-004), Amsterdam UMC, location Academic Medical Center (2018_196 and 2018_138), Leiden University Medical Center (L18.070 and L18.053), Isala hospital Zwolle (180606), and Medisch Spectrum Twente Enschede (H18-81). All patients provided written informed consent, and both studies were conducted in adherence to the principles outlined in the declaration of Helsinki. In order to get a good understanding on the provided dataset a thorough data analysis has been conducted. The full data analysis can be found in Appendix B.1. This section is a summary of the most important findings. The explanation as well as the units of each measured variable in the dataset can be found in Tables 2, 3 and 4.

To begin, the dataset comprises of 249 patients spanning the years 2018-2020, with one person from 2015. Two patients, have been excluded from the analysis due to their lack of data. They only had reported Date of IC. Hence, the total number of patients in the dataset that can be used is 247, with 137 male and 110 female participants. In addition, there were eight patients with missing final response, who will be reported in the Not Available (NA) group. To ensure the selection of the most relevant data, columns with a low number of observations were removed from the dataset, i.e. less than twenty. Despite this step, missing values still exist in the remaining dataset. A summary of the missing data for each variable is provided in Figure 3.1. In addition, the data analysis provided later in this section will exclude missing values depending on the selected tumor or blood marker considered. The final outcome, referring to the final response after chemotherapy treatment using FOLFIRINOX, categorizes the patients into five different categories given in Table 3.1.



```
                  Study Subject ID               Protocol ID    INFORMED_CONSENT_E1_C1             DATE_IC_E1_C1          DATE_BIRTH_E1_C2                   GENDER_E1_C2
                                 0                         0                        27                        27                         0                              0
                      LENGTH_E1_C2              WEIGTH_E1_C2       STAGE_DISEASE_E1_C2         CA199_DIAGN_E1_C2           CEA_DIAGN_E1_C2           FAMILY_HISTORY_E1_C2
                                 5                         2                         0                         9                        64                              0
                     SMOKING_E1_C2             ALCOHOL_E1_C2                  DM_E1_C2    HISTORY_PANCREAT_E1_C2      HISTORY_MALIGN_E1_C2            HISTORY_CHEMO_E1_C2
                                 0                         0                         0                         0                         0                              0
                  DATE_BLOOD_E2_C3    DATE_START_CHEMO_E2_C3         CA199_CHEMO_E2_C3           CEA_CHEMO_E2_C3            HB_CHEMO_E2_C3        TROMBOCYTES_CHEMO_E2_C3
                                 2                         2                        43                        63                         7                              7
            LEUKOCYTES_CHEMO_E2_C3   NEUTROPHILS_CHEMO_E2_C3   LYMPHOCYTES_CHEMO_E2_C3     CREATININ_CHEMO_E2_C3           GFR_CHEMO_E2_C3             SODIUM_CHEMO_E2_C3
                                 6                        28                        75                         6                         7                             59
             POTASSIUM_CHEMO_E2_C3          ASAT_CHEMO_E2_C3          ALAT_CHEMO_E2_C3            AF_CHEMO_E2_C3           GGT_CHEMO_E2_C3          BILIRUBIN_CHEMO_E2_C3
                                61                        23                        11                         9                        26                              8
               ALBUMIN_CHEMO_E2_C3           CRP_CHEMO_E2_C3          DATE_BLOOD_E3_C4    DATE_START_CHEMO_E3_C4         CA199_CHEMO_E3_C4                CEA_CHEMO_E3_C4
                                49                        78                         4                         6                        95                            107
                    HB_CHEMO_E3_C4   TROMBOCYTES_CHEMO_E3_C4    LEUKOCYTES_CHEMO_E3_C4   NEUTROPHILS_CHEMO_E3_C4   LYMPHOCYTES_CHEMO_E3_C4          CREATININ_CHEMO_E3_C4
                                14                        13                        12                        24                       100                             14
                   GFR_CHEMO_E3_C4        SODIUM_CHEMO_E3_C4     POTASSIUM_CHEMO_E3_C4          ASAT_CHEMO_E3_C4          ALAT_CHEMO_E3_C4                 AF_CHEMO_E3_C4
                                14                        56                        56                        41                        15                             21
                   GGT_CHEMO_E3_C4     BILIRUBIN_CHEMO_E3_C4       ALBUMIN_CHEMO_E3_C4           CRP_CHEMO_E3_C4             CT_MADE_E4_C6                  DATE_CT_E4_C6
                                39                        17                        65                       100                         2                             24
                 PROGRESSION_E4_C6          CONT_CHEMO_E4_C6      ADVERSE_EVENTS_E4_C7             CT_MADE_E5_C8           CT_REASON_E5_C8                  DATE_CT_E5_C8
                                24                         2                         2                         8                       166                             89
                 PROGRESSION_E5_C8          CONT_CHEMO_E5_C8      ADVERSE_EVENTS_E5_C9    SURVIVAL_STATUS_E6_C10         DATE_DEATH_E6_C10             CAUSE_DEATH_E6_C10
                                89                         8                         8                         9                       116                            117
            OTHER_TREATMENT_E6_C10     TREATMENT_SPEC_E6_C10        PROGRESSION_E6_C10   DATE_PROGRESSION_E6_C10  CYCLES_FOLFIRINOX_E6_C11  FINAL_RESPONSE_OUTCOME_E6_C11
                                 9                        74                         9                        91                         8                              8
FINAL_RESPONSE_DICHOTOMIZED_E6_C11       G_CSF_YES_NO_E6_C11  G_CSF_PROPHYLAXIS_E6_C11  G_CSF_AFTER_CYCLE_E6_C11                       Age                     Agerounded
                                 8                         9                        62                        61                        27                             27
                               BMI                 SIIbefore                  SIIafter                 NLRbefore                  NLRafter                      PLRbefore
                                 5                        75                       106                        75                       105                             75
                          PLRafter
                               101
```

Figure 3.1: Overview of the variables in the PREOPANC2-iKnowIT dataset and the number of missing values for each variable, n=247.

In the dataset provided, only 3 (1.3%) patients showed complete response to the chemotherapy, 48 (20%) showed

Number	Final reponse	Explanation
0	Complete Response (CR)	The cancer can completely disappear for a time after treatment
1	Partial Response (PR)	The cancer shrinks at least by a third after treatment.
2	Progressive Disease (PD)	The cancer starts to grow again
3	Stable Disease (SD)	The cancer neither shrinks nor grows after treatment
4	Unknown (Un)	-

Table 3.1: Final response to chemotherapy.

partial response, 43 (18%) progressive disease, 122 (51%) stable disease and 23 (9.6%) were unknown. Note that there was a separate number within the final response variable with the label ‘unknown’. These 23 patients are subsequently also removed from the analysis later as the final response is unknown, e.g. because no evaluation scan was performed or available for tumor measurements. Therefore, the total number of unknowns in the entire dataset would be 23+8=31 with missing final response. In addition to that, the final response can also be dichotomized. These two categories are defined as given in Table 3.2.

Number	Dichotomous Final response	Contains
0	Disease Control (DC)	CR, PR, SD
1	Progressive Disease (PD)	PD

Table 3.2: Dichotomized final response to chemotherapy.

In these categories, 173 (72%) patients were classified as DC and 43 (18%) as PD. The data analysis starts with analyzing the patient characteristics, followed by the tumor markers and finally the blood markers grouped by final response to chemotherapy. The most significant findings are summarized in the next subsection.

3.1.1 Overview of significant observations in the dataset

This subsection presents an overview of the significant findings related to patient characteristics, tumor markers and blood markers in the provided dataset. The full analysis can be found in Appendix B.1 containing all the histograms, boxplots, scatterplots, QQ-plots and more.

3.1.1.1 Patient characteristics

Patient characteristics consist of the demographic and clinical characteristics of the patients examined to assess their potential influence on the final response to chemotherapy based on the clinical data provided. The gender distribution in the study cohort was found to be approximately 55% male and 45% female. To assess the impact of gender on the final response to chemotherapy, a stacked bar chart was generated (see Figure B.1). Surprisingly, the analysis indicated that gender did not exert a significant influence on the outcome of chemotherapy treatment. Furthermore, when examining the age range of the patients, it was found to span from 47 (minimum) to 82 (maximum) years, with an average age of 64 years. In order to investigate whether age played a role in the final response to chemotherapy, data visualization was performed, see Figure B.2. Interestingly, the findings suggest that age did not exert a significant effect on the outcome, as there were no notable variations observed across the age range studied.

Assessing the weight distribution among the patients showed that the average weight was 78 kg, ranging from 42 to 124 kg. The distribution of the weight values exhibited a reasonably normal pattern with a slight left skew. Nevertheless, upon analyzing the plot in Figure B.3 no clear association between weight and final response to chemotherapy could be identified. Next, after exploring the height (in the file named ‘Length’) data, it was observed that the patients’ height ranged from 152 to 200cm, with an average height of 174cm. After analysis of the plot in Figure B.4, no significant disparities were observed among the groups in terms of the final response to chemotherapy. This suggests that height is unlikely to be a determining factor in treatment outcomes.

Furthermore, the body mass index (BMI) is calculated for the patients using the equation in Equation (B.2). The values ranged from 15.9 (underweight) to 37.1 (obese) with a mean BMI of 25.5 (slightly overweight). While no distinct pattern linking BMI to chemotherapy response emerged from the data, it is worth noting that patients with higher BMI may have an increased risk of developing PDAC, possibly associated with unhealthy lifestyles. Continuing the analysis of the stage of disease prior to chemotherapy, it revealed that all complete responders (n=3) had resectable pancreatic cancer, as illustrated in Figure B.6. Nevertheless, this finding, despite the small amount of patients, does support the notion that patients with (borderline) resectable PDAC generally exhibit a more favorable disease outcome after chemotherapy.

Moreover, after examination of the influence of a positive history of family history of cancer on the final response to chemotherapy, given in Figure B.7, it was found to have no significant effect. However, a positive family history may serve as an indicator of an individual's susceptibility to developing pancreatic cancer. Smoking is another influential factor that may enhance the probability of developing cancer. From the data analysis, it was seen that approximately half of the patients in the data set were smokers. Nonetheless, the bar chart analysis in Figure B.8 did not show a clear trend linking smoking to the response to chemotherapy. However, it is important to note that smoking can adversely impact other blood markers and a patient's overall health.

Regarding alcohol use, the chart in Figure B.9 displayed that patients who had a complete response to chemotherapy did use alcohol. Though, it remains uncertain whether these patients were occasional drinkers or heavy drinkers. Also, due to the small number of complete responders (n=3), strict conclusions cannot be drawn. In contrast, patients who had stopped drinking tended to exhibit PR or SD as the final treatment outcome. In addition, the presence of diabetes mellitus and its potential influence on the final response to chemotherapy is examined using Figure B.10. However, no significant impact was identified.

After analyzing the history of pancreatitis among the patients in the cohort using Figure B.11, it was observed that a positive history of pancreatitis tended to lead to SD as the final response to chemotherapy. Nonetheless, it is crucial to mention that the dataset only included a limited number of eight patients with a positive of pancreatitis, so clear conclusions cannot be drawn. Furthermore, only 37 patients in the dataset had a positive history of malignancy. Among them, the figure in Figure B.12 shows that all the four final responses have been observed. Hence, no definite conclusions can be drawn from the available data. Lastly, after examining the impact of prior chemotherapy treatment on the final response to the current FOLFIRINOX chemotherapy cycle, it was observed that of the 13 patients who had a positive history of chemotherapy, most of these exhibited either PR or SD. This might be a consequence of a potential increased sensitivity to chemotherapy resulting from previous treatment.

3.1.1.2 Tumor markers

Next, the tumor markers Carbohydrate antigen 19-9 (CA19-9) and Carcinoembryonic antigen (CEA) have been analyzed. In the provided data, the levels of CA19-9 in almost all patients exceeded the healthy threshold of 37 kU/L, based on the advice of medical experts from the Erasmus Medical Center Rotterdam. Surprisingly, the analysis demonstrated that chemotherapy did not have a significant impact on CA19-9 values. However, the biggest difference in values before and after the first chemotherapy cycle, as well as the biggest CA19-9 values, are observed in the patients with PD as final response. Moreover, no discernible gender-based differences were found in CA19-9 levels. Similar to CA19-9, most patients exhibited elevated levels of CEA above the healthy reference range. However, due to the presence of a high outlier value of 2822 in our study, the mean CEA level increased after chemotherapy. In the absence of this outlier, no significant difference was observed in CEA values before and after chemotherapy or between diagnosis and pre-chemotherapy levels. Furthermore, the plot depicted in Figure B.41 did not indicate a clear association between smoking and CEA levels. Nevertheless, current smokers showed a higher mean CEA level compared to non-smokers. The analysis of the relationship between CA19-9 and CEA levels did not reveal a consistent pattern between CA19-9 and CEA levels. Most patients with low CA19-9 levels exhibited correspondingly low CEA levels, while high CA19-9 levels could be associated with low CEA levels before and after chemotherapy. However, a number of outlier values were observed where high levels of both CA19-9 and CEA were present.

3.1.1.3 Blood markers

To continue, the measured blood markers values before and after the first chemotherapy cycle are analyzed. Firstly, the hemoglobin (HB) values exhibited a slight decrease after chemotherapy. Before and after chemotherapy, these values followed a normal distribution, as can be seen from the QQ-plot and performed Shapiro-Wilk test in Figure B.51 and Table B.7. There was no significant difference observed between the final response groups. As expected, females tended to have lower HB values compared to males. In addition, the thrombocytes (platelets) values decreased on average after chemotherapy, with the smallest differences observed in the PR group (they were mostly within the healthy range too). However, there was no significant difference between the TB values of the PD and SD groups. Next, the leukocyte (white blood cell) count significantly increased after chemotherapy surpassing the healthy range. Unlike TB and HB, there was no significant difference observed across the final response groups. Similar to the LK count, the neutrophil count (NP) dramatically increased after chemotherapy for most patients, with no clear distinguishing pattern among the final response groups. On the contrary, the lymphocyte (LC) values were not significantly affected by chemotherapy, as most of them remained within the healthy range before and after treatment.

Incorporating leukocytes, neutrophils, and lymphocytes into a single graph may provide more informative insights than analyzing them separately. Figure 3.2 depicts the values of these variables before the first chemotherapy treatment, while Figure 3.3 displays the values after it. The influence of neutrophils and lymphocytes on leukocyte count per patient can be clearly observed from these plots. Furthermore, note that the scale on the white blood cell variables after treatment is more than twice as large as before treatment. After the first chemotherapy treatment, the neutrophil count appears to dominate the entire leukocyte count, with little contribution from lymphocytes. This suggests that neutrophils are more susceptible to chemotherapy treatment than lymphocytes, although medication given to patients during treatment may also be a contributing factor. However, no clear trend can be inferred between the final response groups and the leukocyte, neutrophil and/or lymphocyte counts.

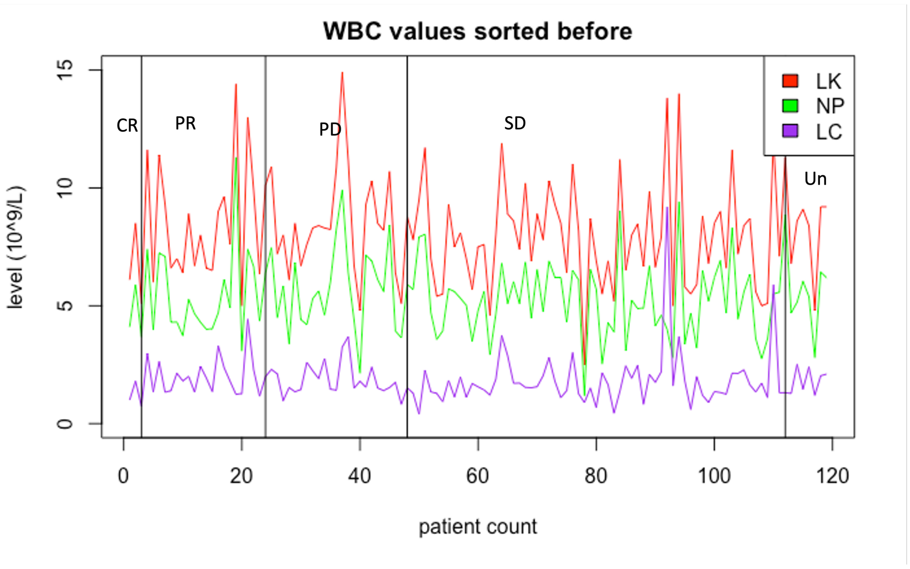


Figure 3.2: White blood cell counts of Leukocytes (LK), Neutrophils (NP) and Lymphocytes (NP) before the first chemotherapy cycle sorted based on their final response in 10^9/L blood, n = 119 (Complete response (CR, n=2), Partial response (PR, n=21), Progressive Disease (PD, n=24), Stable Disease (SD, n=64) and Unknown (Un, n=8).

Furthermore, creatinin values before and after the first chemotherapy cycle seem to follow a normal distribution. However, due to the presence of a few outliers, which can be clearly seen in Figure B.82, the Shapiro-Wilk test rejects the normal distribution. The outliers are likely caused by a couple of patients who were heavily affected by chemotherapy treatment. Overall, chemotherapy did not have a significant impact on the CR values and no clear differences were seen among the final response groups. However, it is worth noting that males tended to have higher CR values compared to females, potentially due to their higher muscle mass. Likewise, the glomerular filtration rate (GFR) was not significantly affected by chemotherapy, with a few exceptions again observed in the individuals who exhibited a strong reaction to treatment. The majority of GFR values fell below the healthy range and the smallest difference before and after chemotherapy was observed in the PR group.

Subsequently, sodium (Na) did not show a significant change before and after the first chemotherapy cycle and

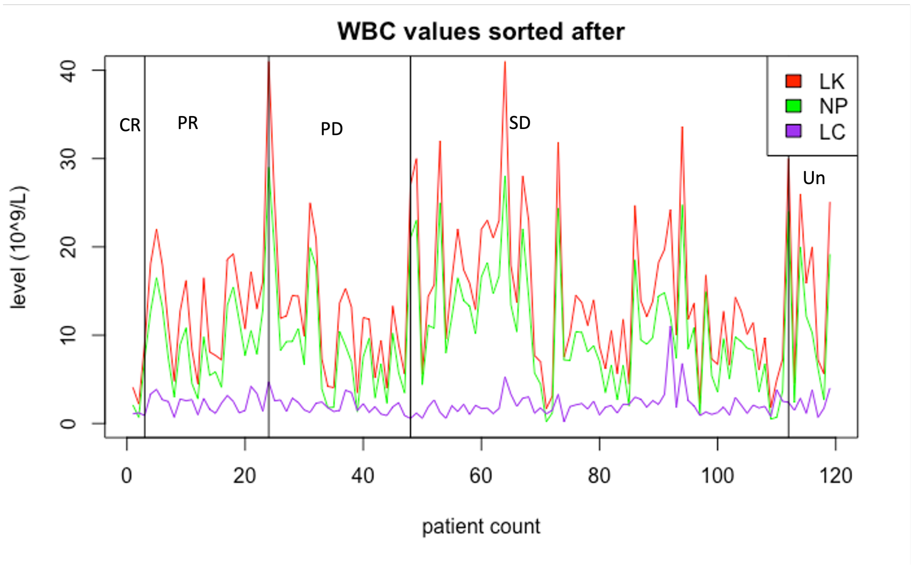


Figure 3.3: White blood cell counts of Leukocytes (LK), Neutrophils (NP) and Lymphocytes (NP) after the first chemotherapy cycle sorted based on their final response in 10^9/L blood, n = 119 (Complete response (CR, n=2), Partial response (PR, n=21), Progressive Disease (PD, n=24), Stable Disease (SD, n=64) and Unknown (Un, n=8).

no clear distinction was observed across the final response groups. Most Na values remained within the healthy range both before and after treatment. However, this might be a consequence of medication given to the patients. Contrasting, potassium (K) values exhibited a slight decrease on average after chemotherapy, but remained within the healthy range for the majority of patients, with only a few falling below the healthy boundary. Also, the potassium values were normally distributed as can be seen from Figure B.103. However, there was no clear difference observed across the final response groups.

Continuing the analysis on the liver variables showed that Aspartate Aminotransferase (ASAT) values slightly decreased in general, especially for the patients with initially high ASAT values before chemotherapy. However, almost all values remained above the healthy range, and the biggest spikes were seen in the patients who exhibited SD as final response. Nevertheless, there was no significant difference observed across the final response groups. Similar to ASAT, Alanine Aminotransferase (ALAT) values also exhibited a slight decrease after chemotherapy, particularly for patients with elevated ALAT levels prior to treatment. Yet, there was no notable difference observed across the final response groups, nor between males and females. In contrast to ASAT and ALAT, Alkaline Phosphatase (AF) values showed a slight increase after chemotherapy, with the largest difference before and after treatment observed in the PD group due to a few outliers. Gamma-glutamyl Transferase (GGT) values generally decreased after chemotherapy, with very few values falling within the healthy range despite the decrease. No significant difference was found between the final response groups, nor between males and females, or across alcohol usage.

Similar to the white blood cells, it may provide valuable insights to depict the ASAT, ALAT, and AF values collectively in a single graph. This analysis was conducted and the results are presented in Figure 3.4 and Figure 3.5. The graphical representations indicate a significant correlation between ALAT and ASAT values. It is evident that patients generally exhibit comparable ALAT and ASAT levels both before and after undergoing chemotherapy. Moreover, the AF values tend to be higher in comparison to the ALAT and ASAT values across all patients. Notably, among patients with elevated levels of ASAT and ALAT, the AF value tends to be higher as well, with the exception of one particular patient who demonstrates exceptionally high pre-chemotherapy ALAT and ASAT values but a comparatively low AF value.

The analysis of the Bilirubin (BR) values shows that these values significantly decreased after the first chemotherapy cycle, transitioning from initially high, unhealthy levels to values within the healthy range for almost all patients. The smallest difference before and after chemotherapy was observed in the PR final response group. Albumin (Alb) levels slightly decreased after chemotherapy for the majority of patients with a few exceptions. Most values fell within the healthy range before the first chemotherapy treatment, but dropped below the healthy range afterwards. Nonetheless, there was no significant pattern observed across the final response groups. On the contrary, C-reactive protein (CRP) levels increased significantly for most patients following chemotherapy, with almost all values exceeding the healthy range both before and after treatment. The largest changes were observed

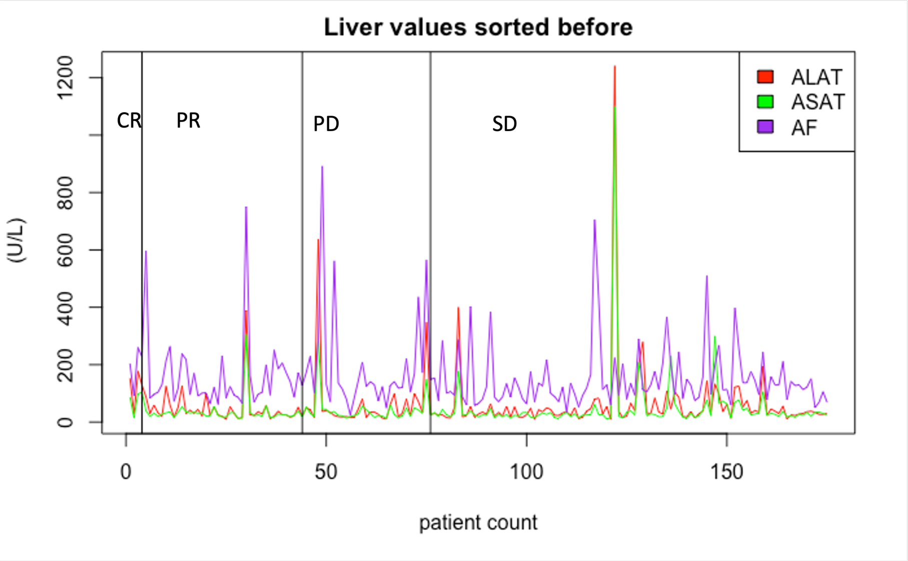


Figure 3.4: ALAT, ASAT and AF values before the first chemotherapy cycle sorted based on their final response in U/L blood, n = 175 (Complete response (CR, n=3), Partial response (PR, n=40), Progressive Disease (PD, n=32) and Stable Disease (SD, n=100).

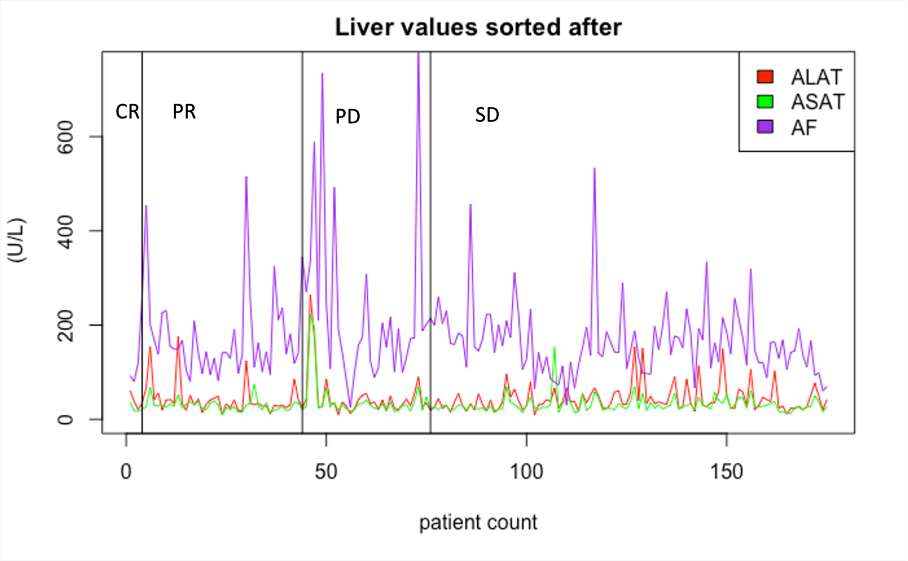


Figure 3.5: ALAT, ASAT and AF values after the first chemotherapy cycle sorted based on their final response in U/L blood, n = 175 (Complete response (CR, n=3), Partial response (PR, n=40), Progressive Disease (PD, n=32) and Stable Disease (SD, n=100).

in the PD and SD groups. Unfortunately, due to the limited amount of available data (n=19) for the international normalized ratio (INR), no meaningful conclusions can be drawn.

Finally, the inflammation indices SII, NLR and PLR are examined. SII, the Systemic Inflammation Index, showed a increasing trend in general after chemotherapy for most patients, shifting from values within the healthy range to values above the healthy boundary of 900, set according to medical experts. From the analysis, it can be seen that the smallest difference before and after the first chemotherapy cycle was found in the PR group. Likewise, the NLR increased in general after chemotherapy, with the most substantial increases observed in the PD and SD groups. However, no significant difference was observed between males and females. In contrast to the NLR, PLR generally decreased after chemotherapy, transitioning from initially high, unhealthy values to lower values, with most falling within the healthy range. Nevertheless, no differences were observed between males and females, nor across the final response groups. Overall, these findings highlight the varied effects of chemotherapy on different blood markers, indicating the complex interplay between treatment and physiological responses. A summary of the most important findings is provided in Table 3.6. Further investigation is required to fully understand the underlying mechanisms and potential implications of these observations.

3.1.2 Overview tables data analysis

Categorical Variables	**Full data (n=247)**	**No missing values (n=216)**	**No outliers based on GFR and BR (n=203)**
Gender			
Male	137	122	113
Female	110	94	90
Stage of disease			
(borderline) Resectable	152	137	130
Locally advanced	54	48	43
Metastatic	41	31	30
Family history			
No	147	131	126
Yes	23	20	18
Unknown	77	65	59
Smoking			
Never	108	97	91
Former	74	64	61
Current	50	43	39
Unknown	15	12	12
Alcohol			
No	78	71	68
Yes	113	98	90
Stopped	38	33	31
Unknown	18	14	14
Diabetes mellitus			
No	189	164	156
Yes	58	52	47
History of pancreatitis			
No	239	208	195
Yes	8	8	8
History of malignancy			
No	210	182	174
Yes	37	34	29
History of chemotherapy			
No	234	204	191
Yes	13	12	12
Cycles of Folfirinox			
0	3	-	-
1	19	6	6
2	13	10	10
3	8	7	7
4	32	32	29
5	5	5	5
6	5	5	4
7	6	6	5
8	128	126	118
9	2	2	2
10	1	1	1
11	2	2	2
12	15	14	14
Unknown	8	-	-
GCSF received			
No	52	44	41
Yes	186	171	162
Unknown	9	1	-
GCSF prophylaxis			
No	23	23	20
Yes	162	147	141
Unknown	62	46	42

Table 3.3: Overview patient characteristics categorical variables

Variables	**Disease Control (n=173)**	**Progressive Disease (n=43)**	**p-value**
Age (years), mean (SD)	64.8 (8.3)	63.5 (8.8)	0.431
BMI, mean (SD)	25.5 (3.8)	25.2 (4.3)	0.453
Length (cm), mean (SD)	174.5 (10.3)	175.2 (9.6)	0.787
Weight (kg), mean (SD)	78.0 (15.0)	77.1 (12.8)	0.826
Gender			0.270*
Male	94	28	
Female	79	15	
Stage of Disease			0.148
(borderline) Resectable	114	23	
Locally advanced	38	10	
Metastatic	21	10	
Family history			0.516
No	103	28	
Yes	15	5	
Unknown	55	10	
Smoking			0.814
Never	78	19	
Former	50	14	
Current	34	9	
Unknown	11	1	
Alcohol			0.552
No	55	16	
Yes	80	18	
Stopped	25	8	
Unknown	13	1	
Diabetes Mellitus			1
No	131	33	
Yes	42	10	
History of pancreatitis			1
No	166	42	
Yes	7	1	
History of malignancy			0.819
No	145	37	
Yes	28	6	
History of chemotherapy			1
No	163	41	
Yes	10	2	
Cycles of Folfirinox			0.0005
1	2	4	
2	6	4	
3	3	4	
4	14	18	
5	3	2	
6	3	2	
7	6	0	
8	120	6	
9	2	0	
10	0	1	
11	1	1	
12	13	1	
GCSF received			0.518
No	33	11	
Yes	139	32	
Unknown	1	0	
GCSF prophylaxis			0.447
No	20	3	
Yes	119	28	
Unknown	34	12	

Table 3.4: Overview categorical variables and p-values for dataset with no missing values split into the dichotomized final response. Continuous variables were calculated using the Mann-Whitney U test and categorical variables using the Fisher's exact test when there was a small frequency number in the contingency matrix, otherwise the chi-square test is used, see Appendix A for more information on these tests.

Patient Characteristic	**Influence on Chemotherapy**	**Difference in Final Response Groups**
Gender	No	No
Age	No	No
Weight	No	No
Length	No	No
BMI	No	No
Stage of disease	CR were (borderline) resectable	No
Family history of PDAC	No	No
Smoking	No	No
Alcohol usage	CR were alcohol users	No
Diabetes mellitus	No	No
Pancreatitis	No	No
History of malignancy	No	No
History of chemotherapy	No	No

Table 3.5: Overview patient characteristics and whether there was a significant difference seen in final response in chemotherapy as well as whether the characteristic itself would influence chemotherapy.

Tumor or Blood Marker	n	Affected by Chemo	Difference in Final Response Groups	Normally Distributed	Male vs Female
CA19-9	131	Yes, slight increase after chemo, few outliers	Yes PD has largest values and biggest difference	No	No
CEA	103	No, slight increase after chemotherapy	No	No	NA
HB	223	Yes, slight decrease	No	Yes	Yes, females tend to have lower values
TB	224	Yes, slight decrease	No, but smallest differences in PR	No	NA
LK	226	Yes, major decrease	No	No	NA
NP	199	Yes, major decrease	No	No	NA
LC	122	No, but maybe a slight increasing trend	No	No	NA
CR	224	No, but maybe a slight increasing trend	No, but outliers in SD group	No	Yes, males have higher values
GFR	223	No	No, but SD has the largest differences	No	NA
Na	160	No	No	No	NA
K	158	Yes, slight decrease	No	Yes	NA
ASAT	192	No, but few outliers decreased after chemo	Yes, biggest differences in SD group	No	NA
ALAT	218	No, but few outliers decreased after chemo	No	No	No
AF	214	No	Yes, largest differences in PD	No	NA
GGT	191	No	Yes, PR and SD have more positive differences	No	No
BR	219	Yes, major decrease after chemo	No	No	NA
Alb	161	Yes, slight decrease after chemo	No	No, but very close	NA
CRP	127	Yes, increased after chemo	No	No	NA
INR	19	No (too few data)	No	No	NA
SII	118	Yes, slight increase in PD and SD	Yes, largest differences and values in PD and SD	No	NA
NLR	119	Yes, increase after chemo	Yes, largest increase in PD and SD groups	No	No
PLR	121	Yes, decrease after chemo	No	No	No

Table 3.6: Summary of the key findings from the data analysis conducted in Section B.1 on tumor and blood markers. 'n' is number of observations for each marker after excluding missing values, 'Affected by Chemo' indicates whether a significant change was observed in the marker value after the initial chemotherapy cycle, 'Difference in Final Response Groups' denotes whether there was a significant variation in values among the final response groups (CR = Complete Response, PR = Partial Response, SD = Stable Disease, PD = Progressive Disease), 'Normally Distributed' is based on the results of the KS-test and QQ-plot analysis in Section B.1, 'Male vs Female' indicates the presence of a significant difference in marker values between males and females (NA = not applicable when such analysis has not been performed).

Variable before chemotherapy	**Disease Control (n=173)**	**Progressive Disease (n=43)**	**p-value**
Tumor markers, median (IQR)			
CA19-9 (kU/L)	136.1 (41.0-403.0)	687.5 (99.5-1970.8)	0.002
CEA ($\mu g/L$)	3.4 (2.1-7.2)	5.1 (2.9-12.4)	0.119
Blood markers, median (IQR)			
HB ($mmol/L$)	7.9 (7.3-8.6)	7.8 (6.8-8.1)	0.032
TB ($10^9/L$)	278.0 (226.2-325.0)	253.0 (220.2-335.2)	0.562
LK ($10^9/L$)	7.9 (6.6-9.2)	8.1 (6.8-9.6)	0.659
NP ($10^9/L$)	8.1 (3.4-13.0)	7.7 (3.7-10.8)	0.206
LC ($10^9/L$)	1.7 (1.3-2.1)	1.8 (1.5-2.2)	0.312
CR ($\mu mol/L$)	69.0 (58.0-79.0)	69.0 (58.5-81.5)	0.632
GFR (mL/min)	90.0 (81.3-90.0)	90.0 (82.0-93.0)	0.352
Na ($mmol/L$)	139.0 (137.0-141.0)	140.0 (137.0-141.0)	0.515
K ($mmol/L$)	4.2 (3.9-4.4)	4.3 (4.1-4.7)	0.195
ASAT (U/L)	27.0 (20.0-36.0)	31.0 (21.0-46.0)	0.590
ALAT (U/L)	35.0 (24.0-57.0)	32.0 (19.0-50.0)	0.391
AF (U/L)	123.0 (90.5-179.0)	130.0 (102.0-187.5)	0.408
GGT (U/L)	72.0 (37.0-140.5)	128.0 (62.0-171.0)	0.083
BR ($\mu mol/L$)	14.0 (7.0-25.0)	17.0 (9.0-24.5)	0.278
Alb (g/L)	40.0 (36.0-43.0)	39.5 (35.3-41.8)	0.299
CRP (mg/L)	4.3 (1.5-9.5)	9.7 (2.7-15.5)	0.133
SII	764.9 (567.0-1131.9)	937.4 (520.5-1249.2)	0.533
NLR	3.0 (2.1-4.3)	3.1 (2.2-4.6)	0.723
PLR	163.5 (122.0-215.9)	148.4 (111.9-252.6)	0.663

Table 3.7: Overview measured tumor and blood variables measured before the first chemotherapy cycle and p-values for dataset with no missing values split into the dichotomised final response: disease control and progressive disease groups. P-values are determined using the Wilcoxon Ranksum test.

Variable after chemotherapy	**Disease Control (n=173)**	**Progressive Disease (n=43)**	**p-value**
Tumor markers, median (IQR)			
CA19-9 (kU/L)	95.0 (33.9-448.3)	1087.0 (158.3-2047.0)	0.001
CEA ($\mu g/L$)	3.7 (2.2-7.2)	4.8 (2.9-12.1)	0.268
Blood markers, median (IQR)			
HB ($mmol/L$)	7.6 (7.0-8.1)	7.4 (7.1-7.9)	0.430
TB ($10^9/L$)	199.5 (151.2-211.9)	223.0 (172.0-291.0)	0.074
LK ($10^9/L$)	11.9 (6.4-18.0)	11.9 (7.3-15.3)	0.854
NP ($10^9/L$)	8.1 (3.4-13.0)	7.7 (3.7-10.8)	0.734
LC ($10^9/L$)	1.9 (1.2-2.6)	1.9 (1.4-2.6)	0.657
CR ($\mu mol/L$)	70.0 (59.0-81.0)	66.5 (56.3-80.8)	0.359
GFR (mL/min)	89.5 (78.0-90.0)	90.0 (82.3-95.0)	0.082
Na ($mmol/L$)	139.0 (137.2-141.0)	140.0 (138.0-141.0)	0.539
K ($mmol/L$)	4.0 (3.6-4.2)	4.0 (3.7-4.2)	0.494
ASAT (U/L)	27.0 (22.0-34.0)	28.0 (21.0-36.0)	0.529
ALAT (U/L)	34.0 (25.0-51.0)	33.0 (27.0-43.0)	0.576
AF (U/L)	146.0 (110.0-193.0)	171.0 (115.0-214.5)	0.165
GGT (U/L)	72.0 (37.0-140.5)	82.0 (53.0-156.0)	0.064
BR ($\mu mol/L$)	7.0 (5.0-11.6)	9.0 (7.0-12.0)	0.138
Alb (g/L)	37.0 (33.0-40.0)	36.0 (33.0-40.0)	0.935
CRP (mg/L)	6.0 (2.6-21.3)	11.0 (3.7-28.2)	0.403
SII	764.8 (474.7-1354.3)	973.7 (737.3-1430.9)	0.173
NLR	4.2 (2.8-6.1)	4.4 (2.3-6.9)	0.835
PLR	107.3 (69.7-149.5)	116.0 (88.0-152.7)	0.280

Table 3.8: Overview measured tumor and blood variables measured after the first chemotherapy cycle and p-values for dataset with no missing values split into the dichotomised final response: disease control and progressive disease groups. P-values are determined using the Wilcoxon Ranksum test.

Variable difference before and after the first cycle	**Disease Control (n=173)**	**Progressive Disease (n=43)**	**p-value**
Tumor markers, median (IQR)			
CA19-9 (kU/L)	0.0 (-44.0-35.0)	-43.3 (-403.5-83.5)	0.078
CEA ($\mu g/L$)	0.0 (-1.1 -0.8)	-0.4 (-0.7 - 0.8)	0.852
Blood markers, median (IQR)			
HB ($mmol/L$)	0.4 (0.0-0.8)	0.2 (-0.2-0.5)	0.041
TB ($10^9/L$)	69.0 (28.8-113.0)	33.5 (0.8-70.8)	0.013
LK ($10^9/L$)	-3.8 (-9.4-1.4)	-2.7 (-7.5-0.9)	0.905
NP ($10^9/L$)	-2.9 (-7.3-1.6)	-1.8 (-4.7-1.6)	0.491
LC ($10^9/L$)	-0.2 (-0.8 - 0.2)	-0.1 (-0.5-0.2)	0.463
CR ($\mu mol/L$)	-1.0 (-7.0 - 3.0)	0.0 (-3.0 - 3.0)	0.091
GFR (mL/min)	0.0 (-1.0 - 3.0)	0.0 (-2.0 - 1.0)	0.183
Na ($mmol/L$)	0.0 (-2.0-2.0)	0.0 (-22.0-1.3)	0.936
K ($mmol/L$)	0.3 (-0.1 - 0.6)	0.3 (-0.1 - 0.5)	0.960
ASAT (U/L)	1.0 (-7.0 - 8.8)	2.0 (-7.8 - 10.8)	0.971
ALAT (U/L)	1.0 (-12.0 - 14.0)	-1.0 (-10.0 - 8.0)	0.589
AF (U/L)	-20.0 (-59.0 - 22.0)	-12.0 (-77.5 - 21.0)	0.551
GGT (U/L)	8.5 (-13.5 - 47.8)	5.0 (-30.0 - 51.8)	0.301
BR ($\mu mol/L$)	6.0 (1.0 - 13.0)	7.0 (1.0 - 11.0)	0.985
Alb (g/L)	3.0 (0.9 - 5.0)	3.0 (-0.3 - 4.3)	0.493
CRP (mg/L)	-1.7 (-11.3 - 1.0)	-0.5 (-7.5 - 6.3)	0.258
SII	-46.6 (-386.5 - 377.9)	56.4 (-520.7 - 387.2)	0.747
NLR	-1.33 (-3.3 - 0.5)	-0.5 (-2.2 - 0.7)	0.230
PLR	62.3 (16.8 - 98.8)	23.2 (-3.1 - 53.0)	0.037

Table 3.9: Overview measured tumor and blood variables difference measured before and after the first chemotherapy cycle and p-values for dataset with no missing values split into the dichotomised final response: disease control and progressive disease groups. P-values are determined using the Wilcoxon Ranksum test.

3.2. Outlier Analysis

In Appendix B.1 with a summary given in section 3.1, we performed a thorough data analysis on the complete dataset and identified missing values for each variable. During this analysis, it was observed that many variables contain (potential) outliers. Outliers are observations with exceptionally high or low values for a particular variable and can significantly impact statistical and machine learning models (like random forests). These outliers can introduce bias in the results, resulting in less accurate and reliable models. Thus, it is essential to identify and appropriately handle outliers. In the case of random forests, removing outliers can improve the model-building process by creating a more balanced and representative dataset. Outliers can skew the distribution of the data and affect the split points chosen by the decision trees in the random forest algorithm. This can lead to overfitting, where the model performs well on the training data but poorly on new, unseen data. By removing outliers, the model can generalize better and make more accurate predictions on new data. However, it is important to keep in mind that removing outliers can also lead to loss of information and potentially biased results. Therefore, it is crucial to carefully evaluate the impact of outliers on the model and consider alternative approaches, such as transforming the data or using robust algorithms that are less sensitive to outliers. The outlier analysis performed on the provided PREOPANC2-iKnowIT dataset will be by use of the inter-quartile range (IQR) method. This section will only include the rationale behind how the outlier analysis is performed as well as a summary of the most important findings, whereas the full detailed outlier analysis can be found in Appendix B.2.

3.2.1 Outlier analysis using Inter-Quartile Range (IQR)

The Inter-quartile Range (IQR) method is a common technique used for outlier analysis in statistics. The IQR is defined as the difference between the lower quartile (25^{th} percentile) of a dataset and the upper quartile (75^{th} percentile). It provides a measure of the spread of the middle 50% of the data, which is more robust to outliers than the standard deviation. To identify outliers using the IQR method, a common rule of thumb is to consider any data point that falls more than 1.5 times the IQR below the lower quartile or above the upper quartile as an outlier. These points are considered to be extreme values that are unlikely to have occurred by chance, and may be indicative of errors in the data or genuine anomalies that require further investigation. By identifying and removing outliers, the IQR method can help improve the accuracy and reliability of statistical analyses, as well as machine learning models that rely on the input data being representative. To summarize, for outlier detection of each of the provided variables, the following steps are taken using the IQR-method:

Algorithm 1 IQR algorithm for outlier detection

```
Require: Data $D$
Find the first quartile, $Q1$
Find the third quartile, $Q3$
Calculate the IQR:
    $IQR = Q3 - Q1$
Define the normal data range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$
Remove any data point outside this range and consider it as an outlier.
```

In the subsequent subsections, the outliers for the tumor markers and blood markers measured before and after the initial chemotherapy cycle, will be determined using the IQR-method. For each variable, tables will be presented displaying the Patient ID and corresponding outlier value. Additionally, boxplots and histograms will be used as graphical representations of the values for each variable both before and after outlier removal. Lastly, a summary table will be provided detailing the count of outliers for each variable, as well as whether the variable exhibits normal distribution both before and after outlier removal, and any noteworthy observations. It is important to note that the IQR method has no medical or biological basis for defining which values should be considered outliers. As such, the interpretation of the results should be considered with caution and in conjunction with medical experts. In the random forest model later on, outliers will be based on medical advice of GFR$<$ 30mL/min and BR$>$ 50μmol/L, as outliers may carry important information or insights into the data, and their removal may lead to a loss of valuable information. Instead, this outlier analysis should be considered as a complementary tool to provide a comprehensive understanding of the dataset and its underlying patterns.

3.2.2 Overview of the most important findings

The findings from the outlier analysis performed on the tumor and blood markers have revealed interesting insights. In the CA19-9 analysis, it was observed that all outlier patients were the same, except for patient 001PANC0011, who only appeared in the outlier table for CA19-9 values before the first chemotherapy cycle. After removing the outlier observations, a clear reduction in the spread of CA19-9 values was observed, as depicted in the provided boxplots and histograms in Figure B.191, Figure B.192, Figure B.193, and Figure B.194. Similar to the CA19-9 marker, many patients who exhibited outlier values for the CEA tumor marker at diagnosis and before the first chemotherapy cycle continued to display such values after it. Removing the outlier observations resulted in a noticeable reduction in the spread of CEA values.

Furthermore, since the the hemoglobin values already showed a normal distribution in the data analysis conducted in Appendix B.1, removal of the two identified outliers did not change this. Conversely, the thrombocyte (TB) values before and after the first chemotherapy cycle were not normally distributed before the removal of outliers. However, after the removal of outliers, TB before the first cycle became normal, with a p-value of 0.61 determined by the Shapiro-Wilk test. On the other hand, the TB values after the first cycle were still not considered normal when, this was also evident in the left-skewed histogram in Figure B.206. In addition, removing the outliers led to a significant decrease in the spread of leukocyte (LK) counts, with barely any outliers remaining in the new dataset. The LK before the first cycle are now considered normally distributed, but the LK after remain left-skewed. A similar trend is observed in the neutrophil counts, with the NP values before the first cycle becoming normally distributed after the removal of outliers. On the contrary, both the lymphocyte counts before and after the first cycle were normally distributed after the removal of outliers, as demonstrated by the histograms in Figure B.216 and Figure B.218.

To continue, the creatinine (CR) values before and after the first chemotherapy cycle followed a normal distribution (p-values: 0.70 and 0.57) after removing outliers. Boxplots indicate that the CR data, without outliers, does not contain any new outliers and falls within the determined IQR. On the other hand, the GFR data had numerous outliers before and after the first cycle. Histograms reveal a non-normal distribution with a mode value of 90 mL/min for many patients. Sodium (Na) values were non-normally distributed and remained so even after removing outliers. However, the p-values for Na before and after increased to 0.008 and 0.004, respectively. Only one outlier was found in the potassium dataset, observed in a single patient (001PANC0015) after the first cycle. The potassium values were already normally distributed and remained so even after removing this outlier, with higher p-values.

The liver variables ASAT, ALAT, AF, and GGT had numerous extreme outlier observations, which were removed using the IQR method. However, despite the reduction in spread in the variables, all variables remained non-normally distributed. In contrast, bilirubin data did not have extreme outliers but removing detected outliers still decreased the spread of values, resulting in a left-skewed histogram. Similarly, albumin values had a few outliers, which, when removed, led to normal distribution for albumin after values and no outliers in the new dataset. On the contrary, C-reactive protein values had many outliers and removing them reduced the range significantly. Moreover, the examination of the inflammation indices SII, NLR and PLR revealed that removal of the outlier observations resulted in the NLR and PLR values becoming normally distributed, both before and after the first chemotherapy cycle. In summary, elimination of the outlier observations led to a normal distribution for some variables, determined by the Shapiro-Wilk test with a significance level of $\alpha = 0.01$, and reduced the spread within nearly all variables under consideration, except for the HB, Na and K values, which exhibited only a few outlier observations.

Finally, to make the analysis complete, the difference between the values before and after the first chemotherapy cycle were also considered as a separate variable on which the exact same outlier analysis using the IQR-method has been performed. Surprisingly, after removing the outliers, many of the previously non-normal distributions of the difference variables exhibited a normal distribution. This was evident in 16 out of the 21 variables considered, as indicated by the fitted density of their histograms resembling a normal distribution. The table provided in Table B.96 provides a table with the patients (referred to by their respective Patient ID) and the number of outliers, as well as the corresponding variables.

Variable	**#Outlier values**	**Normal Distribution before removal**	**Normal Distribution after removal**	**Notes**
CA19-9 *diagnosis*	10	No	No	
CA19-9 *before*	11	No	No	001PANC0011
CA19-9 *after*	10	No	No	
CEA *diagnosis*	9	No	No	Different outliers
CEA *before*	11	No	No	078PANC0001
CEA *after*	10	No	No	
HB *before*	1	Yes	Yes	078PP20007
HB *after*	1	Yes	Yes	078PP20028
TB *before*	7	No	Yes (p-value 0.61)	One low value of 43.0
TB *after*	10	No	No (p-value 0.0022)	Different patients
LK *before*	9	No	Yes (p-value 0.010)	One low value of 2.50
LK *after*	3	No	No	
NP *before*	6	No	Yes (p-value 0.024)	Different patients
NP *after*	5	No	No	
LC *before*	6	No	Yes (p-value 0.078)	
LC *after*	4	No	Yes (p-value 0.090)	All patients in before too
CR *before*	4	No	Yes (p-value 0.700)	
CR *after*	5	No	Yes (p-value 0.572)	001PP20040 huge CR value of 952
GFR *before*	26	No	No	large number of outliers
GFR *after*	25	No	No	
Na *before*	1	No	No (p-value 0.0079)	
Na *after*	7	No	No (p-value 0.0043)	No outliers
K *before*	0	Yes (p-value 0.228)	Yes (p-value 0.244)	No outliers
K *after*	1	Yes (p-value 0.390)	Yes (p-vaue 0.547)	001PANC0015 only outlier

Variable	**#Outlier values**	**Normal Distribution before removal**	**Normal Distribution after removal**	**Notes**
ASAT before	19	No	No	001PANC0052 ASAT of1100
ASAT after	16	No	No (p-value 0.0014)	001PANC0032 ASAT of222
ALAT before	23	No	No	001PANC0052 ALAT of1241
ALAT after	16	No	No	001PANC0032 ALAT of264
AF before	14	No	No	
AF after	12	No	No (p-value 0.0027)	
GGT before	18	No	No	151PP20004 GGT of1110
GGT after	17	No	No	078PP20040 GGT of942
BR before	5	No	No	001PANC0025 BR of132
BR after	5	No	No	
Alb before	2	No	No (p-value of 0.0035)	
Alb after	3	No	Yes (p-value of 0.186)	
CRP before	17	No	No	
CRP after	12	No	No	
SII before	6	No	No (p-value 0.0012)	001PANC0036 SII of 7246.800
SII after	8	No	No	Almost all different outlier patients than before
NLR before	5	No	Yes (p-value 0.013)	
NLR after	9	No	Yes (p-value 0.012)	001PANC0015 and 151PP20004 very high NLR (36.8 and 36.0)
PLR before	6	No	Yes (p-value 0.020)	001PANC0036 PLR of 915.00
PLR after	10	No	Yes (p-value 0.028)	151PP20004 PLR of 725.00

Table 3.10: Overview table of the outlier analysis for all the measured variables. Note that only the p-values of the Shapiro-Wilk test of just (or just not) statistically significant (significance level $\alpha = 0.01$) normal distributed variables are given. Removal refers to removal of the outlier values of that particular variable with the IQR-method.

Variable Difference	**Number of Outlier values**	**Normal Distribution before removal**	**Normal Distribution after removal**	**Notes**
CA19-9	27	No	No	Large positive and negative Difference values
CEA	21	No	Yes (p-value 0.069)	Few patients with big difference
HB	4	No	Yes (p-value 0.162)	similar outlier values
TB	17	No	Yes (p-value 0.040)	
LK	5	No	Yes (p-value 0.074)	All negative
NP	4	No	Yes (p-value 0.044)	All negative
LC	4	No	Yes (p-value 0.514)	151PP20034 only positive Difference value
CR	15	No	Yes (p-value 0.269)	001PP20040 value of -843
GFR	37	No	No	
Na	0	Yes (p-value 0.044)	Yes (p-value 0.037)	No outliers Distribution did change after removal of the other Outlier values from K and Alb
K	6	Yes (p-value 0.171)	Yes (p-value 0.015)	
Alb	3	No	Yes (p-value 0.172)	
ASAT	24	No	Yes (p-value 0.022)	001PANC0052 value of 1079
ALAT	33	No	Yes (p-value 0.056)	001PANC0052 value of 1183
AF	11	No	Yes (p-value of 0.054)	
GGT	29	No	No	
BR	10	No	No	
CRP	16	No	No (p-value 0.001)	
SII	11	No	Yes (p-value 0.558)	
NLR	10	No	Yes (p-value of 0.183)	
PLR	7	No	Yes (p-value of 0.096)	

Table 3.11: Overview table of the outlier analysis of the difference variables. Note that only the p-values of the Shapiro-Wilk test of just (or just not) statistically significant normal distributed variables are given.

3.3. Principal Component Analysis

To gain a deeper understanding of the structures, patterns, and interrelationships among the variables, principal component analysis (PCA) has been conducted. The aim of PCA is to reduce the dimensionality of the dataset, while retaining as much of the original information as possible. Before applying PCA, the measured blood markers were organized into groups based on their specific characteristics. The variable INR was excluded from the analysis due to its limited number of data points. This section begins by elaborating the mathematical background of PCA followed by a summary of the findings of PCA conducted on the entire dataset using both the correlation and covariance matrix and on the dataset excluding outliers based on GFR and BR. The complete PCA using the correlation matrix can be found in Appendix B.3, using the covariance matrix in Appendix B.5 and without outliers based on GFR and BR in Appendix B.6. Furthermore, the analysis will encompass the following additional categories:

- Patient Characteristics: Height, Weight, BMI and Age
- Tumor Markers: CA19-9 and CEA
- All variables
- All variables, Age, BMI and the differences between the variables before and after the first chemotherapy treatment
- Age, BMI and the differences between the variables before and after the first chemotherapy treatment.

3.3.1 Principal Component Analysis Background Information

PCA is a statistical method used to reduce the dimensionality of a large dataset while preserving important information. It achieves this by transforming the original data into a new coordinate system called the principal component space. This transformation is a linear combination of the original variables, with each principal component (PC) capturing the highest variance in the data. By representing the data with a smaller number of uncorrelated variables, PCA enables the identification of key features and patterns while minimizing information loss. The PCs are ranked based on the amount of variance they explain, with PC1 capturing the most variance, followed by PC2, and so on. PCA is an unsupervised learning method, focusing only on the features and excluding the response variable (e.g., final response to chemotherapy). This prevents overfitting and unrealistic expectations. In this analysis, missing values are removed to ensure unbiased results, and the grouping of blood markers is done due to their potential correlations. Overall, PCA is a valuable tool for simplifying complex datasets and identifying the most significant features driving variability in the data.

3.3.1.1 PCA (mathematical details)

PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, the principal subspace, such that the variance of the projected data is maximized [54]. In order to understand the mathematics behind PCA, consider a dataset $\{x_i\}$ of $i = 1, \dots, n$ observations and x_i represents a vector in $\mathbb{R}^p$ with p features. The goal is to project the data onto a subspace having dimensionality m with $m < p$, while maximizing the variance of the projected data. For simplicity, we will focus on the projection onto a one-dimensional space ($m = 1$), which can easily be generalized to any $m \in \mathbb{N}$. Define the direction of this space using a $p-$dimensional vector ϕ_1. Without loss of generality, we can impose the constraint $\|\phi_1\| = 1$, since we are interested in the direction rather than the magnitude of ϕ_1. Each data point x_i is then projected onto the value $\phi_1^T\ x_i$ for $i = 1, \dots, n$. The mean of the projected data is $\phi_1 \overline{x_i}$, where $\overline{x_i}$ is the sample mean given by Equation (3.1) and the variance of the projected data is then given by Equation (3.2) with S the covariance matrix given by Equation (3.3).

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\phi_1^T x_i - \phi_1 \overline{x}\right)^2 = \phi_1 S \phi_1 \tag{3.2}$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T \tag{3.3}$$

To maximize the projected variance $\phi_1^T S \phi_1$ with respect to ϕ_1 while preventing $\|\phi_1\| \to \infty$, the appropriate constraint comes from the normalization condition of $\|\phi_1\| = 1$. This leads to an optimization problem with a Lagrange multiplier, expressed as,

$$J = \phi_1^T S\phi_1 + \lambda_1(1 - \phi_1^T\phi_1). \tag{3.4}$$

To find the stationary points of this optimization problem, we set $\frac{\partial J}{\partial \phi_1} = 0$, while yields $S\phi_1 = \lambda_1\phi_1$. This means that ϕ_1 must be an eigenvector of S. By left-multiplying by ϕ_1^T and considering $\|\phi_1\| = 1$ we find that the variance is given by $\phi_1 S\phi_1 = \lambda_1$, thus the variance is maximum when ϕ_1 is equal to the eigenvector having the largest eigenvalue λ_1. This eigenvector is known as the first principal component loading vector, containing the coefficients of the linear projection of the original data. Subsequently, additional principal component loading vectors can be defined in a similar manner by selecting directions that maximize the projected variance among all possible directions orthogonal to the previous components. In general, for any positive integer $m \in \mathbb{N}$, the linear projection is defined by the m eigenvectors $\phi_1, \ldots, \phi_m$ of the data covariance matrix S corresponding to the m largest eigenvalues $\lambda_1, \ldots, \lambda_m$ [54]. Note that all the $\phi_i, i = 1, \ldots, m$ are real since S is a symmetric matrix.

As a result of projecting the original standardized data onto the PCs, the new data becomes completely uncorrelated, causing the covariance matrix of the PCs to be a diagonal matrix. Specifically,

$$Cov(Z_1, \ldots, Z_m) = Diag(\lambda_1, ..., \lambda_m),$$

where $\lambda_1, ..., \lambda_m$ represent the eigenvalues associated with m PCs. In total there can be $m = min\{n-1, p\}$ PCs. If we denote the set of p features by $X_1, \ldots, X_p$, then PC1 can be defined as a normalized linear combination of these features that exhibits the largest variance as given in Equation (3.5).

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p, \tag{3.5}$$

subject to $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ with the elements $\phi_{11}, \ldots, \phi_{p1}$ being the loadings of the first principal component. Together these loadings make up the principal component loading vector $\phi_1 = (\phi_{11}\phi_{21}\ldots\phi_{p1})^T$, which we found earlier. Then, the scores, which are the projected values of the original data onto the first principal component, is calculated as follows,

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip} = \phi_1^T X \tag{3.6}$$

where the values $z_{11}, \ldots, z_{n1}$ correspond to the scores of PC1 and $X = (X_1, \ldots, X_p)^T$. When PC1 obtained from the features, PC2 can be derived by taking the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all the possible linear combinations that are uncorrelated with Z_1. In general, PC_j can be expressed as:

$$Z_j = \phi_j^T X \tag{3.7}$$

$$\text{subject to}\|\phi_j\|^2 = 1 \tag{3.8}$$

where ϕ_j is the principal component loading vector with elements $\phi_{1j}, \phi_{2j}, \ldots, \phi_{pj}$. The proportion of variance explained by the m^{th} PC is given in Equation (3.9). It is the ratio of the variance explained by the m^{th} principal component divided by the total variance present in the data. The choice of $m < p$ is usually taken such that the cumulative variance explained is larger than 80% or 90%. Equation (3.9) assumes that the features X_j are all centered around the mean zero. To find the cumulative proportion of variance explained we sum up the Equation (3.9) over the first m PCs.

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2} \tag{3.9}$$

There are two versions of PCA depending on whether X is only centered or also standardized and the optimization problem can be solved using either an eigendecomposition or a singular value decomposition. Standardization refers to the technique of scaling the values in such a way to have a distribution with mean zero and unit variance. For each feature X_j with mean μ_j and standard deviation σ_j and x_{ij} an observation, the standardized observation is given as,

$$x_{ij}^{standardized} = \frac{x_{ij} - \mu_j}{\sigma_j}. \tag{3.10}$$

Covariance PCA uses only a centered data matrix X and is used when the variables are on the same or similar scales. On the other hand, correlation PCA uses the standardized data and is preferred when the scales of the data are significantly different.

3.3.1.2 PCA approach in this data analysis

In this dataset, after categorizing the tumor and blood markers in the categories described in section 2.1, PCA is performed using the full data set on both the covariance as well as the correlation matrix and subsequently also after removal of outliers based on GFR and BR. A threshold of a cumulative variance explained of at least 80% is used and determined using scree plots. Subsequently, several other plots are used for the analyses. The interpretation of the plots is explained in Appendix B.3.

3.3.2 PCA on the various groups

This subsection contains a detailed PCA of the various groups described before and stated in Table 3.12. The final response analyzed is the dichotomized final response, where 0 = Disease Control (CR, PR, SD) and 1 = Progressive Disease (PD). The full PCA analysis can be found in Appendix B.3. However, to illustrate how the PCA analysis is performed as well as how to interpret the provided plots, the blood cell group will be examined in detail in the following subsection. Similar analyses have been performed in the other groups for both the full dataset using correlation and covariance matrices as well as after removal of outliers based on GFR and BR.

Group	**Variables**
Blood	HB, TB, LK, NP, LC
White Blood Cells	LK, NP, LC
Kidney	Na, K, CR, GFR
Liver	ASAT, ALAT, AF, GGT, BR and INR*
Nutrition	Alb, Na, K
Inflammation	CRP, LK, SII, NLR, PLR
Patient Characteristics	Height, Weight, BMI, Age
All measured variables	HB, TB, LK, NP, LC, Na, K, CR, GFR, ASAT, ALAT, AF, GGT, BR, Alb, CRP, LK, SII, NLR, PLR
All, Age, BMI and differences	HB, TB, LK, NP, LC, Na, K, CR, GFR, ASAT, ALAT, AF, GGT, BR, Alb, CRP, LK, SII, NLR, PLR, Age, BMI
All, Age, BMI only differences	HBdiff, TBdiff, LKdiff, NPdiff, LCdiff, Nadiff, Kdiff, CRdiff, GFRdiff, ASATdiff, ALATdiff, AFdiff, GGTdiff, BRdiff, Albdiff, CRPdiff, LKdiff, SIIdiff, NLRdiff, PLRdiff, Age, BMI

Table 3.12: Grouped variables used for PCA

3.3.2.1 PCA on the Blood Cell group

The blood cell dataset consists of variables related to HB, TB, LK, NP, and LC before and after chemotherapy, as well as the differences between their values. The dataset contains 109 observations with 15 explanatory variables. The correlation matrix in Figure 3.6 reveals strong positive correlations between HB values before and after the first chemotherapy cycle, as well as between LC values before and after. Additionally, correlations are observed between LK, NP, and LC, which is consistent with their nature as white blood cells. HB shows no significant correlation with other variables, except for TB before and after treatment. Differences in LK, NP, and LC exhibit high correlation, reflecting their common representation as white blood cells. However, no correlation is observed between the difference in HB and TB, except for their own values before and after treatment.

The screeplot in Figure 3.7 shows that the “elbow point" is not clear, but the first five PCs account for 81% of the total variance, which is considered sufficient. Table 3.13 provides the standard deviation, proportion of variance explained, and cumulative proportion for the first five PCs. The standard deviation reflects the variability captured by each principal component, and higher eigenvalues indicate greater variance accounted for than a single original variable. Variance explained indicates the proportion of variability captured by each principal component, while the cumulative proportion represents the cumulative sum of explained variances. Next, the scatter plot in Figure 3.8 represents the observations in the principal component (PC) space. Each individual's position in the plot is determined by their scores along the PCs. The color of each point indicates the quality of representation, with yellow and red indicating high-quality representation and blue indicating low-quality representation. The plot shows that individuals closer to the center are poorly represented, while those further away from PC1 or PC2 have better representation. The proximity of points to the horizontal axis indicates their correlation with PC1, while proximity to the vertical axis indicates their correlation with PC2. Further patterns or clusters can be observed, providing insights into the structure of the dataset.

Moving on, Figure 3.9 displays the contributions of the original variables in the blood dataset to each PC. Variables are represented as arrows, with arrow length and direction indicating their relative importance and contribution to the principal component's direction. The ring in the plot illustrates the total variance captured by PC1 and PC2. Arrows close to the circle's boundary imply less loss of information in that PC direction, while shorter arrows suggest smaller contributions. Variables that are highly positively correlated appear as closely aligned arrows, while highly negatively correlated variables form an angle larger than 90 degrees but smaller than or equal to 180 degrees. Uncorrelated variables are represented by a 90-degree angle. This loading plot can be interpreted as follows: Initially, all variables were defined in a 15-dimensional space (due to the 15 variables in the blood dataset). After PCA, this space is projected onto a 2D space, represented by PC1 and PC2. Variables such as NPdiff, LCdiff, HBdiff, LKafter, and NPafter are positioned close to PC1, indicating strong positive or negative correlations with PC1. These variables also have relatively high loading scores in the loading score matrix. Conversely, variables like TBbefore, TBafter, HBbefore, and HBafter are positioned near the vertical axis (PC2), suggesting high correlations with PC2 and minimal correlation with PC1. These variables exhibit small loading scores in dimension 1 but high loading scores in dimension 2. For more details, refer to the loading matrix shown in Figure 3.11.

As a pre-processing step, all variables in the blood data group were standardized according to Equation (3.10). As a consequence, the cosine of the angle between two variables is equivalent to the correlation coefficient between them, provided that the variables are well-represented. Hence, the plot in Figure 3.9 shows a geometrical representation of the correlation coefficients between the variables in the blood data group. Note that this is only true in the high dimensional space (in this case the 15 dimensional space), while the plot shows a visualization of the correlations between the variables in the projected 2D space. The variables LKafter and NPafter are located near the correlation circle, as can be seen by their long vectors pointing close to its circumference. This indicates that they are well represented and that the projection depicted in the plot is a close approximation of the actual vectors in the high-dimensional space. Similarly, the variables LKbefore, NPdiff, and LKdiff also satisfy this condition. Therefore, the angle between the vectors can be used to visualize the correlation between them. More precisely, the cosine of the angle between the projected values of LKafter and NPafter is almost the same as the cosine of the angle of the original variables in the high dimensional space. Mathematically, the cosine of the angle between the projected values P_A and P_B of variables A and B is almost equal to the cosine of the angle of the original variables in the high dimensional space, as given in equation (3.11), provided that the variables are well-projected.

$$cos(\theta_{A,B}) \approx cos(\theta_{P_A,P_B}) if\ the\ variables\ are\ well\text{-}projected\ (close\ to\ the\ correlation\ circle). \tag{3.11}$$

However, variables that are not well-projected, such as LCdiff and HBdiff (indicated by small arrows and more blue values), may have a small angle in the projection plane while having a much larger angle in the original space. Thus, the cosine of the angle in the projection plane does not accurately estimate the cosine of the angle in the original space, making it difficult to determine the correlation coefficient between poorly represented variables. Only well-projected variables, those close to the correlation circle, can be reliably interpreted. Consequently, it is not possible to estimate the correlation coefficient between two poorly represented variables based solely on the projection.

Moreover, the biplot shown in Figure 3.10 combines the scatter plot from Figure 3.8 and the loading plot from Figure 3.9, and can be interpreted similarly. Firstly, the position of the points represents the observations projected onto the principal component space defined by PC1 and PC2. The proximity of observations in the biplot indicates their similarity or dissimilarity. The direction of the arrows represents the direction of highest variance for each variable, while the length of the arrows reflects the strength of the variables in that direction. Positive correlations between variables are depicted by arrows pointing in the same direction, while negatively correlated variables have arrows pointing in opposite directions. Uncorrelated variables have perpendicular arrows. Additionally, the proximity of observations to a variable indicates the correlation between the observation and the variable. For instance, vectors representing variables like NPdiff and LKdiff predominantly point to the right in the horizontal direction, indicating high positive correlations with PC1. Thus, observations close to these variables have high values for NPdiff and LKdiff, while those far to the left have smaller values. On the other hand, variables such as LKafter and NPafter exhibit a correlation close to -1 with PC1, resulting in values moving in the opposite direction. Consequently, the first dimension of the biplot separates patients with high LK and NP after values (on the left) from those with high NPdiff and LKdiff values (on the right), representing the primary source of variability. The second dimension indicates the high representation of variables LKbefore, NPbefore, and TBbefore. Observations close to these vectors exhibit high correlation with these variables and have high values for them. Conversely, HBbefore has a correlation coefficient close to -1, causing observations with small values in the second dimension to be positioned on the same side as the corresponding variables with high values but opposite to variables with low values. Consequently, observations can be far from variables for which they have small values. For example, variables with large negative values near the vertical axis will have large negative values for HBbefore but small values for NPdiff or LKafter.

The correlation matrix and the $\cos^2$ bar chart in Figures 3.12 and 3.13, respectively, provide insights into the correlation between the original variables and the PCs. The correlation plot shows absolute correlation values between the original variables and the PCs. It confirms previous findings that PC1 is highly correlated with variables LKafter, NPafter, LKdiff, and LPdiff. Additionally, in the second dimension, variables LKbefore and NPbefore exhibit high correlations. The quality of representation is visualized in the $\cos^2$ plot, where variables LK, NP, and LC are identified as the most important in the blood dataset. A high cos2 value indicates a good representation of the variable in that dimension and the variable will also be positioned close to the circumference of the correlation circle given in Figure 3.9. Conversely, low $\cos^2$ values indicate poor representation, and the variables will be closer to the center of the correlation circle. Therefore, $\cos^2$ values are used to assess the quality of representation of the original variables in the principal component space. Variables highly correlated with the first two PCs are crucial in explaining the total variability in the dataset, as PC1 and PC2 capture the most variance. Variables with small correlations or high correlations in the last dimension may be removed to simplify the analysis. Furthermore, the contribution of the original variables to the PCs is visualized in Figure 3.14, with darker colors and larger sizes indicating greater contributions of the original variables to specific dimension

Finally, in Figure 3.15 a PCA plot is shown with the data points projected onto the first two PCs. The percentage of variance explained by each PC is displayed on the axes, indicating how much of the variability in the data is captured by each PC. Each data point represents an observation from the original dataset with the color and shape determined by the final response. 95% Prediction ellipses are plotted to indicate the regions of the plot where the majority of the points are expected to fall. From the figure it can be seen that the ellipses overlap mostly, indicating that the first two dimensions are not able to separate the blood data based on the final response. However, it is also possible that the overlap is due to high variability within each group or due to the fact that the original chosen variables may not be the best ones to use for distinguishing between the groups. Logically, it makes sense, since blood values alone are most likely insufficient to predict a final response to chemotherapy. Since the first two dimensions were not able to clearly separate the two final response groups, the plot in Figure 3.16 shows a matrix scatter plot of the two groups in the first four dimensions. It is apparent from this graph that there is no clear trend. Closer inspection only shows that the DC group seems to be more scattered throughout the spaces compared to the PD groups based on the blood values. Nevertheless, this might a consequence of the smaller sample size in the PD group. The 3D-plots in Figure 3.17 and Figure 3.18 show the scores of the observations projected onto the first three PCs in a 3D visualization.

Blood	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	2.174	1.586	1.359	1.307	1.189
Proportion of Variance Explained	0.315	0.168	0.123	0.114	0.094
Cumulative Proportion	0.315	0.483	0.606	0.720	0.814

Table 3.13: PCA summary values of the blood group (n=109 (Disease control (n=85), Progressive disease (n=24))). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five PCs.

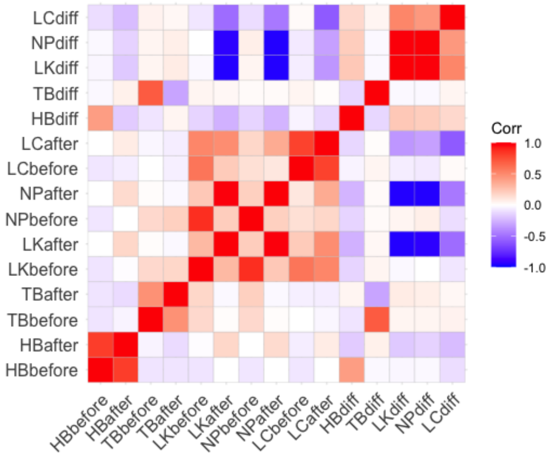


Figure 3.6: Correlation Matrix of the various variables in the Blood group, n=109 (Disease control (n=85), Progressive disease (n=24)). Red = Positive correlation, Blue = negative correlation and White = No correlation.

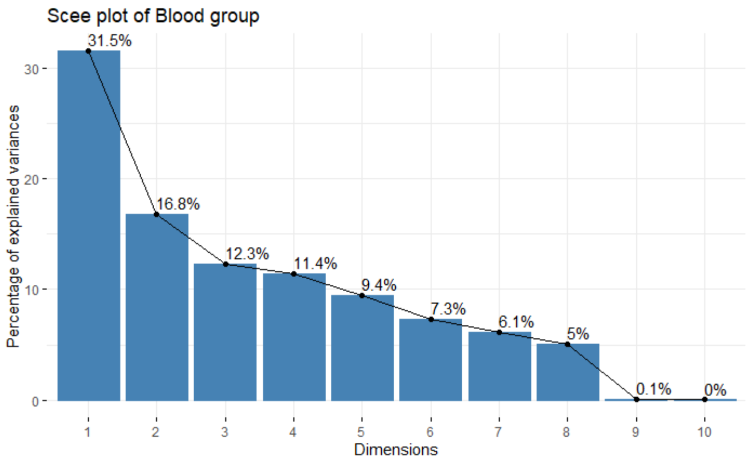


Figure 3.7: Scree plot of the blood group variables, n=109 (Disease control (n=85), Progressive disease (n=24)).

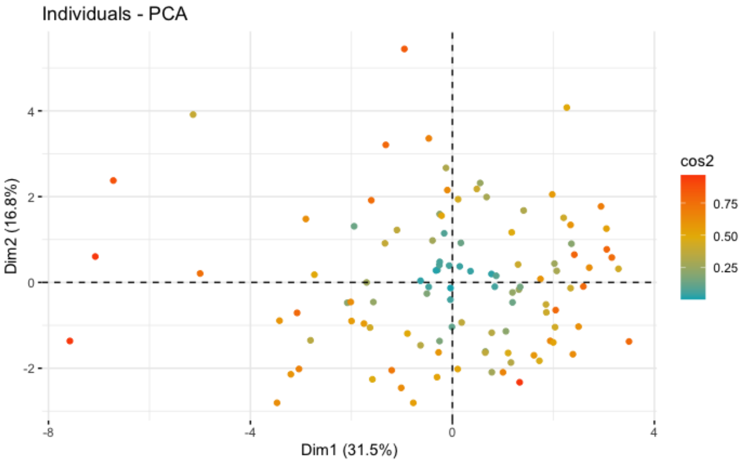


Figure 3.8: Scatter plot of the blood group variables, n=109 (Disease control (n=85), Progressive disease (n=24)).

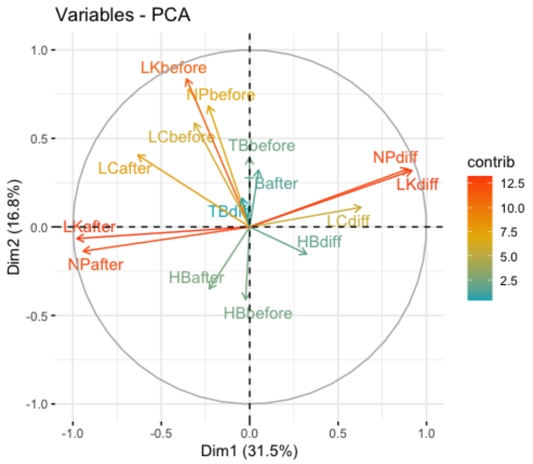


Figure 3.9: Loading plot of the blood group variables n=109 (Disease control (n=85), Progressive disease (n=24)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

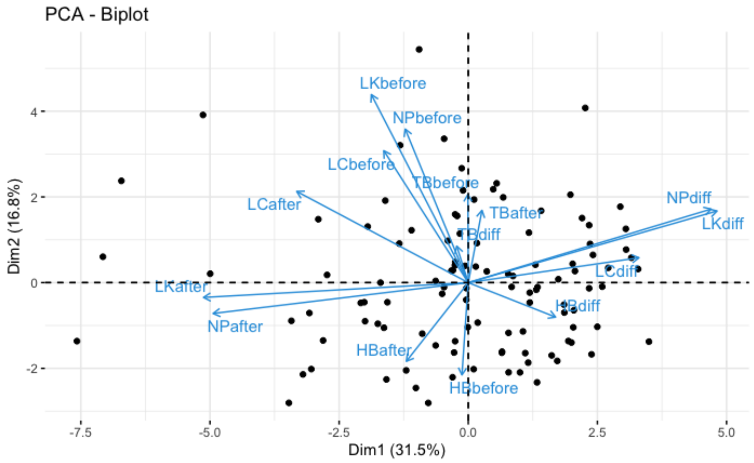
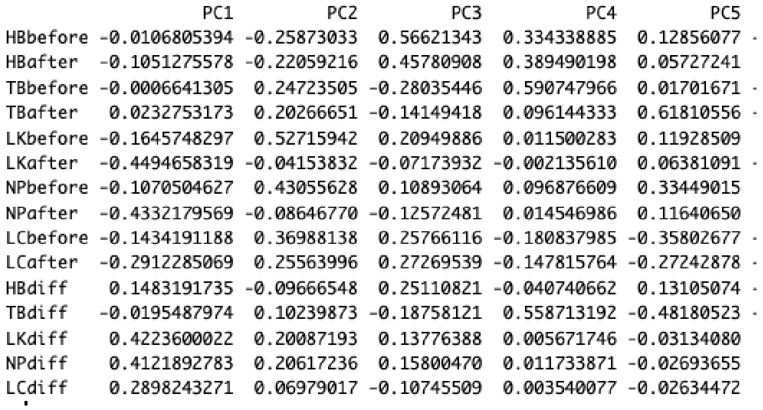


Figure 3.10: Biplot plot of the blood group variables n=109 (Disease control (n=85), Progressive disease (n=24)).



```
                   PC1         PC2         PC3          PC4         PC5
HBbefore -0.0106805394 -0.25873033  0.56621343  0.334338885  0.12856077 -
HBafter  -0.1051275578 -0.22059216  0.45780908  0.389490198  0.05727241
TBbefore -0.0006641305  0.24723505 -0.28035446  0.590747966  0.01701671 -
TBafter   0.0232753173  0.20266651 -0.14149418  0.096144333  0.61810556 -
LKbefore -0.1645748297  0.52715942  0.20949886  0.011500283  0.11928509
LKafter  -0.4494658319 -0.04153832 -0.07173932 -0.002135610  0.06381091 -
NPbefore -0.1070504627  0.43055628  0.10893064  0.096876609  0.33449015
NPafter  -0.4332179569 -0.08646770 -0.12572481  0.014546986  0.11640650
LCbefore -0.1434191188  0.36988138  0.25766116 -0.180837985 -0.35802677 -
LCafter  -0.2912285069  0.25563996  0.27269539 -0.147815764 -0.27242878 -
HBdiff    0.1483191735 -0.09666548  0.25110821 -0.040740662  0.13105074 -
TBdiff   -0.0195487974  0.10239873 -0.18758121  0.558713192 -0.48180523 -
LKdiff    0.4223600022  0.20087193  0.13776388  0.005671746 -0.03134080
NPdiff    0.4121892783  0.20617236  0.15800470  0.011733871 -0.02693655
LCdiff    0.2898243271  0.06979017 -0.10745509  0.003540077 -0.02634472
```

Figure 3.11: Loading matrix of the blood group variables of the first 5 PCs, n=109 (Disease control (n=85), Progressive disease (n=24)).

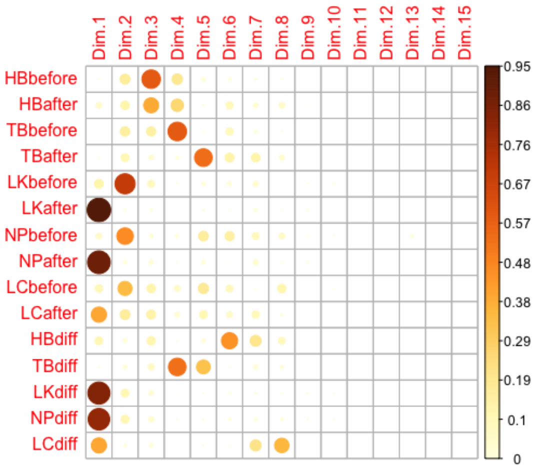


Figure 3.12: Correlation of the original blood group variables with the PCs, n=109 (Disease control (n=85), Progressive disease (n=24)).

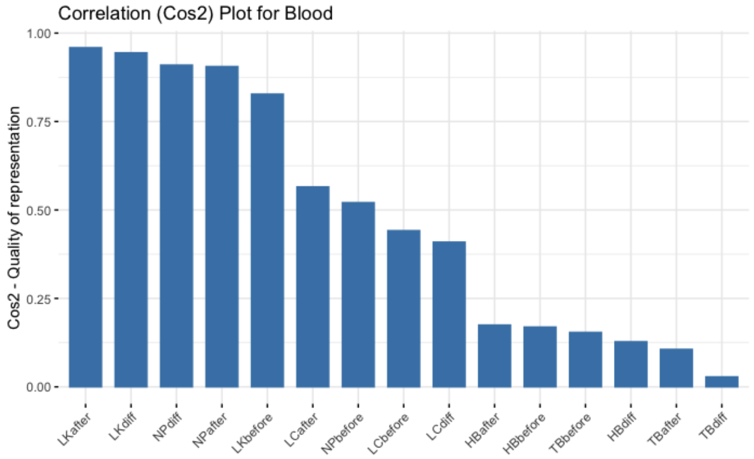


Figure 3.13: Cos2 bar chart of the original blood group variables, n=109 (Disease control (n=85), Progressive disease (n=24)).

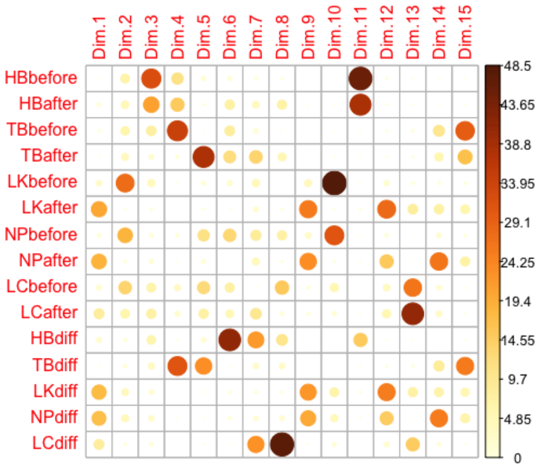


Figure 3.14: Contribution of the original blood group variables to the PCs, n=109 (Disease control (n=85), Progressive disease (n=24)).

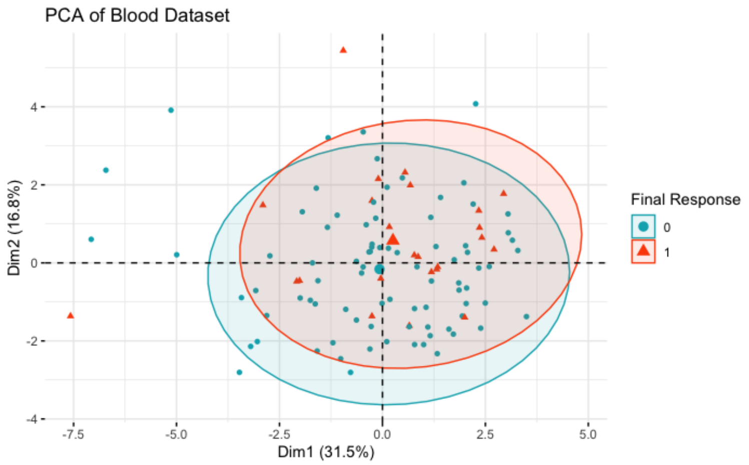


Figure 3.15: PCA plot of the blood group with prediction ellipses, n=109 (Disease control (n=85), Progressive disease (n=24)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

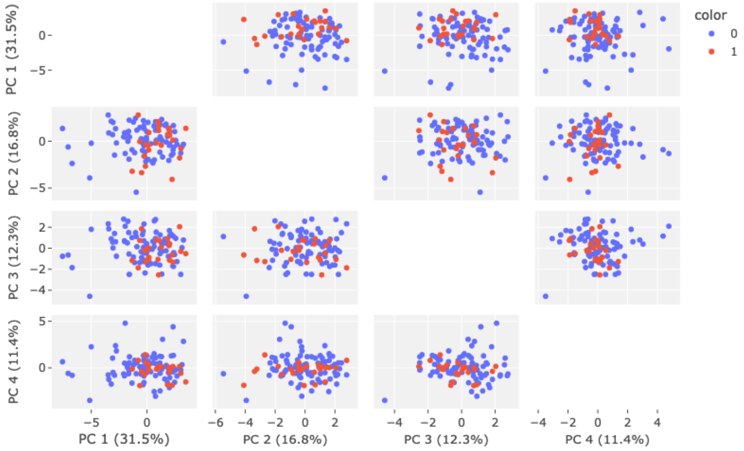


Figure 3.16: Scatter matrix plot of the first four PCs, n=109 (Disease control (n=85), Progressive disease (n=24)). The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the blood group variables with the final dichotomized response is coloured, Blue = 0 (Disease Control) and Red = 1 (Progressive Disease).

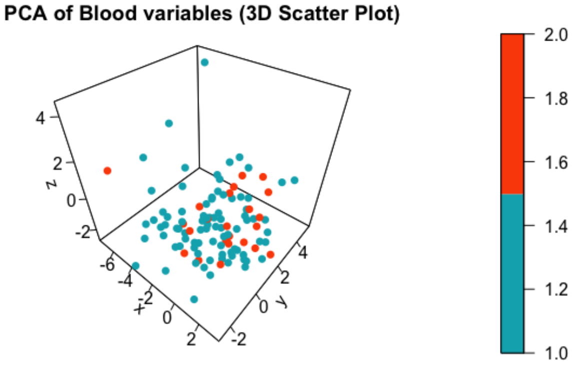


Figure 3.17: 3D plot of the blood group variables projected onto the first three PCs, n=109 (Disease control (n=85), Progressive disease (n=24)).

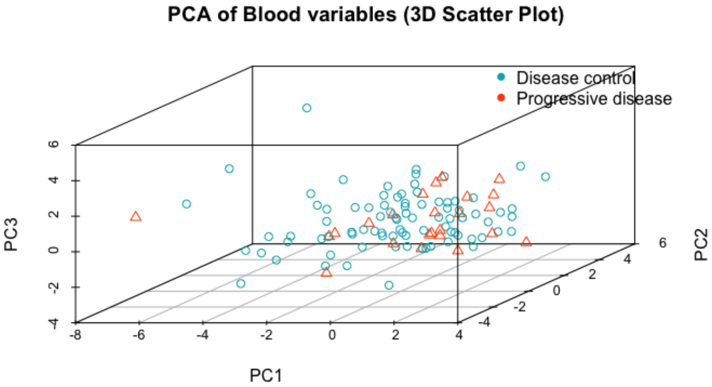


Figure 3.18: 3D plot of the blood group variables with the final dichotomized response is coloured and shaped, where Blue = 0 (Disease control) and Red = 1 (Progressive disease), n=109 (Disease control (n=85), Progressive disease (n=24)).

3.3.3 Overview of significant PCA findings

This section presents a brief overview of the key results obtained from the PCA conducted on the original standardized variables using the correlation matrix. A summary of the main findings is provided in Table 3.14. The first column of the table lists the group of variables that were analyzed using PCA. The second column reports the number of observations in each data set with the number of patients classified as DC and PD in brackets respectively, along with the number of explanatory variables and PCs needed to explain at least 80% of the total variance in the data. The subsequent column presents the variables that are most influential in driving the variation in the data (i.e., those with the highest correlation with the first principal component), which can be viewed as the primary factor that explains the structure in the data if the majority of the information is captured by the first principal component. The prediction ellipse column provides information on the shape and overlap of the final response 95% prediction ellipses, while the last column includes any notable correlations or other observations that emerged from the analysis.

The blood data group showed positive correlations among many before and after values of the same variables. However, HBdiff and TBdiff had no correlation with each other or with differences in the white blood cell variables. Moreover, LKdiff and NPdiff exhibited a strong positive correlation, which aligns with their biological connection, as NP are a type of LK. The most important variables in determining PC1 were LKafter, NPafter, LKdiff, and NPdiff. Of these, LKafter and NPafter were highly positively correlated with each other, while strongly negatively correlated with LKdiff and NPdiff. Strong correlations among variables indicate dependencies and shared factors in the data. When considering only the white blood cell variables, the same variables as before were the key variables capturing variation in PCA. However, projecting the data onto the first two or three dimensions did not distinguish between the DC and PD groups.

In contrast, the kidney function dataset showed no clear correlations among variables, except for their respective before, after, and difference values. CRbefore, CRafter, GFRbefore, and GFRafter were most influential for PC1, Kbefore for PC2, and Kdiff and Kafter for PC3. Nevertheless, no clear separation between DC and PD groups was found after projecting the variables onto the first two or three PCs. This lack of distinct separation suggests that the variability in kidney function as well as the (white) blood cell groups is not sufficient for distinguishing between the two response groups.

Moreover, the liver group dataset exhibits strong positive correlations among most variables. Surprisingly, the before and after values of ASAT and ALAT showed weak correlations. However, there was a notable positive correlation between ASAT and ALAT values, as well as between their respective differences. PCA identified ALATbefore, ASATbefore, ASATdiff, and ALATdiff as influential variables for PC1, while AFafter and GGTafter predominantly contribute to PC2. Despite these correlations and variable importance, the 2D and 3D projections of the data fails to provide a clear distinction between the DC and PD groups. Unlike the liver group dataset, the nutrition data group demonstrated weak correlations among variables, except for the before, after, and difference values of each variable, which displayed either positive or negative correlations. Naafter and Albafter were identified as the most influential variables for PC1, capturing significant variation in the dataset. PC2 was mainly influenced by Kafter, while PC3 was determined by Kdiff. Despite the significance of these variables and PCA, projecting the data onto the first two or three PCs did not distinctly separate the DC and PD groups. Thus, the analyzed liver and nutrition variables alone may not possess sufficient discriminatory power to differentiate between the DC and PD groups. Additional factors need to be considered for a comprehensive assessment.

Similar to the liver group dataset, the inflammation dataset exhibits several correlations among the variables, particularly for SII, NLR, and PLR. These correlations are expected, given the calculation methods for these inflammation ratios. However, the CRP difference variable showed no correlation with the LK difference and only weak negative correlations with SII, NLR, and PLR differences. On the other hand the before, after and difference values of each variable displayed strong positive or negative correlations with each other, consistent with the definition of these variables. These correlation patterns are evident in the loading and biplots. The most important variables in determining PC1 were SIIdiff, NLRdiff, and PLRdiff, which can be explained by the strong correlations among them. The high importance of these variables reflects their contribution to the primary dimension capturing significant variation in the dataset. However, despite the presence of these correlations and the importance of the identified variables, projecting the data onto the first two or three dimensions did not result in a clear distinction between the DC and PD groups. Moreover, in the patient characteristics dataset, weight and height exhibited a strong positive correlation, as expected due to their inherent relationship. Similarly, BMI and weight showed a strong positive correlation, consistent with the calculation of BMI. Age, on the other hand, showed no correlation with other variables, as anticipated given its independence in this context. Weight and BMI were identified as the most influential variables in determining PC1, capturing significant variation in the dataset. PC3 was primarily influenced by age. However, considering only these four patient characteristics is insufficient to effectively distinguish between the DC and PD groups. Despite using three components to explain nearly all the variation in the data, projecting the scores onto the first two or three PCs did not reveal a clear distinction between the two response groups. Besides the blood markers, the tumor markers CA19-9 and CEA, play a crucial role in the prognosis of PDAC. Upon examining the correlation matrix in Figure B.393 it becomes evident that the diagnosis, before, and after values of the tumor markers exhibit very strong positive correlations, while the difference values do not show significant correlations. This correlation pattern is reflected in the determination of PC1 and PC2. PC1 is primarily determined by the CEA and CA19-9 diagnosis, before, and after values, whereas PC2 is influenced by the CEAdiff and PC3 is mainly determined by the CA19-9diff. Despite these correlations and the ability of the PCs to explain a substantial amount of variation, projecting the data onto two or three dimensions did not result in a clear distinction between the DC and PD groups. This suggests that relying solely on the tumor marker values, specifically CA19-9 and CEA, may not be sufficient to effectively differentiate between the final response groups.

After analyzing the blood markers and tumor markers separately, we now combine all these variables into one comprehensive dataset. The correlation matrix did not reveal any new correlations between the variables, except for a strong negative correlation between CR before and GFR before. The screeplot illustrates that no single principal component captures a substantial amount of variability in the data. Examination of the cos2 bar chart reveals that the variables best represented in dimensions 1 and 2 are CEAbefore, GGTafter, AFafter, CEAafter, and GGTbefore. Again, it becomes apparent that projecting the data onto the first two or three PCs does not yield a clear separation between the DC and PD groups. Furthermore, the patient characteristics dataset was expanded to include additional variables, namely Age and BMI, while excluding Weight and Height due to the strong correlation observed between them. However, the screeplot showed that no individual principal component dominates the overall variability. The most important variables determining PC1 and PC2 include ALATdiff, ALATbefore, ASATdiff, ASATbefore, GGTdiff, GGTbefore, LKafter, NPafter and BRbefore. As previously observed, ALAT and ASAT values exhibit a strong positive correlation, as do LKafter and NPafter. Consequently, when projecting the data onto the first two or three dimensions, no clear differentiation between the DC and PD groups is apparent. This analysis emphasizes that despite incorporating patient characteristics and a comprehensive set of blood and tumor markers, the PCs derived from the data do not exhibit distinct separability between the two final response groups.

Finally, PCA was further conducted on the dataset containing only the patient characteristics Age and BMI alongside the difference before and after the first chemotherapy cycle values of the blood and tumor markers. The analysis showed a strong correlation between LK and NP was observed as well as between ASAT and ALAT values. Furthermore, positive correlations were identified among AF, GGT and BR variables, all conforming previous observations. Despite the presence of these correlations, no single principal component was able to account for a substantial portion of the total variation in the dataset. The most influential variables in determining PC1 were ASATdiff, ALATdiff, AFdiff, GGTdiff and BRdiff while PC2 was primarily determined by LKdiff and NPdiff. Nevertheless, projecting the data onto the first two or three dimensions did not reveal a clear distinction between the DC and PD groups.

In summary, PCA was performed on various datasets, including bloodcell values, tumor markers, liver function, nutrition data, inflammation data, patient characteristics, and a comprehensive dataset combining all variables as well as their differences. The results revealed positive correlations among many before, after and difference values of the same variables in data groups. The most important variables in the blood dataset and white blood cell data were the same. Surprisingly the before and after values of both ASAT and ALAT showed weak or no correlation, while the two variables themselves showed an extreme positive correlation. Moreover, the variables in the nutrition dataset exhibited weak correlations with each other. Furthermore, the inflammation data showed strong correlations among specific inflammation ratios, and patient characteristics demonstrated expected correlations between weight and height, and BMI and weight. Overall, this analysis has the overarching conclusion that the studied variable groups as well as all the variables together do not exhibit distinct separability between the final response groups in the first two or three dimensions, despite having explored various combinations of the variables. However, PCA assumes *linear relationships* between variables. Thus, if there are non-linear relationships present, PCA may not accurately capture the underlying structure and patterns in the data. Moreover, exploring non-linear modeling approaches and incorporating more comprehensive datasets could provide further insights into the complex relationships between variables and improve the understanding of the factors influencing the response to chemotherapy in PDAC patients.

3.3.4 PCA with covariance matrix overview

The previous PCA was conducted using the correlation matrix after centering and scaling the original measured values of all variables. However, considering unscaled values in PCA could provide additional insights. Unscaled values may be informative because variables with small variance can lead to larger values during standardization, because each value is centered and divided by its standard deviation. The choice between using a covariance or correlation matrix in PCA can impact the resulting PCs in terms of magnitude and direction. PCs derived from a covariance matrix reflect the covariance structure of the original variables, while those derived from a correlation matrix reflect the correlation structure after standardization. However, the overall pattern of variability across the PCs remains the same regardless of the matrix used. Using a correlation matrix is more suitable when seeking relationships between the original variables, while a covariance matrix is more appropriate when focusing on the overall variability of the variables.

If a single PC captures a significant portion of the total variation in the data, it suggests the dominance of a specific factor driving the variation. For example, suppose we are analyzing a dataset that contains measurements of various physical properties of different fruits, such as weight, color, sweetness, and acidity. If PC1 captures 97% of the total variation in the data, it means that one of these properties (or a combination of them) is driving most of the variation in the dataset. We might interpret this to mean that the weight of the fruit, for example, is the most important factor in determining the overall structure of the dataset, and that other properties like sweetness and acidity are less important. Note that these properties might be dominant in high order PCs. However, it's important to note that reducing the dimensionality to a single principal component may result in the loss of information about other important factors contributing to the variation. Careful consideration is needed to balance dimensionality reduction and information preservation when interpreting PCA results for a specific analysis or application. An overview of all the findings using PCA on the covariance matrix can be found in Table 3.15 with the full analysis presented in Appendix B.5. This analysis reveals a significant dominance of CA19-9, as it appears that almost all the variance in the dataset is predominantly captured by CA19-9 when considering the entire dataset.

3.3.5 PCA after removal of outliers overview

The PCA was performed without removal of outliers due to the small sample size and the potential loss of valuable information or distortion of the data structure. Including outliers in the analysis ensures the integrity of the dataset and provides a robust assessment of the PCA results. However, outliers can have a significant influence on PCA as they can distort the estimation of the PCs and the representation of the majority of the data points. They can increase the variance and covariance in certain dimensions, leading to a distortion of the covariance matrix. Moreover, outliers can bias the directions of greatest variation towards themselves, potentially distorting the overall representation. Additionally, outliers with large variances can dominate the total variance, causing certain components to explain a disproportionate amount of the variance. This can lead to an overestimation of their importance and an underestimation of others. Finally, outliers can significantly impact the projection of the data onto the selected PCs, as they skew the data distribution. This was observed in many variables in the analysis conducted in Appendix B.1.

A brief examination of the PCA results on the dataset, excluding outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values, as suggested by medical experts from the Erasmus MC Rotterdam, before chemotherapy, is presented in this section. The rationale for selecting these values is further elaborated in Section 5. An overview of the PCA results after outlier removal is summarized in Table 3.16. In this table, the data groups are defined as explained earlier, with ‘n’ representing the total number of observations. The values within parentheses denote the breakdown of observations into DC and PD groups. For instance, ‘105 (83,22)’ indicates a group comprising 105 observations after removing missing values, with 83 patients classified as DC and 22 patients as PD. Similar to the overview table in Table 3.14, ‘EV’ denotes the number of explanatory variables within each data group. ‘PCs’ represents the minimum number of PCs required to explain at least 80% of the total variation in the data. The cumulative variance explained is presented in the column labeled ‘Cum var expl’. Furthermore, the most significant variables in each data group, based on their cos2 values, are identified in the column ‘Important variables’. Additionally, a brief description of the shape of the 95% prediction ellipses is provided in the ‘Prediction Ellipse’ column and the final column compares the PCA results with the analysis conducted in Appendix B.3, encompassing scree plots, loading plots, and other relevant plots used in the analysis of different groups. These comparisons reveal minimal differences in certain groups, while notable changes are observed in others. The plots that have changed compared to the previous PCA are presented in Appendix B.6.

Data Group	**n**	**#EV**	**#PCs**	**Cumulative variance explained**	**Important Variables**	**Prediction Ellipse**	**Notes**
Blood	109(88,23)	15	5	81.4%	NPdiff, LKdiff LKafter, NPafter LKbefore, NPbefore	Overlap almost entirely, most variation in PC1	High correlation in LK and NP values
White Blood Cells	111(87,24)	9	3	90.5%	Npafter, LKafter, LKdiff, NPdiff, LKbefore	Overlap almost entirely, most variation in PC1	High correlation in LK and NP values
Kidney	145(114,31)	12	5	80.5%	CRbefore, GFRbefore CRafter, GFRafter	Overlap, DC within PD	Almost no correlations between the different variables
Liver	166(135,31)	15	4	82.5%	ASATbeofre ALATbeofre ASATdiff ALATdiff	Overlap, DC within PD	ASAT/ALAT, GGT/AF, show positive correlations
Nutrition	102(80,22)	9	5	90.8%	Naafter, Albafter, Kafter	Overlap, DC within PD DC circular	No strong correlations between the original variables
Inflammation	90(71,19)	15	4	83.1%	SIIdiff, NLRdiff PRdiff	Overlap	SIIdiff, NLRdiff, PLRdiff highly correlated
Patient Characteristics	189(151,38)	4	3	99.9%	Weight, BMI	Overlap	Weight and Height show positive correlation
Tumor Markers	96(76,20)	8	2	82.8%	CA199after, CA199before, CA199diag, CEAdiag, CEAbefore, CEAafter	Overlap, DC within PD	CA199diff and CEAdiff show strong positive correlation
All blood & tumor variables	59(44,15)	34	11	82.7%	CEAbefore, GGTafter, AFafter, CEAafter, GGTbefore	Overlap, DC within PD	Most important variables are: CEA, GGT, AF, ALAT, ASAT, GFR, CA19-9
All variables + Age + BMI + Differences	59(44,15)	53	13	81.8%	ALATdiff, ASATdiff, GGTdiff, LKafter, NPafter, BRbefore	Overlap, DC within PD	Most important variables are: ALAT, ASAT, GGT, LK, NP, BR and AF
Age + BMI + Differences	59(44,15)	19	9	81.6%	ALATdiff, GGTdiff, ASATdiff, BRdiff, AFdiff, LKdiff, NPdiff	Overlap, DC within PD	NPdiff/LKdiff, ALATdiff/ASATdiff, are highly correlated variables

Table 3.14: Overview table of PCA performed on the correlation matrices of the analyzed data sets, EV = Explanatory Variables, PC = Principal Component.

Group	**#PCs**	**Cumulative variance explained**	**Important Variables**	**Ellipse overlap?**
Blood	2	99%	TBbefore, TBdiff, TBafter	Yes
WBC	1	92%	LKafter, LKdiff, NPafter, NPdiff	Yes
Kidney	1	92%	CRafter, CRdiff, CRbefore	Yes
Liver	3	91%	GGTbefore, GGTafter GGTdiff, AFbefore, ALATdiff, ALATbefore	Yes
Nutrition	3	91%	Albafter, Albbefore, Albdiff, Naafter, Nadiff, Nabefore	Yes
Inflammation	2	99%	SIIdiff, SIIafter, SIIbefore	Yes
Patient Characteristics	3	100%	Weight, Height, Age	Yes
Tumor Markers	1	97%	CA19-9after, CA19-9before, CA19-9diag	Yes
All blood & tumor variables	1	97%	CA19-9after, CA19-9before	Yes
All variables + Age + BMI + Differences	2	99%	CA19-9after, CA19-9before	Yes
Age + BMI + Differences	1	96%	CA19-9diff	Yes

Table 3.15: Overview table of PCA performed on the covariance matrix. The first column contains the group of variables considered in the PCA analysis that have been centered, but not scaled to unit variance. The second column shows how many PCs are necessary to explain at least 90% (or more) of the total variation the data. The third column contains the variables that are most important in driving this variation in the data and can be interpreted as the dominant factor that explains the structure in the data. Finally, the last column contains a Yes if either the prediction ellipses completely overlap or one ellipse is contained entirely within the other (in this case DC prediction ellipses are contained within the progressive disease ellipses).

Data Group	**n**	**#EV**	**#PCs**	**Cum var expl**	**Important variables**	**Prediction Ellipse**	**Same as full**
Blood	105 (83,22)	15	5	81.4%	LKafter, NPafter LK diff, NP diff, LK before	Almost completely overlap	Yes, exactly the same
White Blood Cells	111 (87,24)	9	3	90.5%	LKdiff, LK after, NPdiff, NPafter, LKbefore	Almost completely overlap	Yes, exactly the same
Kidney	139 (110,29)	12	5	80.2%	CRafter, GFRbefore, CRbefore CRdiff	DC within PD	No, representations and directions changed for a few, PD ellipse more elongated
Liver	160 (131,29)	15	4	82.8%	ALATdiff ASATdiff GGTafter ALATbefore	DC almost entirely within PD	barely, PC1 explains 37.2%, but for the rest, almost no difference
Nutrition	99 (78,21)	9	5	90.7%	Naafter, Kafter, Nabefore	DC circle, PD ellipse with overlap	No, directions have changed, same for ellipse
Inflammation	87 (69,18)	15	4	83.1%	SIIdiff, NLRdiff, PLRdiff	Overlapping ellipses in opposte directions	Yes, barely any difference. Scree plot %s might differ with + or - 0.1%
Patient Characteristics	179 (143,36)	4	3	99.9%	Weight, BMI, Age	Overlapping ellipses	Yes, barely any difference. Only difference in %s in screeplot
Tumor Markers	93 (74,19)	8	2	83.7%	CA19-9 after CEAdiag CA19-9before	DC completely within PD	Yes, barely any difference PC1 explains 68.4% only difference
All blood and tumor markers	57 (43,14)	34	10	80.5%	CEAbefore AFafter, GGTafter	DC more circular within PD	No, directions mirrored wrt PC2 and representations changed
All variables + Age + BMI + Differences	57 (43,14)	53	13	82.7%	ALATdiff, ASATdiff, ALATbefore	DC completely within PD ellipse	No, directions mirrored wrt PC1 and representations changed, ellipses also changed
Age + BMI + Differences	57 (43,14)	19	9	81.7%	ALATdiff, LKdiff ASATdiff NPdiff	DC ellipse within PD	No, directions changed slightly for some variables and also representation

Table 3.16: Overview table of PCA performed on the correlation matrices of the analyzed data sets after removal of outliers based on GFR< 30mL/min and BR> $50\mu mol$/L. EV = Explanatory Variables, PC = Principal Component.

4 | Methodology

4.1. Random Forest background information

The objective of this thesis is to develop a predictive model for disease or tumor response in PDAC patients undergoing FOLFIRINOX chemotherapy using the provided dataset. Artificial Intelligence (AI) and Machine Learning (ML) have gained significant attention in the medical domain. Selecting the appropriate ML model depends on various factors, such as the trade-off between underfitting and overfitting, the type of learning method (supervised or unsupervised), and the specific application context. Underfitting occurs when the ML model is overly simplistic relative to the complexity of the problem and the dataset, resulting in poor performance and a high bias. Conversely, overfitting arises when the model is excessively complex with numerous parameters relative to the problem's complexity and dataset size. This leads to excellent performance on the training data but poor generalization on the test set, characterized by high variance. The ideal scenario is to achieve a balance with low bias and low variance. In this case, we have labeled data and thus need for a supervised learning method for the classification of PDAC patients. We aim to find a model that exhibits good performance, ease of interpretability and implementation, and has no strong underlying assumptions. Among the available options, Random Forest stood out as the most suitable choice. This method effectively handles noisy data, is well-suited for large and heterogeneous datasets, and demonstrates robust classification performance. It also implicitly performs feature selection by ranking features during the classification process. Compared to more complex methods like Neural Networks or Deep Learning, Random Forests are easier to comprehend and require less computational power. Due to the small dataset size, deep learning models are also more prone to overfitting and regularization techniques or extensive hyperparameter tuning would be required. On the other hand, simpler methods such as k-nearest neighbors or support vector machines are prone to overfitting, sensitive to outliers, and face challenges in handling noisy or missing data. Furthermore, these often rely on strong assumptions, such as linear separability [55].

This section focuses on constructing a random forest model using the PREONPANC2-iKnowIT dataset, excluding missing values and identified outliers. The model generates a binary response (0 = disease control, 1 = progressive disease). To comprehend the model, it is important to establish background knowledge about decision trees and random forests. The following subsection provides this information along with mathematical details. Subsequently, a random forest will be trained and evaluated.

4.1.1 Decision Trees

Decision trees are a class of supervised learning algorithms used in machine learning. They recursively split a dataset into categories based on whether a condition is true or false until there are only pure leaf nodes left, which have data with only one type of class. There are two types of decision trees: classification and regression trees. While the former is used for classifying data, the latter is used for predicting numeric values. In this case, the objective is to predict the final response as either disease control (0) or progressive disease (1), making classification trees more suitable. Hence, the focus will be on classification trees. Both numeric and categorical data can be used to classify observations in classification trees. Furthermore, different numeric thresholds can exist for the same data. For example, CA19-9 levels < 80000 kU/L can be in one decision node, and CA19-9 levels < 5000 kU/L can be another decision node below that decision node. In general, the rule is moving left if a statement is *true* and right if it is *false.* The top node of a decision tree is called the root, the internal nodes are called branches or decision nodes, and the leaf nodes represent the final classified observations. An example of a decision tree can be found in Figure 4.1.

The decision tree algorithm creates a tree-like model that uses nodes for input features, branches for decisions, and leaves for class labels. The tree is built top-down, starting from the root node and recursively splitting it based on an input feature's value. Impurity measures like entropy or Gini impurity help identify the best splits. The goal is to minimize impurity by optimal data splitting. To classify new data, the decision tree is traversed from root to leaf. At each node, the algorithm compares the input feature's value to a threshold and follows the appropriate branch until a leaf with the predicted class label is reached. To understand how a decision tree works, consider a fruit dataset that needs to classify apples and pears based on features like color, shape, and weight. The first step is to choose a feature as the root node, such as color with two possible values: red and green. This creates two branches. Then, another feature is selected, like roundness for red fruits and pear-shape for green fruits, to further split the data. This process continues with additional features until reaching the leaf nodes, which represent the final decisions (apple or pear) based on fruit characteristics. Once the decision tree is trained on the dataset, it can be used to classify new fruits by traversing the tree from the root to the leaf node, which determines the final decision.

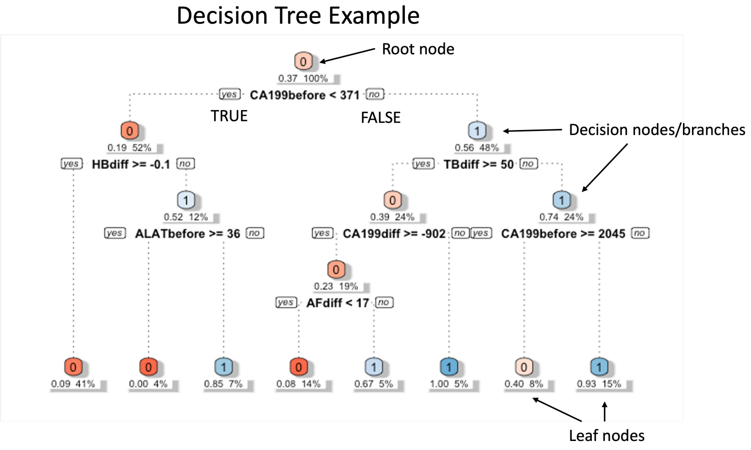


Figure 4.1: Example of a Decision tree created using the original training dataset, n=194, 0=Disease Control (n=120), 1=Progressive disease (n=74).

Decision trees have several advantages, such as interpretability, ease of use, and the ability to handle different types of data. However, they also have limitations, including overfitting, sensitivity to data changes, and handling complex relationships between features. To address these limitations, ensemble methods like random forests are commonly used. Random forests combine the predictions of multiple decision trees to improve accuracy and stability. In the next subsections, we will explore key concepts to better comprehend the construction of decision trees.

4.1.1.1 Entropy

Entropy is a measure of data impurity or randomness. In decision trees, it helps determine the feature for data splitting at each node. To elaborate, entropy quantifies the level of uncertainty or heterogeneity in a set of samples based on their classifications. In a binary classification scenario (which we have), entropy can be computed using Equation (4.1) as follows:

$$E(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2), \tag{4.1}$$

where S is the set of observations, p_1 is the proportion of samples belonging to one class and p_2 the proportion of samples belonging to the other class, under the condition that $p_1, p_2 \in [0, 1]$ and $p_1 + p_2 = 1$. To be more specific $p_1 = \frac{\text{number of samples in class 1}}{\text{total number of samples in the data}}$ and $p_2 = \frac{\text{number of samples in class 2}}{\text{total number of samples in the data}}$. When all the samples belong to the same class, the entropy is zero, indicating that the set is completely homogeneous and there is no uncertainty in classification. Conversely, if the samples are equally divided split between the two classes, the entropy is one. The higher the entropy value, the higher the level of disorder indicating a lower level of purity.

4.1.1.2 Gini Impurity

Another measure of impurity or randomness in a dataset is the Gini impurity. The Gini impurity of a set of observations (samples) is the probability of misclassifying a randomly chosen sample in that set, if it were labeled randomly according to the distribution of labels in the set. For the binary classification problem, Gini impurity can be calculated as:

$$Gini(S) = 1 - p_1^2 - p_2^2 \tag{4.2}$$

where S, is the set of samples, p_1 is the proportion of samples belonging to class 1 and p_2 the proportion of samples belonging to class 2, under the conditions that $p_1, p_2 \in [0, 1]$ and $p_1 + p_2 = 1$. When all the samples belong to the same class, the Gini impurity is zero, indicating that the set S only contains observations from one class. On the other hand, if the samples are split 50/50 between the two classes, the Gini impurity is 1, indicating the highest degree of impurity or randomness in the dataset.

Gini impurity can also be used to determine the best feature to split the data on at each node of the decision tree. In general, the feature with the lowest Gini impurity is chosen as the split feature. The formula for calculating the Gini impurity of a split on feature A is:

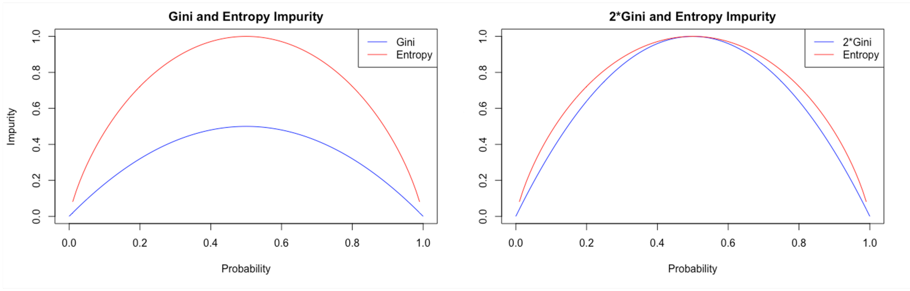
$$Gini_A(S) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Gini(S_v) \tag{4.3}$$

where S is the set of samples, A is the feature or variable being considered, S_v is the subset of S in which feature A has value v, and $Values(A)$ is the set of possible values that feature A can take. To calculate the Gini impurity of the split on feature A, take the weighted sum of the Gini impurities of its subsets S_v, where the weight is the proportion of samples in S_v with respect to the total number of samples in S. Then the feature with the lowest Gini impurity is selected as the split feature at that node.

4.1.1.3 Difference Gini impurity and entropy

The main difference between Gini impurity and entropy is the way they measure the impurity of a split in a decision tree. The Gini impurity measures the probability of misclassifying a randomly chosen element if it is randomly labeled according to the distribution of labels in the set. It is based on the idea of the probability of two random elements in the same set being of different classes. In addition, it is computationally faster to compute the Gini impurity than entropy because it does not involve logarithmic calculations. Entropy, on the other hand, measures the amount of uncertainty or randomness in a set of data. It calculates the amount of information needed to convey the class distribution of the data. Moreover, entropy tends to penalize the tree for having too many branches, as it gives more weight to smaller subsets of data with fewer classes.

When looking at the graphs, the Gini index has values inside the interval [0, 0.5] while Entropy on the other hand has values in their interval [0,1] by definition. Actually, $Gini(S) \approx \frac{1}{2}E(S)$. This is illustrated in Figure 4.2 using the two class case where p = probability of class 1 and $1-p$ = probability of class 2. Note that $1-p^2-(1-p)^2 = 1-p^2-1+2p-p^2 = 2p-2p^2 = 2p(1-p)$.



(a) (b)

Figure 4.2: Graphs of the Entropy and Gini impurity with the probability of class 1 p on the x-axis for the two class case: (a) both original graphs $Gini(p) = 2p(1-p)$ and $E(p) = -plog_2(p) - (1-p)log_2(1-p)$ (b) 2×Gini and the original Entropy graph $2 \times Gini(p) = 4p(1-p)$ and $E(p) = -plog_2(p) - (1-p)log_2(1-p)$.

There is no consensus on which impurity measure is better as it depends on the problem and dataset. Both measures yield similar results in practice, and the choice is often based on computational efficiency. In general, Gini impurity is preferred for large datasets or complex decision trees, where efficiency is crucial. It favors balanced partitions and is less sensitive to outliers compared to entropy, which considers the log-likelihood of the class probabilities.

4.1.1.4 Information Gain

Information gain is a metric used in decision tree algorithms to assess the usefulness of a feature in predicting the target variable. It quantifies the reduction in entropy, Gini impurity, or other impurity measures achieved by splitting the data on a specific feature at a particular level. Higher information gain corresponds to lower entropy or Gini impurity, indicating more useful splitting. Conversely, lower information gain indicates higher impurity. Mathematically, information gain can be defined as the difference between the entropy of the original set of samples and the weighted average of the entropy of its subsets. The subsets are formed based on the possible values that the feature being considered can take. The feature with the highest information gain is chosen as the split feature at that node of the decision tree. The information gain is calculated using the following equation:

$$InformationGain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{4.4}$$

where S is the set of samples, A is the feature being considered, S_v is the subset of S in which feature A has value v, and $Values(A)$ is the set of possible values that feature A can take. In summary, information gain is a measure of the reduction in the disorder in the target variable or class given additional information provided by a particular feature. It is a useful tool in determining which variable to split the data on at each node of the decision tree, with the feature that gives the highest information gain being selected as the split variable.

Calculation Example of classifying pets

To clarify the usage of Gini impurity, entropy, and information gain, consider an example of a decision tree classifying pets as either “dog" or “cat" based on their characteristics. We have a dataset that includes information on weight, fur length, and tail presence. To build the decision tree, we need to determine the root node, which can be determined by calculating the Gini impurity and information gain. Let us focus on the weight feature. Suppose the dataset has 100 pets, with 60 dogs and 40 cats. We split the dataset based on weight, creating two subsets: pets under 10 pounds and pets over 10 pounds. The subsets are as follows: 70 pets under 10 pounds (40 dogs, 30 cats) and 30 pets over 10 pounds (20 dogs, 10 cats). To calculate the Gini impurity of the weight feature, we first need to calculate the Gini impurity of the original dataset, denoted as $G_{original}$:

$$G_{original} = 1 - (\frac{60}{100})^2 - (\frac{40}{100})^2 = 0.48 \tag{4.5}$$

Then calculate the Gini impurity of the two subsets after splitting the dataset based on weight, call these G_{w_1}, G_{w_2} :

$$G_{w_1} = 1 - (\frac{40}{70})^2 - (\frac{30}{70})^2 \approx 0.49 \tag{4.6}$$

$$G_{w_2} = 1 - (\frac{20}{30})^2 - (\frac{10}{30})^2 \approx 0.44 \tag{4.7}$$

To calculate the information gain of the weight feature, take the weighted average of the Gini impurities of the two subsets and subtract it from the Gini impurity of the original dataset:

$$Information\ gain = 0.48 - (\frac{70}{100} * 0.49 + \frac{30}{100} * 0.44) \approx 0.048 \tag{4.8}$$

In this example, the information gain for weight was found to be 0.048, which suggests that weight may not be the best feature to use as the root node. This process can be repeated for the other features (fur length and tail), and the feature with the highest information gain is chosen as the root node of the decision tree. This is because the feature with the highest information gain contributes the most to the classification of the dataset. Note that instead of Gini impurity, entropy or any other impurity measure can also be used to calculate the information gain.

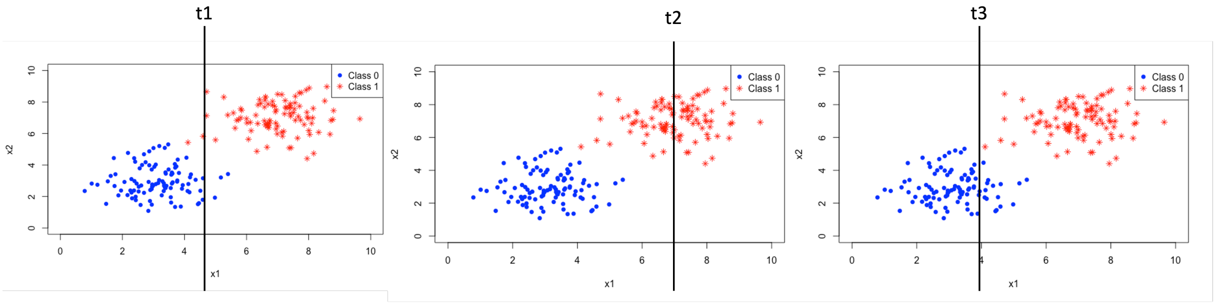
4.1.1.5 Pruning

Another important concept to discuss in decision tree learning is pruning. Pruning is a technique used to address the problem of overfitting in decision trees. Overfitting occurs when a decision tree is too complex and becomes too closely fit to the training data, resulting in poor performance on new data. Pruning works by selectively removing branches or nodes from a decision tree that do not significantly contribute to the overall accuracy of the tree, essentially “cutting" the tree at a certain depth. There are two primary approaches to pruning: pre-pruning and post-pruning. Pre-pruning involves setting a stopping criterion based on an impurity measure, such as Gini impurity or entropy, or an error measure before the tree is fully grown. This can be an effective way to prevent overfitting and reduce the size of the tree, but it can also lead to underfitting if the stopping criterion is overly strict. In contrast, post-pruning involves growing the decision tree to its full size and then removing branches or nodes that do not improve the performance of the tree on a validation set. This approach is often more effective than pre-pruning, as it allows the tree to reach its maximum potential size before removing any unnecessary branches or nodes.

4.1.1.6 Choosing the variables and thresholds

In decision tree algorithms, the variable and threshold selection for splitting rely on information gain, which is based on the reduction of an impurity measure like entropy or Gini. The objective is to find the best combination that effectively partitions the data, maximizing class classification. To illustrate this concept, we can refer to the graphs shown in Figure 4.3. These graphs depict the distribution of data points, with blue dots representing one class and red stars representing another class. The decision boundary is determined by a specific threshold value applied to the variable x_1 represented on the x-axis. By visually examining the graphs, it appears that the split at $x_1 = t_3$ provides the most favorable classification outcome. This split results in only blue dots on the left side, translating to a 'pure node' in a decision tree, containing only blue points, while maintaining the smallest number of both red and blue points on the other side. In other words, if a randomly chosen point has a value of $x_1 < t_3$, it is likely to belong to the blue class. Similarly, if a randomly chosen point has a value of $x_1 > t_2$, it is likely to belong to the red class. However, such a clear inference cannot be made at t_1. To make an informed decision about the optimal threshold, we evaluate the distribution of data points on either side of each threshold and compare it to the distribution before the split. The objective is to identify the threshold that yields the most favorable data distribution, ideally placing a majority of points from one class (blue in this case) on one side while minimizing the presence of the other class (red in this case) on the other side of the split. Mathematically, this involves assessing the impurity measure of the data prior to the split and selecting the threshold that maximizes the reduction in impurity. In the decision trees generated in the coming models, this impurity measure will be represented by the Gini impurity.

Therefore, the selection of the threshold in decision tree algorithms involves an evaluation of the impurity measures before and after the split, aiming to identify the threshold that results in the most significant improvement in data purity and classification effectiveness.



(a) (b) (c)

Figure 4.3: Illustration of defining the threshold for the 'best' split of the two groups of data. (a) t1 = threshold 1 (b) t2 = threshold 2 (c) t3 = threshold 3.

To further illustrate this, consider the following dataset given in Table 4.1 with two explanatory variables $X1$ and $X2$ and response y. The datasets has four points for class 0 and four for class 1. Thus the Gini impurity before the split can be calculated as given in Equation (4.9):

$$\begin{aligned} G(\text{before split}) &= 1 - (\frac{n_0}{N})^2 - (\frac{n_1}{N})^2 \\ &= 1 - (\frac{4}{8})^2 - (\frac{4}{8})^2 \\ &= 1 - \frac{1}{4} - \frac{1}{4} = 0.5 \end{aligned} \tag{4.9}$$

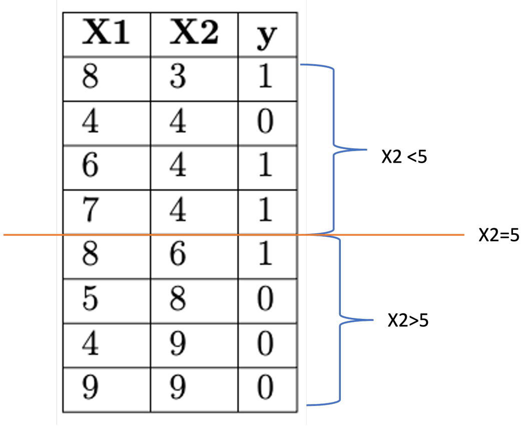
where n_0 denotes the number of points of class 0 ($n_0 = 4$ in this case) and n_1 the number of points of class 1 ($n_1 = 4$ in this case) and N the total number of points ($N = n_1 + n_2 = 8$ in this case). If we consider the split at a threshold of $X2 = 5$ the data distribution would look like Figure 4.4. Then the Gini impurity for both the splits is given in Equations (4.10)

X1	X2	y
8	3	1
4	4	0
6	4	1
7	4	1
8	6	1
5	8	0
4	9	0
9	9	0

Table 4.1: Example data with X2 sorted by magnitude.

and (4.11).

$$G(X2 < 5) = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375 \tag{4.10}$$



X1	X2	y
8	3	1
4	4	0
6	4	1
7	4	1
8	6	1
5	8	0
4	9	0
9	9	0

Figure 4.4: Example data with X2 sorted by magnitude and the boundaries for splitting at $X2 = 5$.

$$G(X2 > 5) = 1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.375 \tag{4.11}$$

Thus, the overall information gain is:

$$IG(G) = \Delta G = G(\text{before split} - \frac{G(X2 < 5) + G(X2 > 5)}{2} = 0.5 - \frac{0.375 + 0.375}{2} = 0.125 \tag{4.12}$$

This can be done for all the possible thresholds of X2. Four examples are listed below in Table 4.2, but note that an algorithm would range over the entire range of possible values for a variable (in this case X2) from minimum to maximum value. This results into different thresholds with different calculated gains. The same process can repeated for the other feature, X1 and an example table is provided in Table 4.3.

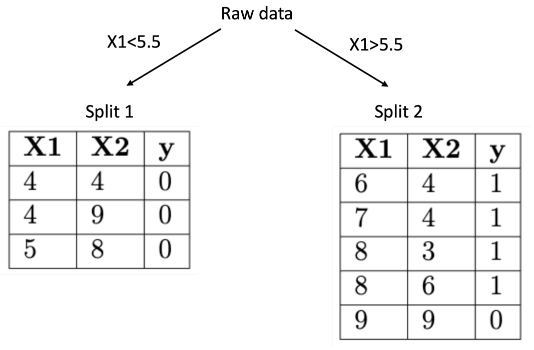
Threshold	**G(before split)**	**G(< threshold)**	**G(> threshold)**	**Gain**
3.5	0.5	0	0.49	0.255
5.0	0.5	0.375	0.375	0.125
7.0	0.5	0.320	0	***0.340***
8.5	0.5	0.528	0	0.236

Table 4.2: Example table of threshold values for variable X2, the calculated Gini impurity before the split (G(before split)), the Gini impurity less than the given threshold (G< threshold), the Gini impurity bigger than the given threshold (G> threshold) and the information gain (Gain).

Threshold	**G(before split)**	**G(< threshold)**	**G(> threshold)**	**Gain**
4.5	0.5	0	0.444	0.278
5.5	0.5	0	0.32	***0.340***
6.5	0.5	0.375	0.375	0.125
7.5	0.5	0.48	0.444	0.038
8.5	0.5	0.49	0	0.255

Table 4.3: Example table of threshold values for variable X1, the calculated Gini impurity before the split (G(before split)), the Gini impurity less than the given threshold (G< threshold), the Gini impurity bigger than the given threshold (G> threshold) and the information gain (Gain).

In this specific example, the highest information gain value (0.340) remains the same whether the data is split at $X1 = 5.5$ or $X2 = 0.340$. When faced with multiple features offering the same maximum gain, the feature located furthest to the left (in the original data) is chosen. Following this principle, the optimal splitting point for this example is at $X1 = 5.5$. Thus, by setting our decision node at $X1 = 5.5$, we can construct a decision as shown in Figure 4.5. Then, repeat these steps for each split dataset until all divisions consist solely of pure samples from a single class, indicated by an impurity measure of zero.



X1	X2	y
4	4	0
4	9	0
5	8	0

X1	X2	y
6	4	1
7	4	1
8	3	1
8	6	1
9	9	0

Figure 4.5: Example decision tree after the first split based on $X1 < 5.5$.

4.1.2 Random Forest

One approach for improving decision trees is to use a random forest. A random forest is an ensemble of decision trees, where each tree is trained on a randomly selected subset of the original dataset (with replacement). This process, known as bootstrapping, ensures that the model does not rely solely on the same data for every tree, thereby mitigating its sensitivity to the specific characteristics of the original training data. In order to make the final class label prediction, majority voting among the generated forest of decision trees is used. This averaging of the results decreases the variance in the classification model. Together, the two techniques are called ‘bagging’, which is short for bootstrap aggregation and explained in more detail below. Aside from bagging, a random subset of features is used to split the data at each node within the decision trees. This practice effectively diminishes the correlation among the trees. The rationale behind this feature selection process is to alleviate the potential problem of having identical trees possessing identical decision nodes when all features are used. Consequently, the trees would act in a similar manner, increasing the variance of the model. To be specific, each new tree is built using a random sample of say $m < p$ features chosen as split candidates from the full set of p features. The split in turn only uses one of those m features. This results in a forest of decorrelated trees. As an example, suppose that there is one very strong feature in the dataset, along with a number of other moderately strong features. Then in the collection of bagged trees, most or all of the trees will use this strong feature in the top split. In our dataset, the most likely feature that will be used in the root node will be CA19-9 value. As a consequence, all of the bagged trees will look quite similar to each other. Hence the predictions from the bagged trees will be highly correlated and averaging these results will not lead to a reduction in variance. Therefore, random forests force each decision tree to consider only a subset of the features, so on average $\frac{p-m}{p}$ of the trees will not even consider the strong feature and so other features will have more of a chance. This makes the trees less correlated, thereby making the average of the resulting trees less variable and more reliable. However, since a substantial number of trees are generated within the random forest, ‘suboptimal’ predictions tend to cancel each other out. By combining bagging and random feature selection, random forests create trees that are trained on different subsets of the data with different subsets of features. As a result, each tree in the random forest learns different aspects of the data and makes decisions based on different sets of features. This diversity helps to reduce the correlation between the trees and reduces the tendency for all trees to make similar predictions. Moreover, each tree is trained on $\approx \frac{2}{3}$ of the bootstrapped sample, called the ‘in bag’ sample and performance is tested on the remaining Out-Of-Bag (OOB) sample. The optimal splitting criterion in each decision tree is based on an impurity measure like Gini or entropy, as explained in the previous subsection. There are some common stopping rules to prevent overfitting in random forests, namely:

1. n_{tree} = Number of trees in the forest
2. m_{try} = Number of features considered for splitting a node
3. max_{depth} = Maximum number of levels in each decision tree
4. $min_{samples-split}$ = Minimum number of data points in a node before the node is split
5. $min_{samples-leaf}$ = Minimum number of data points allowed in a leaf

Subsequently, the test dataset is used to make predictions and these predictions are averaged at the end to result in a final prediction, which in classification is determined using majority voting. A pseudo-code of the random forest algorithm is provided in Algorithm 2 and the algorithm is visualized in Figure 4.6. In the random forests build in Section 5 only the first two are considered, namely n_{tree} and m_{try}. This because of small training and test sizes.

4.1.2.1 Bagging

Bagging, short for bootstrap aggregation, is a procedure that reduces the variance of a statistical learning method by training multiple models on different subsets of the training data. These models' predictions are then aggregated to make a final prediction. In bagging, each subset is created by sampling with replacement from the original training dataset. Random forest is an extension of bagging that further enhances model diversity. In addition to creating models using different subsets of the training data, random forest also randomly selects a subset of features for each model. The number of selected features, denoted as $m < p$, is less than the total number of features (p). When $m = p$, random forest and standard bagging are equivalent. Thus, the key distinction between bagging and random forest lies in the random selection of feature subsets in addition to generating bootstrapped samples from the original training data.

4.1.2.2 Out-Of-Bag Error (OOB error)

The fundamental principle behind bagging involves fitting models (in this case decision trees) repeatedly to subsets of the observations using bootstrapping with replacement. This implies that some observations in the original training dataset may not be included in any of the subsets used to train a specific model, and these observations are called out-of-bag (OOB) observations for that model. The OOB error estimate is a way to assess the performance of a model without requiring a separate validation set or the use of cross-validation. The OOB error is calculated by predicting the outcome of each observation in the original dataset using only the trees that were built without incorporating that observation in their bootstrap sample. It is expressed as the proportion of misclassified observations. In an ideal scenario, the OOB sample comprises around 37% ($\approx \frac{1}{3}$) of the total training data. To explain its functioning in more detail, consider each observation in the original dataset. If an observation was not part of the bootstrap sample used to train a specific tree, it is labeled as an “out-of-bag" observation for that tree. The OOB observations are subsequently used to assess the prediction capability of the corresponding tree. The prediction process involves passing the out-of-bag observation through the tree and consolidating the outcomes based on the majority vote (for classification problems) or averaging (for regression problems) of predictions from all trees that did not use that particular observation during training. Then, the OOB error is computed by comparing the predicted values for the out-of-bag observations with their true values, with the use of the appropriate error metric such as misclassification rate or mean squared error. It provides an unbiased estimate of the model's performance, facilitates accuracy assessment, and simplifies model training by eliminating the need for a separate validation set.

Algorithm 2 Random Forest Algorithm (Classification)

```
Require: Training dataset $D$, number of trees $T$, number of features $p$, number of data points $n$
1: for $t = 1$ to $T$ do
2:     Randomly select a subset of of the training dataset of $D$ with replacement, this is called bootstrapping. This creates a new dataset of the same size $|D|$ as the original, but with some repeated observations and other omitted. Call this bootstrap sample $D_t$.
3:     Randomly select a subset $m \subset p$ features from the Feature set for each tree. Generally $m = \sqrt{p}$, the square root of the total number of features is used as the maximum number of features to select.
4:     Train a decision tree using each bootstrap dataset $D_t$ independently using the ‘in bag’ samples and the selected $m$ features. Each tree is grown to the largest extent possible and there is no pruning. Performance is tested on the OOB samples and reported in the OOB-estimate.
5:     Store the trained tree in the Random Forest model $RF$
6:     Repeat steps 1-3 to create the Random Forest $RF$ (default is 500 trees) for each tree.
7:     To make a prediction for a new observation, pass the test set through each decision tree in the forest and store the predicted class label. Then take the majority vote over all the predictions for the final prediction (in the classification case). This process of combining results from multiple models is called aggregation.
```

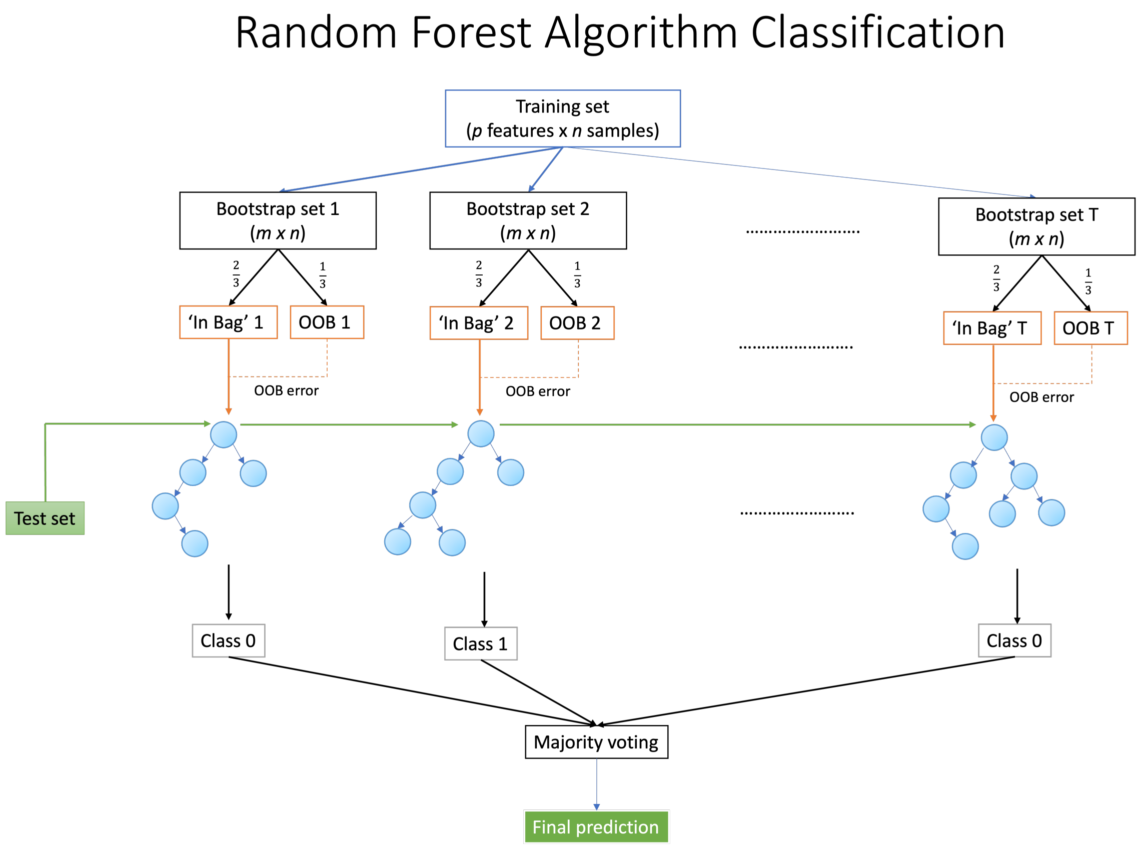


Figure 4.6: Random forest algorithm for classification with a dataset D containing n observations, p features with a chosen number of T trees and $m < p$ features to consider per tree.

4.1.3 Imbalanced Data

The dataset used in this study is imbalanced, with a majority of patients (173/216) categorized as DC (0) and 43/216 labeled as having PD (1). This imbalance poses a challenge for classifiers like random forest, which tend to prioritize accuracy. As a result, the random forest classifier may classify all observations as DC to achieve a high overall accuracy. However, this approach is not ideal, as it fails to accurately identify patients with PD (the minority class) who may not respond well to chemotherapy. Misclassifying DC patients as having PD should also be avoided, as it could prevent them from receiving potentially beneficial treatment. On the other hand, misclassifying PD patients as having DC is also costly, since we give ineffective treatment to these patients. Therefore, different approaches to address the issue of imbalanced data will be explored in the following section.

- **Under-sampling or downscaling**: This technique involves randomly removing instances that belong to the over-represented class. As an example, in this case, patients classified as DC are the over-represented or majority class. By reducing the number of DC samples, more emphasis is placed on accurately classifying individuals with PD. However, a significant drawback of random under-sampling is the potential loss of valuable information, particularly when the minority class (PD) is considerably smaller relative to the majority class.
- **Over-sampling or upscaling**: Similar to under-sampling, over-sampling aims to address class imbalance by randomly duplicating instances from the minority class. By adding more copies of PD patients (in this case), the classifier focuses on correctly classifying this group. However, a significant drawback of over-sampling is the increased risk of overfitting, as it generates exact replicas of the minority class samples, potentially leading to an overly optimistic estimation of model performance.
- **Combination of under- and over-sampling**: This approach involves applying a balanced combination of under-sampling and over-sampling techniques. It includes a moderate amount of oversampling for the minority class and a modest amount of undersampling for the majority class. This strategy aims to improve the bias towards the minority class while also reducing the bias towards the majority class, striking a balance between preserving information and addressing class imbalance.
- **Cost-sensitive learning (CSL)**: Cost-sensitive learning involves assigning weights or costs to misclassified instances based on their class membership. It highlights the imbalanced learning problem by utilizing cost matrices that describe the cost of misclassification. For example, misclassifying a healthy person as unhealthy might be considered less costly than misclassifying an ill patient as healthy. In the context of random forests, these weights can be incorporated into the process of finding splits and the voting mechanism for each classification, effectively adjusting the importance of different instances based on their associated costs.

These approaches offer solutions to address the dataset's imbalance and enhance classification for PD patients while considering the risk of misclassifying DC patients. The selection of a specific approach depends on dataset characteristics and the balance between preserving information, mitigating overfitting, and effectively handling class imbalance. In the upcoming section, all these methods will be employed and explored in training the random forest. Furthermore, the ROSE and Complete Class inclusion methods will also be utilized and discussed below.

4.1.3.1 Random Over-Sampling Examples (ROSE)

In addition to the aforementioned techniques, Random Over-Sampling Examples (ROSE) is another method that can be used to tackle imbalanced datasets. ROSE aims to address the class imbalance by generating synthetic samples based on the existing minority class samples. This technique involves creating a modified dataset that combines the original minority class samples with the newly generated synthetic samples, which can be considered as oversampling with random noise. The synthetic samples are generated by randomly selecting existing minority class samples and introducing slight modifications while preserving the overall characteristics of the minority class. These modifications can include variations in feature values or the introduction of random noise to the samples. The primary goal of the ROSE method is to increase the representation of the minority class in the dataset, thus mitigating the class imbalance issue through the generation of synthetic samples.

However, ROSE's effectiveness depends on the quality of the synthetic samples generated and their representation of the minority class. While ROSE can be effective in some scenarios, it is not a universal solution and its performance may vary based on dataset characteristics and the chosen machine learning algorithm. Evaluation, trade-offs, and limitations should be considered when using ROSE for imbalanced datasets. The ROSE algorithm can be summarized as follows:

1. *Class imbalance ratio calculation:* The initial step involves calculating the imbalance ratio (IR) between the majority and minority classes. This ratio is determined by dividing the number of majority class samples by the number of minority class samples.
2. *Determination of synthetic sample count:* Using the computed IR, the desired number of synthetic samples for the minority class is established. This count is obtained by multiplying the number of original minority class samples by (IR - 1).

3. *Nearest neighbor identification:* For each minority class sample, a process is undertaken to identify its k nearest neighbors within the minority class. The parameter k is user-defined and determines the number of neighbors considered.
4. *Synthetic sample generation:* With the nearest neighbors identified, a random selection is made for each minority class sample. The chosen neighbor acts as a base for generating a synthetic sample. Slight modifications are applied to the feature values of the base sample to create the synthetic sample. Techniques such as adding random noise or employing interpolation methods are commonly used for these modifications.
5. *Iterative repetition:* Steps 3 and 4 are repeated for each minority class sample until the desired count of synthetic samples is generated.
6. *Integration of original and synthetic samples:* The final step involves combining the original minority class samples with the newly generated synthetic samples, thereby forming a modified dataset. Through this integration, a more balanced representation of the minority class is achieved.

Mathematically, let the binary response variable be denoted as y_i, where $i \in \{0, 1\}$ represents the class labels belonging to the set $\{Y_0, Y_1\}$, and $\mathbf{x} \in \mathbb{R}^d$ represents a vector of predictors for a dataset consisting of n subjects indexed as $i, i = 1, \ldots, n$, with an unknown probability density function $f(\mathbf{x})$. Let the training set of size n is denoted as $\mathbf{T}_n$ defined by the pairs $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. Also let the number of observations in each class Y_0 and Y_1 be denoted by $n_j < n, j \in \{0, 1\}$. The ROSE procedure for generating a new artificial example involves the following steps:

1. Select $y^* = Y_j$ with probability π_j. In this step, we choose the class label for the artificial example. The selection is made based on the probabilities associated with each class. For example, if π_0 is the probability of class Y_0 and π_1 is the probability of class Y_1, we select Y_0 or Y_1 with their respective probabilities.
2. Select $(\mathbf{x}_i, y_i) \in \mathbf{T}_n$, such that $y_i = y^*$ with probability $\frac{1}{n_j}$. Once we have determined the class label y^*, we randomly choose an example from the original dataset $\mathbf{T}_n$ that belongs to the selected class $y_i = y^*$. The selection is done uniformly at random from the examples of the selected class. For instance, if we chose Y_0 in the previous step, we randomly pick an example $(\mathbf{x}_i, y_i)$ from $\mathbf{T}_n$ where $y_i = Y_0$.
3. Sample $\mathbf{x}^*$ from $K_{H_j}(\cdot, \mathbf{x}_i)$ with K_{H_j} a probability distribution centered at $\mathbf{x}_i$ and covariance matrix H_j. Finally, we generate a new artificial example $\mathbf{x}^*$ by sampling from a probability distribution K_{H_j}, which is centered at the selected example $\mathbf{x}_i$. The distribution KH_j is defined by a covariance matrix H_j specific to the selected class y^*. This step introduces some variation and randomness to create a new example $\mathbf{x}^*$ that resembles the original example $\mathbf{x}_i$.

By repeating these steps for multiple iterations, we can generate a set of artificial examples for the minority class, thereby balancing the class distribution in the dataset. Essentially, the ROSE method selects an observation that belongs to one of the two classes and generates new samples $(\mathbf{x}^*, y^*)$ within its neighborhood, where the width of the neighborhood is determined by H_k, and the shape is determined by the contour sets of K. Given the selection of the class label Y_j, the generation of new samples from Y_j using ROSE corresponds to generating data from a kernel density estimate of $f(\mathbf{x}|Y_j)$, where the kernel is represented by K and the smoothing matrix is denoted as H_j [56]. Repeating steps 1-3 m times produces a new synthetic training set $\mathbf{T}^*_m$ of size m where the imbalance level is defined by the probabilities π_j. The size m can be set to the original training set size n or any other choice. Note that if $\pi_j = \frac{1}{2}$ then approximately the same number of samples belong to each of the two classes. Note that when $\mathbf{H}_j \to 0$, ROSE 'converges' to a standard combination of over- and under-sampling [56].

4.1.3.2 Cost sensitive learning

When dealing with imbalanced datasets, cost-sensitive learning approaches can be used to address the bias caused by the skewed class distribution. In the context of random forests, cost-sensitive learning involves adjusting misclassification costs associated with different classes during training and prediction. Misclassification costs assign penalties or weights to different types of misclassifications based on their consequences. By incorporating misclassification costs, a classification model can optimize its predictions considering the specific costs associated with different types of misclassifications. In binary classification, a 2×2 misclassification cost matrix is used to assign costs to the four possible outcomes: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). An example of a misclassification cost matrix is provided in Table 4.4. In this matrix, the costs are specified as follows:

- c_{00}: The cost of misclassifying an actual Class 0 instance as Class 0 (TP).
- c_{01}: The cost of misclassifying an actual Class 0 instance as Class 1 (FP).
- c_{10}: The cost of misclassifying an actual Class 1 instance as Class 0 (FN).
- c_{11}: The cost of misclassifying an actual Class 1 instance as Class 1 (TN).

Customizable costs based on domain knowledge can be assigned in cost-sensitive learning. Misclassification costs enable optimization of predictions by considering the total cost rather than overall accuracy. For example, in a medical diagnosis scenario, misclassifying a patient with disease as non-disease (FN) might have more severe consequences than misclassifying a non-disease patient as having a disease (FP). Therefore, generally higher costs are assigned to costly misclassifications, prioritizing their minimization. In imbalanced data, the minority class is typically assigned a higher misclassification cost for more accurate predictions. The objective is to optimize a cost-based criterion, such as weighted misclassification cost or performance measures like F1 score or area under the Precision-Recall curve. In random forests this can be achieved through class weights or modified splitting criteria.

Once a misclassification cost matrix is determined, we can for instance find the 'optimal' threshold p^* can to classify an observation x as either 0 or 1. In this case we let $p = \mathbb{P}(j = 0|x)$, the probability of being classified as class 0 given an observation x. This p^* can be determined as follows. Consider the cost matrix $c = (c(i, j)(x))$, where $c(i, j)(x)$ represents the cost of classifying instance x from class j as class i given in Table 4.4. For simplicity, we assume constant and class-dependent cost functions.

		Predicted Class	
		Class 0	**Class 1**
Actual Class	**Class 0**	$C(0,0) = c_{00}$	$C(0,1) = c_{01}$
	Class 1	$C(1,0) = c_{10}$	$C(1,1) = c_{11}$

Table 4.4: Misclassification Cost Matrix

Naturally, the following two conditions (also called the "reasonableness" conditions) have to be satisfied [57]:

$$\begin{aligned} &(1)c_{01} > c_{00} \\ &(2)c_{10} > c_{11} \end{aligned} \tag{4.13}$$

Then a natural question arises:
Considering the cost matrix presented in Table 4.4 and a function $f : X \to [0, 1]$ that models the probability $\mathbb{P}(y = 1|x)$, which criteria should be used to determine whether x should be classified as class 0 or 1?
Intuitively, we classify x as 1 if the expected cost of classifying it as 1 is lower than that of classifying it as 0. In other words, the decision rule can be defined as follows:

$$\mathbb{P}(j = 0|x)c_{10} + \mathbb{P}(j = 1|x)c_{11} \leq \mathbb{P}(j = 0)c_{00} + \mathbb{P}(j = 1|x)c_{01} \tag{4.14}$$

if we write $p = \mathbb{P}(j = 1|x)$, we obtain:

$$(1 - p)c_{10} + pc_{11} \leq (1 - p)c_{00} + pc_{01} \tag{4.15}$$

If we have equality between the expected costs, it implies that predicting either class would be equally optimal. Therefore, we need to determine the optimal threshold p^* that ensures the expected values are equal. Mathematically, this can be expressed as:

$$(1 - p^*)c_{10} + p^*c_{11} = (1 - p^*)c_{00} + p^*c_{01} \tag{4.16}$$

solving this for p^* gives,

$$p^* = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}} \tag{4.17}$$

This result was proven by Elkan in [57] in more detail. In practice, this threshold is taken to be 0.5 in general.

4.1.3.3 Complete Class Inclusion

The 'Complete Class Inclusion' (CII) method addresses dataset imbalance by including the entire minority class in the training dataset. This helps reduce bias towards the majority class and improves the prediction of the minority class. By retaining all instances of the minority class during training, valuable information specific to that class is preserved. This approach can be seen as a form of 'oversampling' but does not necessarily achieve a 50/50 ratio between the majority and minority classes, as observed in traditional oversampling techniques. CII offers advantages such as exposure to the full set of minority class instances, enabling the model to learn distinctive patterns and establish effective decision boundaries. This comprehensive learning improves classification performance by capturing complex relationships and patterns between the minority class and features. This is especially useful when dealing with a very small minority set, as many machine learning models have troubles learning from a small sample. As a consequence, incorporating the minority class in training enhances the model's generalization ability to unseen samples during testing. However, CII alone may not fully address class imbalance. The biggest disadvantage is the risk of overfitting and not having a totally independent test set to validate the model performance on. Additional strategies like oversampling, undersampling, or adjusting class weights may be required to further balance the class distribution and improve overall model performance.

4.1.4 Evaluation Metrics

In the provided random forest model, various evaluation or performance metrics are used. In this subsection a brief explanation on the interpretation of these metrics is provided in the context of the random forest model. Note that many metrics can be generalized.

4.1.4.1 Confusion Matrix

The confusion matrix is a common tool for evaluating machine learning classification problems. It is a table that summarizes the predicted and actual values. It includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From the confusion matrix, various performance metrics can be calculated, such as recall, precision, specificity, and accuracy.

- True Positive (TP): The model correctly predicted a value as true, and it is indeed true (correctly classified).
- True Negative (TN): The model accurately predicted a value as negative or false, and it is indeed negative or false (correctly classified).
- False Positive (FP) or Type I error: The model incorrectly predicted a value as positive or true, but in reality, it is negative or false.
- False Negative (FN) or Type II error: The model erroneously predicted a value as negative or false, but in reality, it is true or positive.

Consider the example of a confusion matrix depicted in Figure 4.7. Formulas for computing different performance measures are provided below:

- **Recall:** Recall is also referred to as sensitivity, detection rate, or true positive rate (TPR), is calculated using Equation (4.18) and a higher recall value is more desirable.

$$TPR = Recall : \frac{TP}{TP + FN} \tag{4.18}$$

- **Precision:** Precision is also known as positive predictive value (PPV), is determined by Equation (4.19). Precision signifies the proportion of predicted positives that are truly positive. Hence, a higher precision value is preferable.

$$PPV = Precision : \frac{TP}{TP + FP} \tag{4.19}$$

- **Accuracy (ACC):** Accuracy is defined by Equation (4.20). It reflects the proportion of correctly classified instances over the total number of instances and provides a general measure of the model's overall correctness. However, accuracy alone might not be sufficient, especially in the presence of class imbalance.

$$ACC = Accuracy : \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.20}$$

- **The F1-score:** The f-1 score denoted as the harmonic mean of precision and sensitivity, is given by Equation (4.21). Comparing models with low precision and high recall, or vice versa, can be challenging. The F1-score serves as a composite metric that balances both precision and sensitivity. The highest achievable score is 1.0, while the lowest is 0. A value closer to 1 indicates a better-performing model. when dealing with imbalanced class distributions or when the costs associated with false positives and false negatives significantly differ, the F1-score is particularly valuable .

$$\text{F1 score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4.21}$$

- **Specificity or True Negative Rate (TNR):** Specificity or True Negative Rate measures the proportion of true negatives (class 1 predicted as class 1) out of all actual negative instances (class 1), as given in Equation (4.22).

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{4.22}$$

- **Negative Predictive Value (NPV):** Negative Predictive Value represents the proportion of true negatives (class 1 predicted as class 1) out of all instances predicted as negative (class 1).

$$NPV = \frac{TN}{TN+FN} \tag{4.23}$$

- **Prevalence:** Prevalence represents the proportion of total positive instances in the dataset.

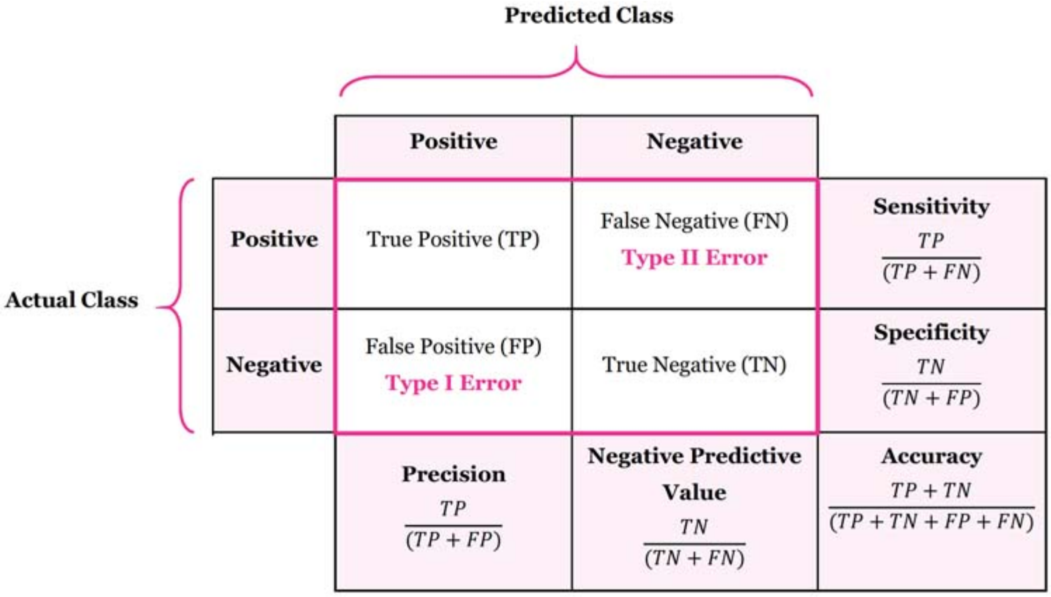
$$Prevalence = \frac{P}{P+N} \tag{4.24}$$

- **Detection Prevalence:** Detection prevalence represents the proportion of instances predicted as positive by the model.

$$\text{Detection Prevalence } = \frac{TP+FP}{TP+FP+FN+TN} = \frac{P}{P+N} \tag{4.25}$$

- **Balanced Accuracy:** Balanced accuracy is the average of sensitivity and specificity. Higher balanced accuracy values indicate better overall classification performance, taking into account both the positive and negative instances.

$$BA = \frac{TPR+TNR}{2} \tag{4.26}$$



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 4.7: Confusion matrix with the formulas of the performance metrics: sensitivity, specificity, accuracy, precision and negative predictive value. P = Total number of positives = TP + FP, N = Total number of negatives = FN + TN.

4.1.4.2 ROC and AUC

The Receiver Operating Characteristic (ROC) curve is a graph showing the trade-off between a classifier's sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). It illustrates how well the classifier distinguishes between positive and negative classes. The curve plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. A classifier with a curve closer to the top-left corner is considered to have better performance. The Area Under the Curve (AUC) measures the separability of the classes depicted by the ROC curve. A higher AUC value (closer to 1.0) indicates better predictive performance. For example, an AUC of 0.8 means an 80% chance of correct class separation, while an AUC close to 0.5 suggests poor separation. Figure 4.8 illustrates three scenarios for the ROC curve and AUC. A perfect classifier with an AUC of 1.0 shows an ROC curve in the top-left corner, indicating complete class distinction. In contrast, when there is significant overlap between the classes, resulting in limited separability, the AUC is approximately 0.5, and the ROC curve resembles a diagonal line.

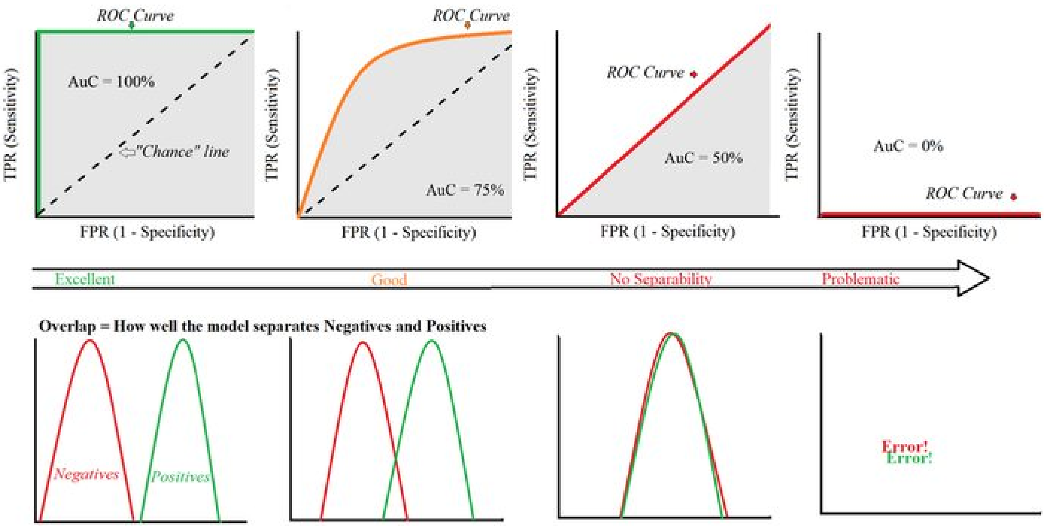


Figure 4.8: The top graph shows the ROC and corresponding AUC from 'best' model to 'worst' with on the bottom the overlap between the positive and negative classes presented. [58]

4.1.4.3 Other important metrics used in the random forest output

When evaluating the performance of the random forest model on an independent test set, next to the aforementioned evaluation metrics based on the confusion matrix, other metrics are used to asses its performance as well. Note that these are all in the context of the random forest model, since for instance 'p-values' in general have a much broader interpretation as well as confidence intervals. A brief explanation based on the interpretation of such an output as in Figure 4.9 will be provided at the end.

1. **95% confidence interval:** A 95% confidence interval is a range of values that is likely to contain the true unknown parameter of interest. It provides a measure of uncertainty or variability associated with estimating a parameter based on a sample dataset. In the context of random forest model performance accuracy, the confidence interval quantifies the plausible values within which the true accuracy is expected to lie with a 95% level of confidence. This means that if the experiment were to be repeated numerous times, the calculated confidence intervals would include the true 'accuracy' of the random forest model in approximately 95% of those experiments. Larger sample sizes generally lead to narrower confidence intervals, as they provide more precise estimates. Similarly, lower data variability and higher confidence levels result in narrower intervals, signifying increased confidence in the estimated accuracy. It is important to interpret the confidence interval properly, recognizing that it does not provide information about specific values within the range. Rather, it conveys a range of plausible accuracy values, emphasizing the uncertainty inherent in the estimation process.

2. **No Information Rate (NIR):** The No Information Rate (NIR) is a benchmark used to assess the performance of a predictive model. It quantifies the accuracy attained by a model that consistently predicts the majority class, such as disease control in this case. The NIR functions as a baseline for comparison and characterizes the accuracy achieved by a simplistic model that solely predicts the majority class, devoid of any sophisticated analysis or learning. By establishing the NIR, practitioners and researchers can evaluate the efficacy of their models relative to this trivial baseline. Comparing the accuracy of a predictive model against the NIR provides insights into its ability to surpass the simplistic approach of merely predicting the majority class. Models that exhibit accuracy higher than the NIR demonstrate their capacity to capture patterns and make informed predictions beyond what can be achieved by a rudimentary baseline model. It is worth noting that the NIR is particularly valuable in imbalanced datasets, where the majority class overwhelmingly dominates the minority class(es). In such cases, the trivial model that predicts the majority class at all times would yield a high accuracy. Consequently, a more sophisticated model that surpasses this accuracy threshold signifies its effectiveness in identifying meaningful patterns and making accurate predictions that go beyond what is achieved by the most basic strategy.

3. **P-Value [Acc > NIR]:** The p-value [Acc > NIR] serves as a statistical metric that assesses the significance of a predictive model's accuracy relative to the NIR explained before. In this context, the null hypothesis states that the accuracy of the model is equal to the NIR. By calculating the p-value, we can determine the likelihood of obtaining the observed accuracy solely due to random chance, assuming that the model's accuracy is no different from the baseline represented by the NIR. A higher p-value, e.g. 0.6741, indicates that there is insufficient evidence to reject the null hypothesis and suggests that the model's accuracy does not significantly differ from the NIR. Conversely, lower p-values, e.g. < 0.05, indicate stronger evidence against the null hypothesis, suggesting a significant difference between the model's accuracy and the NIR. In such cases, researchers can infer that the observed accuracy is unlikely to occur by chance alone, implying that the model's performance is better than the NIR.

The p-value is calculated using a permutation test called the “Out-of-Bag (OOB) Estimate of Error Improvement" also referred to as the “Permutation Test of Variable Importance". This test measures the improvement in accuracy when the predictions of the random forest are permuted for a particular variable while keeping other variables constant. The variable importance is calculated based on the decrease in accuracy resulting from the permutation. To obtain the p-value for Acc > NIR, the random forest algorithm performs permutations of the predicted class labels and computes the accuracy for each permutation. Then it compares the observed accuracy with the distribution of permuted accuracies to calculate the p-value. If the observed accuracy is significantly higher than the accuracies obtained by permutations, the p-value will be small, indicating a statistically significant difference between accuracy and NIR. The following algorithm is used.

(a) Calculate the observed accuracy difference: $Acc - NIR$.

(b) Perform a large number of permutations (e.g., 10 000) of the predicted class labels. For each permutation:

   i. Randomly shuffle the predicted class labels while keeping the actual class labels unchanged.

   ii. Calculate the accuracy difference for the permuted labels: $Perm_{Acc} - NIR$.

(c) Count the number of permutations where the accuracy difference from step 2 is greater than or equal to the observed accuracy difference from step 1.

(d) Divide the count from step 3 by the total number of permutations (e.g., 10 000) to obtain the proportion of permuted accuracies that are as extreme as or more extreme than the observed accuracy.

(e) The p-value is the proportion calculated in step (d).

4. **Cohen’s Kappa:** Cohen’s Kappa is a statistical measure designed to assess the degree of agreement between two classifiers while considering the potential agreement that could occur by chance alone. This metric is particularly valuable when evaluating the performance of a classification model, especially in scenarios where there might be an imbalance in the distribution of classes, which is the case in this study. The kappa coefficient ranges from -1 to +1 and provides insight into the level of agreement between the classifiers under examination. In the context of the random forest model output, Cohen’s kappa compares the agreement between the predicted class labels generated by the random forest algorithm and the true class labels of the data. Different values within this range have specific interpretations:

- A kappa coefficient of +1 indicates perfect agreement between the two classifiers. This implies that the observed agreement exceeds what would be expected by chance alone, indicating a high level of consistency and reliability in their classifications.
- A kappa coefficient of 0 denotes agreement equivalent to what would be expected by chance alone. In other words, there is no systematic agreement or disagreement between the classifiers beyond what random chance would produce.
- Negative values of the kappa coefficient indicate agreement that is worse than what would be expected by chance alone. This suggests that the classifiers exhibit a level of disagreement that is more substantial than random chance would account for, indicating a poor level of agreement.

In general, a higher kappa value corresponds to a better performance of the classifiers. A higher value indicates a greater level of agreement between the classifiers, surpassing what would be expected by chance alone. Consequently, a higher kappa value suggests a higher degree of consistency and reliability in the classifications made by the classifiers. To calculate Cohen’s Kappa, the formula in Equation (4.27) is used,

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.27}$$

where P_o is the observed agreement probability and P_e is the expected agreement probability by chance alone. These are based on an agreement matrix or contingency table that shows the observed frequencies of the different categories for each classifier. The observed agreement P_o is calculated by summing the diagonal elements of the contingency table and dividing it by the total number of observations. On the other hand, the expected agreement by chance P_e is calculated by considering the probabilities of each classifier selecting each category independently. Since the interpretation of Cohen’s kappa can vary depending on the field or context, generally the following guidelines, proposed by Landis and Koch [59] are used.

Kappa	**Level of Agreement**
> 0.8	Almost perfect
> 0.6	Substantial
> 0.4	Moderate
> 0.2	Fair
> 0	Slight
< 0	No agreement

Table 4.5: Cohen’s Kappa and level of agreement by Landis and Koch [59]

.

Example calculation Suppose we have two different classifiers to classify a patient as disease control or progressive disease. And the contingency table in Figure 4.9 is provided. Then in this case

$$P_o = \frac{38}{61} + \frac{9}{61} = \frac{47}{61} \approx 0.770$$

$$P_e = \frac{42}{61} \cdot \frac{48}{61} + \frac{19}{608} \cdot \frac{13}{61} \approx 0.61$$

thus κ is

$$\kappa = \frac{P_o - P_e}{1 - P_e} \approx \frac{0.770 - 0.608}{1 - 0.608} \approx 0.414$$

which is the same as the Cohen's kappa given in Figure 4.9.

5. **McNemar's Test P-Value:** McNemar's test is a statistical procedure used to compare the performance of two models or classifiers using paired data. It is commonly used when evaluating the effectiveness of a model in relation to a baseline or when comparing two different models. The test aims to determine whether there exists a significant difference in the classification performance between the two models. The test relies on a 2 × 2 contingency table that represents the counts of the four possible outcomes, as given in Table 4.6 or Figure 4.7. This table captures the number of observations where both models agree, where the first model predicts correctly while the second model predicts incorrectly, where the first model predicts incorrectly while the second model predicts correctly, and where both models predict incorrectly. To compute McNemar's test statistic, the chi-squared statistic is used. This statistic is derived from the contingency table and is calculated using the formula given in Equation (4.28), where A, B, C, and D represent the counts from the contingency table in Table 4.6 as described earlier. Once the chi-squared statistic is obtained, it is compared to the chi-squared distribution with 1 degree of freedom to determine the associated p-value. The p-value reflects the probability of observing the given data, or data more extreme, under the assumption that the two models are equally effective. If the calculated p-value is lower than a predetermined significance level (e.g., 0.05), it suggests a significant difference in performance between the two models.

   It tests the null hypothesis: $H_0 = p_B = p_C$ stating that the proportion of the frequencies in cells B and C are equal, while the alternative hypothesis is $H_1 = p_B \neq p_C$. So it only uses the cells of the contingency table where there is a disagreement between the two classifiers. Then using the Chi-squared test statistic given in Equation (4.28) the value is calculated. Then the area to the right-hand side of this value in a chi-square distribution with one degree of freedom is calculated. This area represents the p-value of the test.

	Model 2 Correctly Classified	**Model 2 Misclassified**
Model 1 Correctly Classified	A	B
Model 1 Misclassified	C	D

Table 4.6: McNemar's Test Contingency Table

$$\chi^2 = \frac{(|B - C| - 1)^2}{B + C} \tag{4.28}$$

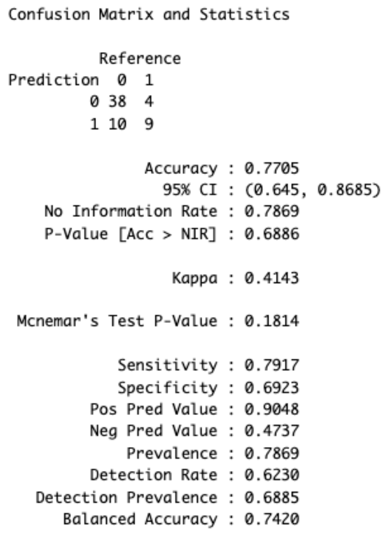
Example calculation: In the contingency table provided in Figure 4.9 to calculate the χ^2 statistic with one degree of freedom used in McNemar's test, we get:

$$\chi^2 = \frac{|(4 - 10| - 1)^2}{4 + 10} = \frac{25}{14} \approx 1.79 \tag{4.29}$$

Then the corresponding p-value is calculated to be ≈ 0.18 (using statistical software like R), which is exactly the same as the P-value given in Figure 4.9.

Here is a brief interpretation of the aforementioned metrics in relation to the example confusion matrix provided in Figure 4.9.

1. **Accuracy:** The accuracy is a measure of the overall correctness of the classification model. In this case, the accuracy is 0.7705, which means that the model correctly predicts the outcome for approximately 77.05% of the cases.
2. **95% CI:** The 95% confidence interval provides a range of plausible values for the true accuracy of the model. In this case, the confidence interval is (0.645, 0.8685), suggesting that we can be 95% confident that the 'true' accuracy falls within this interval.
3. **No Information Rate:** The no information rate is the accuracy that would be achieved by simply predicting the majority class. In this case, the no information rate is 0.7869, indicating that if we always predicted the majority class, we would achieve an accuracy of approximately 78.69%.
4. **P-Value [Acc > NIR]:** This p-value tests whether the observed accuracy is significantly different from the no information rate. In this case, the p-value is 0.6886, suggesting that there is no significant difference between the observed accuracy and the no information rate.



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  4
         1 10  9

               Accuracy : 0.7705
                 95% CI : (0.645, 0.8685)
    No Information Rate : 0.7869
    P-Value [Acc > NIR] : 0.6886

                  Kappa : 0.4143

 Mcnemar's Test P-Value : 0.1814

            Sensitivity : 0.7917
            Specificity : 0.6923
         Pos Pred Value : 0.9048
         Neg Pred Value : 0.4737
             Prevalence : 0.7869
         Detection Rate : 0.6230
   Detection Prevalence : 0.6885
      Balanced Accuracy : 0.7420
```

Figure 4.9: Example of an output of a confusion matrix after testing a fitted random forest model on an independent test set.

5. **Kappa:** Kappa is a measure of the agreement between the predicted and actual classes, taking into account the agreement that would be expected by chance. A kappa value of 1 indicates perfect agreement, while a value of 0 indicates agreement due to chance. In this case, the kappa value is 0.4143, indicating a fair agreement between the predicted and actual classes.
6. **McNemar's Test P-Value:** McNemar's test is used to assess whether there is a significant difference between the predicted classes when comparing paired observations. In this case, the p-value is 0.1814, suggesting no significant difference in the predicted classes.
7. **Sensitivity:** Sensitivity (also known as the true positive rate or recall) is the proportion of actual positive cases correctly identified by the model. In this case, the sensitivity is 0.7917, indicating that the model correctly identifies approximately 79.17% of the positive cases.
8. **Specificity:** Specificity is the proportion of actual negative cases correctly identified by the model. In this case, the specificity is 0.6923, indicating that the model correctly identifies approximately 69.23% of the negative cases.
9. **Pos Pred Value:** The positive predictive value is the proportion of predicted positive cases that are actually positive. In this case, the positive predictive value is 0.9048, suggesting that when the model predicts a positive case, there is a 90.48% chance that it is indeed positive.
10. **Neg Pred Value:** The negative predictive value is the proportion of predicted negative cases that are actually negative. In this case, the negative predictive value is 0.4737, indicating that when the model predicts a negative case, there is a 47.37% chance that it is indeed negative.
11. **Prevalence:** Prevalence is the proportion of positive cases in the dataset. In this case, the prevalence is 0.7869, indicating that approximately 78.69% of the cases are positive.
12. **Detection Rate:** The detection rate (also known as the true positive rate or recall) is the proportion of actual positive cases correctly identified by the model. In this case, the detection rate is 0.6230, which is the same as the sensitivity value.
13. **Detection Prevalence:** Detection prevalence is the proportion of predicted positive cases. In this case, the detection prevalence is 0.6885, indicating that approximately 68.85% of the cases are predicted as positive.
14. **Balanced Accuracy:** The Balanced Accuracy value is 0.7420. This indicates thatthe model generally classifies approximately 74.20% of the instances correctly, considering both the sensitivity (ability to correctly identify positive instances) and specificity (ability to correctly identify negative instances).

4.1.5 Variable Importance

The variable importance in the Random Forest models is determined using two metrics, namely: mean decrease in accuracy and mean decrease in Gini index. When dealing with imbalanced datasets, it is generally recommended to use the mean decrease in Gini index rather than accuracy. This is because accuracy can be misleading in imbalanced datasets where the majority class dominates, leading to biased variable selection. Conversely, the Gini index takes into account the relative frequencies of the classes and is less sensitive to class imbalance. For a binary classification problem:

$$Gini = 1 - (p^2 + q^2),$$

where p is the probability of selecting a sample from the positive class and q is the probability of selecting a sample from the negative class and $p + q = 1$, assuming that the data is split into two classes: positive and negative. A Gini index close to 0 indicates a 'pure' node, predominantly from a single class, while a value close to 0.5 indicates an 'impure' node corresponding to a balanced distribution between the classes. Gini is less affected by class imbalance, because it considers the squared probabilities of each class and aims to minimize impurity within each node, giving more weight to correctly classifying instances of the minority class. On the other hand, accuracy is defined as the ratio of correctly classified samples to the total number of samples,

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)},$$

with TP, TN, FP and FN derived from the confusion matrix. It treats all classes equally and does not consider the distribution of classes. In imbalanced datasets, where one class dominates the dataset, a model that always predicts the majority class can achieve high accuracy, while performing poorly on the minority class. This can mask the poor performance on the minority class, which is often more critical.

Moreover, the mean decrease in accuracy measures the reduction in the accuracy of the model when a particular predictor variable is randomly permuted. This metric reflects the importance of a variable in predicting the outcome of the model. If permuting a variable results in a significant drop in model accuracy, then that variable is considered to be more important in predicting the model outcome. In contrast, if permuting a variable has no effect on model accuracy, then the variable is deemed less important in predicting the model outcome. When computing the mean decrease in accuracy, the variable importance measure is determined by permuting the values of each predictor variable in the out-of-bag (OOB) samples of the random forest model. Permutation means that the values of a particular predictor variable are randomly shuffled, breaking the relationship between the predictor variable and the outcome variable (e.g. CA19-9 before value and the final response to chemotherapy). The permuted OOB samples are then re-evaluated using the trained random forest model to obtain the new accuracy score. The difference between the original accuracy score and the permuted accuracy score for each predictor variable is then averaged over all trees in the random forest to obtain the mean decrease in accuracy value for that variable.

Similarly, the mean decrease in Gini index measures the total decrease in impurity in the dataset when a particular predictor variable is selected for a split in a decision tree. This metric assesses the importance of a variable in splitting the data into smaller, more homogeneous groups. A higher decrease in Gini index indicates that the variable is more important in splitting the data and is thus more influential in the model's predictions. For mean decrease in Gini index, the variable importance measure is determined based on the number of times a particular predictor variable is used to split the data in the decision trees of the random forest model. Each time a variable is used to split the data, the decrease in impurity (measured by the Gini index) is recorded. This is repeated across all trees in the random forest model, and the total decrease in impurity is averaged over all trees to obtain the mean decrease in Gini index value for that variable.

4.1.5.1 Ranking the variables - Ranksum

After having found an optimal random forest model for the classification of PDAC patients, the performance of the random forest model as well as the variable importance is determined by running a simulation in which in every iteration a new optimized random forest model is created and subsequently the top 15 variables are determined based on the mean decrease in Gini impurity. This metric is chosen above the mean decrease in accuracy as explained previously. Accuracy can be misleading in imbalanced datasets where the majority class dominates the classification and therefore lead to a biased variable selection. On the other hand, the Gini index takes into account the relative frequencies of the classes and is less sensitive to class imbalance. Then in order to find the most important variables, the 'ranksum' of each variable is computed using the following:

$$rs_i = \sum_{j=1}^{n_i} \frac{1}{r_{ij}} \times f_{ij} \tag{4.30}$$

where, rs_i is the rank sum of the i^{th} variable, n_i is the total number of different ranks variable i gets, r_{ij} is the rank of i for the j^{th} rank and f_{ij} the corresponding frequency. So in words, every iteration each variable gets an importance rank. At the end of the simulation run (in this case we choose $k = 100$ runs), every variable will have 100 (potentially different) ranks. To calculate the ranksum, we first count the number of times a variable gets a certain rank. Suppose that 'CA199-before' gets ranked '1', $\frac{98}{100}$ runs and '2' $\frac{2}{100}$ runs. Then the ranksum of 'CA19-9 before' is computed as:

$$rs_{\text{CA19-9 before}} = \frac{1}{1} \times 98 + \frac{1}{2} \times 2 = 99. \tag{4.31}$$

This is then done for each of the variables and based on this ranksum, the top 10 most important variables are determined. Thus the ranksum varies between $[0, k]$, with k the number of runs.

4.2. Partial Dependence and Accumulated Local Effects

Notation: the following section uses random vectors $X = [X_1, \dots, X_p]^T$, where each random variable $X_i, i = 1, \dots, p$ has distribution $F_{X_i}(x_i)$, assuming that the vector has p predictor variables. To distinguish between a random variable and a value of a particular random variable, we use capital letters X to denote a random variable or vector and small letters x to denote the values a random variable or vector takes on. Note that in the vector case, x is also a vector of values. Depending on whether the set S consisting of the feature of interest contains 1 or more variables, X_S is a random variable in case $|S| = 1$ and a random vector is $|S| > 1$, similarly for the values x_S. f is the true underlying function of the model, $\hat{f}$ refers to the prediction function based on the data in the model, while $\tilde{f}$ refers to the approximated function.

4.2.1 Partial dependence plots (PDP)

A partial dependence plot (PDP) is a graphical representation that illustrates the marginal impact of a feature on the predicted outcome of a machine learning model. It provides insights into the relationship between the target variable and a particular feature. For example PDPs always exhibit a linear relationship for linear regression models. A PDP is able to visualize the relationship between a feature and the predicted outcome, while marginalizing the effects of the other features in the model by holding all other features constant and only varying the value of that particular feature. As the feature of interest varies across a range of values, the effect of that variation on the predicted outcome is shown.

Let $X = [X_1, X_2, \dots, X_p]^T$ represent a random vector of the p predictors or features in a model with prediction function is $\hat{f}(X)$ based on the observed data. Then partition X into $X = X_S \cup X_{S^c}$, with $S \subset \{1, \dots, p\}$ the set containing the feature of interest (or two or more features of interest) and S^c the set containing the remaining features, the complement of S. The definition of a partial dependence function $f_{S,PD}(X_S)$ with respect to the marginal distribution of the features X_{S^c} at a value $X_S = x_S$ is as follows,

$$f_{S,PD}(x_S) = \mathbb{E}_{X_{S^c}}(\hat{f}(x_S, X_{S^c})) = \int \hat{f}(x_S, X_{S^c}) d\mathbb{P}(X_{S^c}) \tag{4.32}$$

where $\hat{f}(x_S, X_{S^c})$ is the predictor function from the model. In simpler terms, given a value of $X_S = x_S$, we calculate the average predicted outcome while averaging over all the other remaining features. This involves integrating over the marginal distribution of all the other features. The term $d\mathbb{P}(X_{S^c}) = \int p(X_{S^c}) dX_{S^c}$, is the marginal probability density of X_{S^c}. Furthermore, Equation (4.32) can be estimated using the Monte Carlo method from a training dataset consisting of n data points using the following formula:

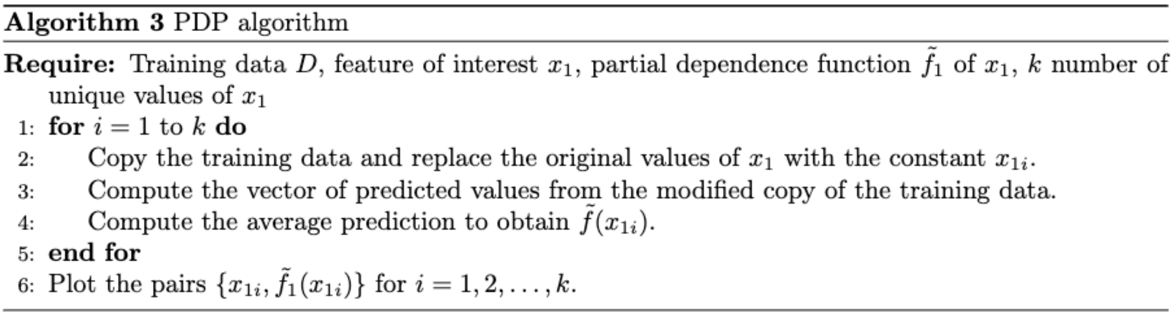
$$\tilde{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_{i,S^c}) \tag{4.33}$$

where x_{i,S^c} ($i = 1, 2, \dots, n$ the number of observations) are the values of X_{S^c} that occur in the training sample. In other words, we average out the effects of all the other predictors in the model in which we are not interested.

In classification using the random forest model, the PDP shows the probability associated with a specific class when varying the values of the feature(s) of interest in a set S. It provides an estimation of the average effect on the predicted outcome when a particular feature variable, such as X_1, is changed. The plot allows for a global interpretation of the relationship between the feature and the predicted outcome, considering all instances in the dataset. However, it assumes that the features in S^c are independent of the features in S, and extrapolation issues may arise if this assumption is violated. The ‘extrapolation’ issue refers to the problem that a PDP may incorporate data points that are highly unlikely or even impossible to occur.

4.2.1.1 PDP algorithm

In practical terms, constructing a PDP involves the following steps provided in Algorithm 3. For simplicity, consider only one feature of interest, denoted as $X_S = X_1$. Suppose that we have a dataset of n observations of which X_1 takes on k unique values, that is $\{x_{11}, x_{12}, \cdots, x_{1k}\}$, then the partial dependence of the response on feature X_1 can be constructed as follows:



```
Algorithm 3 PDP algorithm
Require: Training data $D$, feature of interest $x_1$, partial dependence function $\tilde{f}_1$ of $x_1$, $k$ number of unique values of $x_1$
for $i = 1$ to $k$ do
    Copy the training data and replace the original values of $x_1$ with the constant $x_{1i}$.
    Compute the vector of predicted values from the modified copy of the training data.
    Compute the average prediction to obtain $\tilde{f}(x_{1i})$.
end for
Plot the pairs $\{x_{1i}, \tilde{f}_1(x_{1i})\}$ for $i = 1, 2, \ldots, k$.
```

The algorithm for constructing PDP can be summarized in words as follows:

1. Select a range of values for the variable of interest, denoted as X_1. Typically, this range encompasses the minimum and maximum values observed for that variable in the dataset. Replicate the entire dataset (excluding the feature of interest) for each unique value of X_1. For instance, if the dataset contains 100 observations and X_1 has 20 unique values, the resulting sample will consist of 100×20 instances as each unique value of X_1 is assigned to 100 samples.
2. For each unique value of the variable, calculate the model's predictions while keeping all other variables constant.
3. Compute the average of the predicted outcomes for each unique value of the variable. These averaged values represent the PDP values, indicating the average effect of the variable on the model's predictions. In the given example, if there are 100 predictions per unique value of X_1, the average is taken over these 100 outputs to obtain the final output value for each unique value of X_1. So in this example we would have 20 average values for the 20 unique values of X_1.
4. Plot the values of the variable of interest against the averaged predicted outcomes obtained from the model. This plot is the Partial Dependence Plot, where the unique values of X_1 are plotted against the corresponding averaged predictions.

4.2.1.2 Advantages and Disadvantages of PDPs

PDPs are an easy way to get an understanding of the marginal effect of a particular feature on the predicted outcome. However, as mentioned earlier, one of the strong underlying assumptions in the interpretation of a PDP is the assumption of independence. This assumption may not hold when features are correlated, potentially leading to the inclusion of unrealistic data points in the average calculation (the extrapolation issue). In this section, we will discuss some of the advantages and limitations of PDPs and propose better alternatives in the next subsections.

Advantages of PDPs

- **Interpretability:** PDPs provide a clear interpretation of the relationship between a specific feature and the average prediction in the dataset. Assuming independence among features, PDPs show how changing the feature of interest affects the average prediction and thus offer a straightforward understanding of feature influence.
- **Visual representation:** PDPs offer a visual representation of the relationship between a feature and predictions. They provide a concise summary that helps identify trends and patterns more easily.
- **Identification of non-linear relationships:** PDPs can reveal non-linear relationships between variables and the outcome. They can uncover complex interactions that may not be evident in simple linear analyses.
- **Feature importance ranking:** PDPs assist in ranking the importance of variables by examining the magnitude and shape of their effects. Variables with steeper slopes or larger effects are likely to have greater influence on the model's outcome.

Limitations of PDPs

- **Independence assumption and Extrapolation issue:** PDPs assume variable independence, which can be problematic when variables actually interact and influence each other's predictions. Improbable data points may be created, leading to inaccurate representations. For instance, consider a model attempting to predict an individual's running speed based on their height and weight. In the PDP for height, the assumption is made that weight is not correlated with height, which is an incorrect assumption. When computing the PDP at a specific height (e.g., 200cm), the average is taken over the marginal distribution of weight. This can include improbable weight values (e.g., 35kg) for a person of such height, resulting in the creation of new data points in regions of the feature distribution with extremely low actual probabilities.

- **Concealment of heterogeneous effects:** PDPs only show the average marginal effects of a variable, potentially hiding heterogeneous effects where different parts of the dataset exhibit opposite associations.
- **Misinterpretation:** PDPs represent average effects and can be misleading, particularly when the model contains interactions. Due to the probable inclusion of improbable data points, they can lead to biased interpretations.

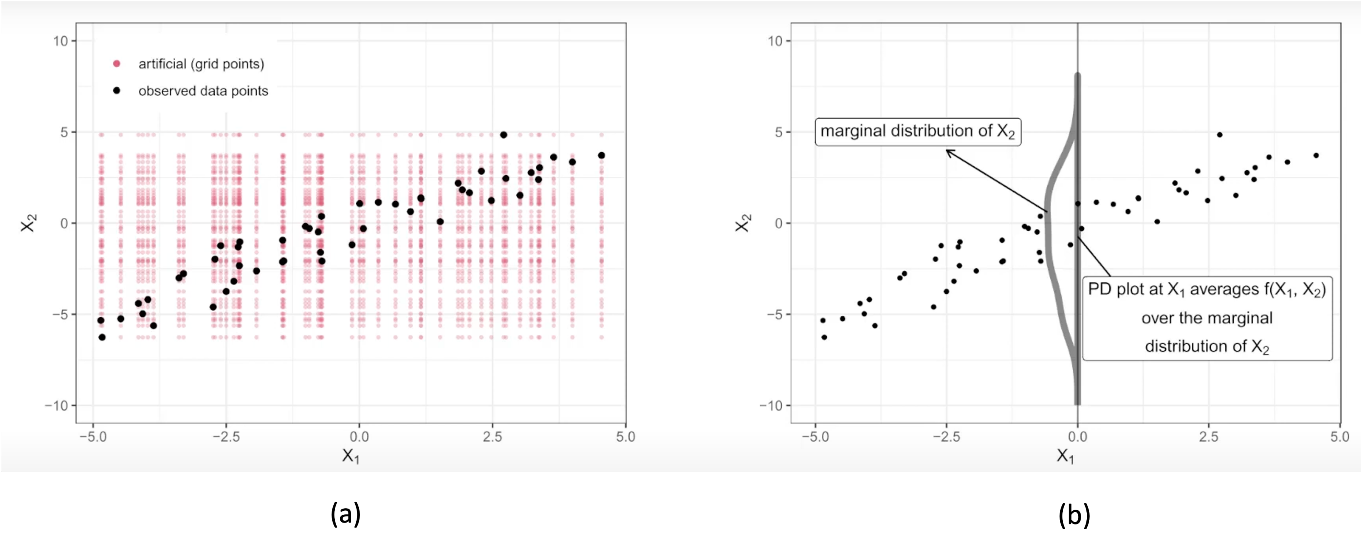


Figure 4.10: Motivation of correlated features and the PDP plot. (a) red points represent the artificial grid points and the black points the observed data points. PD plots average over predictions of artificial points that are out of distribution, which are the red points. This can lead to biased interpretation, especially when the model contains interactions too. (b) black points are the observations and the curve drawn in grey is the marginal distribution of x_2 at $x_1 = 0.0$.

To give a more visual representation of the problem of correlation in PDPs, consider the two figures in Figure 4.10. In Figure 4.10(a), the plot on the left, both the artificial grid points (depicted in red) and the corresponding observed data points (depicted in black) are displayed. PDPs average predictions over the artificial points, of which some have a low or no probability of occurrence in reality. For instance, the earlier example of an individual who is 200cm tall but weighs 35kg exemplifies this scenario. Consequently, such instances can lead to potentially misleading interpretations, particularly when the model contains interactions. To delve further into this matter, we can refer to the plot in Figure 4.10(b), which is presented on the right side. Consider the fixed value of feature $X_1 = x_1$ at $x_1 = 0.0$. The PDP averages predictions using the marginal distribution of X_2, accounting for all possible feature values of X_2. As a result, the PDP also averages predictions in regions where there is a scarcity or absence of observed data points, as indicated by the tails of the plot. Consequently, such averaging can contribute to misleading interpretations when utilizing PDPs and create unrealistic scenarios. This is the *extrapolation issue.*

4.2.2 Marginal effect plots (M-plots)

One potential solution to address the issue of the independence assumption in PDPs, as well as the usage of the entire marginal distribution of other variables, is to use Marginal Effect Plots (M-plots). By using M-plots, we calculate averages using the conditional distribution instead of the entire marginal distribution of the other variables, ensuring that only points close to the conditional distribution of a specific variable (e.g., X_2) given a certain value of another variable (e.g., $X_1 = x_1$) are considered. Rather than incorporating the entire range of values for X_2 and averaging over all these values, we define a neighborhood set $N(x_1) = \{i : x_{1i} \in [x_1 - \epsilon, x_1 + \epsilon]\}$ for some $\epsilon > 0$ consisting of observations in close proximity, allowing us to solely average points within the neighborhood of $X_1 = x_1$. This approach enables the approximation of the conditional distribution of X_2 given $X_1 = x_1$. Mathematically, in terms of expectations, PDPs average predictions over the marginal distributions. If we only consider the features X_1 and X_2 in this case at a fixed value of $X_1 = x_1$:

$$\mathbb{E}_{X_2}(\hat{f}(x_1, X_2)) \approx \tilde{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_1, X_{2,i}) \tag{4.34}$$

Contrastingly, in M-plots the expectation is represented by a conditional expectation and estimated using a neighborhood set, as given in Equation (4.35),

$$N(x_1) = \{i : x_{1i} \in [x_1 - \epsilon, x_1 + \epsilon]\} \quad \text{for some } \epsilon > 0 \tag{4.35}$$

and the following approximation,

$$\mathbb{E}(\hat{f}(x_1, X_2|X_1 = x_1)) \approx \tilde{f}_{1,M}(x_1) = \frac{1}{|N(x_1)|} \sum_{i \in N(x_1)} \hat{f}(x_1, X_{2,i}) \tag{4.36}$$

In general, M-plots average the predictions over the conditional distribution, that is:

$$f_{S,M}(x_S) = \mathbb{E}[\hat{f}(X_S, X_{S^c}|X_S = x_S)] = \int_{X_{S^c}} \hat{f}(x_S, X_{S^c}) d\mathbb{P}(X_{S^c}|X_S = x_S) \tag{4.37}$$

with $\hat{f}(x_S, X_{S^c})$ in turn estimated by,

$$\tilde{f}_S(x_S) = \frac{1}{|N(x_S)|} \sum_{i=1}^{n} \hat{f}(x_S, X_{i,S^c}|X_S = x_S) \tag{4.38}$$

with $N(x_S) = \{i : x_{1i} \in [x_1 - \epsilon, x_1 + \epsilon]\}$ the neighborhood set around $X_S = x_S$ for a given $\epsilon > 0$ and $|N(x_S)|$ denoting the number of observations in this given set.

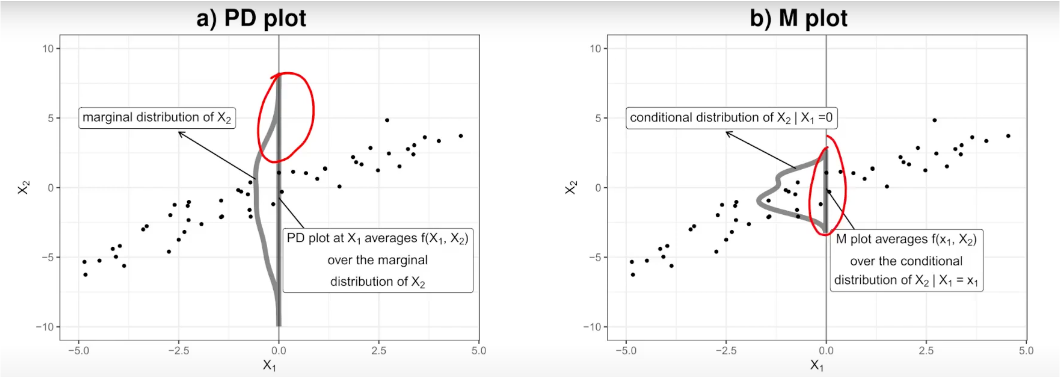


Figure 4.11: (a) PDP of X_1 against X_2, where the black dots represent the observations, the grey density at $X_1 = 0.0$ is the marginal distribution of X_2 at this point. The red circle indicates a position where the problem lies in PDPs, as it also takes into account these values which are not observed at $X_1 = 0.0$. (b) M-plot of the same observations, but with the conditional distribution of $X_2|X_1 = 0.0$ drawn in grey. M-plot averages $f(X_1, X_2)$ over the conditional distribution of $X_2|X_1 = x_1$ defined in the neighborhood $N(X_1 = x_1)$.

In summary, M-plots involve averaging predictions based on conditional distributions (e.g. $\mathbb{P}_{X_2}(\cdot|X_1 = x_1)$) as an alternative to PDPs, thereby mitigating extrapolation issues. However, M-plots are susceptible to omitted variable bias (OVB) as they incorporate effects of other dependent features, rendering them ineffective for assessing the marginal effect of a feature when feature dependencies exist. To illustrate this concern, consider the following simulation example. A linear model is fitted to 500 identically and independently distributed observations with features X_1 and X_2 following a normal distribution $N(0, 1)$, with $Cor(X_1, X_2) = 0.9$. The true underlying function is given by

$$y = -x_1 + 2 \times x_2 + \epsilon, \epsilon \sim N(0, 1). \tag{4.39}$$

Equation (4.39) represents a linear relationship between the dependent variable and the independent variables, with a negative coefficient for X_1 and a positive coefficient for X_2. This scenario is depicted in Figure 4.12(a). The goal is to examine the marginal effect of X_1 on y, while holding X_2 constant. It is evident from the correlation of 0.9 that X_1 and X_2 are strongly correlated. To illustrate this concern, an M-plot is created to visualize the marginal effect of X_1. The M-plot compares the predicted values of y based on the observed values of X_1 while keeping X_2 constant at its observed values. However, the M-plot significantly deviates from the *true marginal function of* X_1, which is represented by

$$f_1(x_1) = -x_1, \tag{4.40}$$

a linear line with a negative slope. In this context the *true marginal function of* X_1 refers to the part of the original true function y that only depends on the variable X_1, which in this case is only $f_1(x_1) = -x_1$, as this captures the marginal effect of X_1 on the function y in Equation (4.39).

Consequently, when estimating the marginal effect of X_1, one would anticipate a linear line with a slope of -1. Nevertheless, the M-plot exhibits a substantial deviation as it takes into account the effect of the second feature. As a result, the effects of both X_1 and X_2 are aggregated, resulting in a sum of $-1 + 2 = +1$, causing the green line to deviate entirely from the true marginal effect of x_1. It should be noted that in the absence of any interactions (e.g., no cross-terms $x_1 \times x_2$), PDPs do not suffer from extrapolation issues and perform well. This can be observed in Figure 4.12(b), where the blue line (PDP) closely aligns with the red line (true function).

The deviation observed in the M-plot, where the estimated marginal effect of X_1 differs from the true marginal function, can be attributed to *Omitted Variable Bias* (OVB). OVB occurs when relevant variables (X_2 in this case) that are correlated with both the independent variable of interest (X_1) and the dependent variable (y) are not included in the model. In this example, the correlation between X_1 and X_2 is strong (correlation coefficient of 0.9), indicating a high degree of association between the two variables. Thus, when fitting the linear model with only X_1 as an independent variable, the effect of X_2 is *omitted* from the model. As a result, the estimated marginal effect of X_1 in the M-plot is confounded by the omitted variable X_2, leading to biased estimates. Mathematically, when fitting the linear model with only X_1 as the independent variable, the estimated model becomes:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1, \tag{4.41}$$

where $\hat{y}$ represents the predicted values of y based on the estimated model coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. Since X_2 is omitted from the mode, the estimated coefficient $\hat{\beta}_1$ captures both the true marginal effect of X_1 and the association between X_1 and X_2. Consequently, the estimated marginal effect of X_1 in the M-plot is affected by the presence of X_2 in the underlying data. As a result, the M-plot deviates from the true marginal function of X_1, which is represented by $f_1(x_1) = -x_1$. The effect of X_2 causes the estimated marginal effect to differ from the true effect, resulting in a deviation from the expected slope of -1. This deviation in the M-plot highlights the omitted variable bias, as the effect of the omitted variable (X_2) influences the estimation of the marginal effect of X_1.

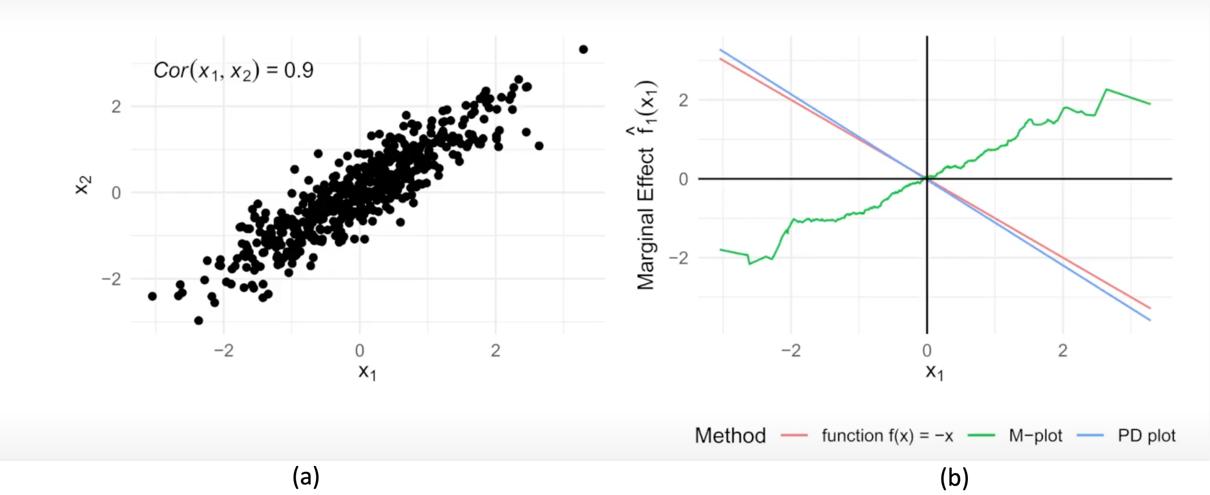


Figure 4.12: (a) Scatterplot of the generated i.i.d. observations from $y = -x_1 + 2 \times x_2 + \epsilon, \epsilon \sim N(0, 1)$, $X_1, X_2 \sim N(0, 1)$ with $Cor(X_1, X_2) = 0.9$. (b) Red = $\hat{f}_1(x_1) = -x_1$, Green = M-plot, Blue = PDP.

However, the green line of the M-plot is correct when the two features would be independent. This can be explained using the construction of the two correlated random variables X_1 and X_2 based on a linear transformation of two independent distributed variables X_1 and U, where $X_1, U \sim N(0, 1)$ using the following. Since the joint distribution of (X_1, X_2, y) alone does not specify what we want. The construction is as follows,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \sqrt{1-\alpha^2} & \alpha \end{bmatrix} \begin{bmatrix} X_1 \\ U \end{bmatrix} \tag{4.42}$$

Then

$$X^T X = \begin{bmatrix} 1 & 0 \\ \sqrt{1-\alpha^2} & \alpha \end{bmatrix} \begin{bmatrix} 1 & \sqrt{1-\alpha^2} \\ 0 & \alpha \end{bmatrix} = \begin{bmatrix} 1 & \sqrt{1-\alpha^2} \\ \sqrt{1-\alpha^2} & 1 \end{bmatrix} \tag{4.43}$$

For the determination of α, since $Cor(X_1, X_2) = 0.9$, we need to have that $\sqrt{1-\alpha^2} = 0.9$. Hence, $\alpha = \sqrt{1-(0.9)^2} \approx 0.44$. Subsequently, if we replace X_2 with $X_2 = \sqrt{1-\alpha^2}X_1 + \alpha U$, then we obtain the following:

$$\begin{aligned} f(X_1, X_2) &= -X_1 + 2X_2 + \epsilon \\ f(X_1, U) &= -X_1 + 2(\sqrt{1-\alpha^2}X_1 + \alpha U) + \epsilon \\ f(X_1, U) &= (2\sqrt{1-\alpha^2} - 1)X_1 + 2\alpha U + \epsilon \end{aligned} \tag{4.44}$$

From this we can see that the coefficient in front of X_1 is $2\sqrt{1-\alpha^2}-1$ with $\alpha \approx 0.44$. Calculating this gives $2(\sqrt{1-(0.44)^2} - 1) = 0.8$ which is exactly the slope of the green line. This shows that if the two features would be independent, the M-plot would be correct.

4.2.3 Accumulated Local Effects (ALE)

4.2.3.1 Intuition behind ALE

To gain a better understanding of the limitations of PDPs and M-plots and the intuition behind Accumulated Local Effects (ALE), consider an example involving a machine learning model used to predict the price of a house based on the number of rooms and the size of the living area. We are specifically interested in assessing the effect of the living area on the predicted price. Suppose we want to predict the price of an apartment and have a data set with ranging apartments from $25 - 1000m^2$ and rooms ranging from $1 - 20$. In the case of PDPs, the approach involves selecting (a) the feature of interest (size of living area), (b) defining a grid, and (c) for each unique grid price, replacing the feature with that price and averaging the predictions. Finally, (d) draw a curve. However, when calculating the first grid value of the PDP, $25m^2$, the living area for all the observations is replaced by $25m^2$, including houses with a large numbers of rooms. Consequently, the PDP incorporates unrealistic houses, such as a house with an area of 25 m^2 having 15 rooms (this is most likely not possible), thereby affecting the estimation of the feature's effect on the predicted price. M-plots, on the other hand, average the predictions of houses with similar living areas (around 25 m^2), thereby 'solving' the unrealistic scenario problem. But they estimate the *combined* effect of the living area and the number of rooms due to their correlation. Note that we assume that larger houses have more rooms, thus a positive correlation between these two variables. However, even if we suppose that the living area has no direct effect on the predicted price, with only the number of rooms having an impact, the M-plot would still indicate that an increase in living area leads to an increase in the predicted price. This occurs because the M-plot incorporates the effects of all correlated features and thus mixes the effect of the feature of interest with the effects of other correlated features. In this case the M-plot still shows that the size of the living area increases the predicted price, since the number of rooms increases with the size of the living area.

A solution to address these limitations is the use of Accumulated Local Effect (ALE) plots. ALE plots overcome these issues by calculating the differences in predictions instead of averaging them. For instance, when examining the effect of a living area of $25m^2$, the ALE method considers all houses with similar sizes in the neighborhood of the area of consideration. So for a living area of around $25m^2$, it obtains model predictions assuming these houses had an area of for instance $26m^2$, and subtracts the predictions assuming they had for instance an area of $24m^2$. By using these differences, ALE plots provide a pure estimation of the effect of the living area, without confounding it with the effects of correlated features. The use of differences helps to isolate the effect of the feature of interest and blocks the influence of other features. In mathematical terms, the concept of ALE plots involves the usage of partial derivatives. The idea behind ALE plots is to eliminate the undesired marginal effects from other features by first computing the partial derivative of the model's prediction function, $\hat{f}$, with respect to the feature of interest, say X_j. Subsequently, these partial derivatives are integrated with respect to the same feature. This process can be summarized as follows:

1. *Removal of other main effects:* The partial derivative of $\hat{f}$ with respect to X_j, $\frac{\partial \hat{f}}{\partial X_j}$, is computed. This step aims to eliminate the influence of other features and isolate the effect of X_j.
2. *Recovery of the main effect of X_j:* The resulting partial derivatives $\frac{\partial \hat{f}}{\partial X_j}$ are integrated with respect to the variable X_j, $\int \frac{\partial \hat{f}}{\partial X_j} dX_j$. This integration process recovers the main effect of X_j.

To acquire a mathematical intuition of ALE, we shall consider an illustrative example involving a predictive function with additive properties that depend solely on two variables, denoted as X_1 and X_2 given in Equation (4.45), similar to the example given in the simulation example concerning the problem with M-plots in the previous subsection. However, the only difference is that the noise ϵ is now replaced with the interaction term $4x_1x_2$. Let us further assume our objective is to examine only the impact of X_1 on the predicted outcome. In order to do so, we calculate the partial derivative of the predicted function $\hat{f}$ with respect to X_1, yielding the expression given in Equation (4.46). To make it clear that we are dealing with the values of the random variables X_1 and X_2, we shall use the small x_1 and x_2 in the following notation:

$$\hat{f}(x_1, x_2) = -x_1 + 2x_2 - 4x_1x_2 \tag{4.45}$$

$$\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = -1 - 4x_2 \tag{4.46}$$

From Equation (4.46) it can be seen that by taking the partial derivative with respect to x_1, the marginal effect of x_2 is removed, as the term $2x_2$, present in the original Equation (4.45) has disappeared. However, the interaction effect of x_1 and x_2 is still present. To continue, integrate the aforementioned partial derivative with respect to x_1, resulting in the following,

$$\int \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = \int -1 - 4x_2 dx_1 = -x_1 - 4x_1x_2 + C, \quad \text{with } C \text{ an arbitrary constant.} \tag{4.47}$$

Therefore, by integrating the partial derivative, we successfully eliminate the main effect of x_2, which was our objective and recovered only the terms of the true underlying function which are affected by x_1. To be more specific, the only terms that remain in this marginal effect function of x_1 depend on x_1. As a result, ALE plots avoid the OVB concerns that are present in M-plots. It is important to note that if a third feature, such as x_3, were included in the function and an interaction term, such as x_2x_3, taking the partial derivative would have eliminated both terms. However, the constant C in this case is actually a function of all the other variables that we 'removed'. In this case $C = C(x_2)$, a function of the variable x_2. If there would have been more features, say x_3 and x_4, then $C = C(x_2, x_3, x_4)$. The point is, is that this constant C is independent of x_1 and when changing a value of x_1, this constant is not affected. Hence, we have obtained the marginal effect of x_1.

4.2.3.2 Accumulated Local Effects mathematical details

Having developed an intuitive understanding of ALE plots, we shall delve into the mathematical details in this section. Recall that ALE plots are constructed by integrating partial derivatives. This unique approach circumvents two key issues: the extrapolation encountered in PDPs and the OVB problems that arise when features are dependent in M-plots. Suppose the feature of interest is X_S. Then ALE plots can be constructed based on the following steps:

1. **Estimation of the local effects:** Compute the partial derivative $\frac{\partial \hat{f}(x_S, X_{S^c})}{\partial x_S}$ evaluated at specific points $X_S = x_S$ defined within a neighborhood $N(x_S)$. These local effects quantify the impact of variations in x_S while keeping the remaining features, X_{S^c}, constant.
2. **Averaging of local effects:** Average the local effects over the conditional distribution $\mathbb{P}_{X_{S^c}}(\cdot|X_S = x_S)$, similar to the approach used in M-plots. By taking the expectation the issue of extrapolation encountered in PDPs is addressed.
3. **Integration of averaged local effects:** to estimate the global effect of X_S and mitigate the OVB issue, integrate the averaged local effects x drawn from the distribution $\mathbb{P}_{X_S}(x_S)$. This accumulation of local effects provides an estimation of the overall impact of x_S, while simultaneously removing any undesired main effects that were eliminated during the initial step (1).

To be more specific, denote X_S as the feature of interest, where $min(X_S = x_S) = z_0$, and X_{S^c} represents all other features (complement of S). The uncentered first-order ALE function, denoted as $\hat{f}_{S,ALE}(x)$, at a specific feature value $x \sim \mathbb{P}(X_S)$, can be defined as in Equation (4.48):

$$\hat{f}_{S,ALE}(x) = \int_{z_0}^{x} \mathbb{E}\Big(\frac{\partial \hat{f}(z_S, X_{S^c})}{\partial z_S}|X_S = z_S\Big) dz_S \tag{4.48}$$

Unlike PDPs, the first-order ALE function is centered by subtracting the average value of the uncentered ALE curve. Note that this average is a constant value. By doing so, the centered ALE curve, denoted as $\hat{f}_{S,ALE}$, attains a mean of zero with respect to the marginal distribution of the feature of interest, x_S. That is,

$$\hat{f}_{S,ALE}(x) = \hat{f}_{S,ALE}(x) - \int_{-\infty}^{\infty} \hat{f}_{S,ALE}(x_S) d\mathbb{P}(x_S). \tag{4.49}$$

4.2.3.3 ALE estimation

In practice, taking partial derivatives may not be feasible, particularly in tree-based models such as random forests. In random forests, the function $\hat{f}$ is piece wise constant so ALE is more a 'smoothing' method as every partial derivative will be zero locally. To address this problem, we resort to approximating the partial derivatives using finite differences based on predictions obtained from partitioning the range of values of the specific feature of interest, denoted as X_S, into $K+1$ grid points $z_{i,S}$, for $i = 0, \dots, K$, such that

$$min(X_S) = z_{0,S} < z_{1,S} < \ldots < z_{K,S} = max(X_S). \tag{4.50}$$

Thus we can write,

$$\begin{aligned} x \in (\min(X_S), \max(X_S)] &= \cup_{i=1}^{K}(z_{i-1,S}, z_{i,S}] \\ &\iff x \in (z_{0,S}, z_{1,S}] \vee x \in (z_{1,S}, z_{2,S}] \vee \ldots \vee x \in (z_{K-1,S}, z_{K,S}] \end{aligned} \tag{4.51}$$

where $z_{0,S} = \min(X_S)$, the minimum value of the feature of interest, and $z_{K,S} = \max(X_S)$ the maximum value. One approach to create K disjoint intervals for the feature of interest X_S is to use its quantiles. In this method, the interval bounds $z_{1,S}, \dots, z_{K-1,S}$ are determined by the $K-1$ quantiles, excluding the 0^{th} quantile (minimum) and the maximum quantile. More explanation on quantiles can be found in Appendix D. To continue, let us focus on the estimation of the uncentered effect first,

$$\tilde{f}_{S,ALE_{\text{unc}}}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:x_S^{(i)} \in N_S(k)} [\hat{f}(z_{k,S}, x_{S^c}^{(i)}) - \hat{f}(z_{k-1,S}, x_{S^c}^{(i)})] \tag{4.52}$$

where $k_S(x)$ denotes the interval index a feature value $x \in X_S$ falls in, $n_S(k) = |N_S(k)|$ denotes the number of observations inside the k^{th} interval of X_S and with $N_S(k) = \{i : z_{i,S} \in [x_S - \epsilon, x_S + \epsilon]\}$ the neighborhood set around $X_S = x_S$ for a given $\epsilon > 0$.

The breakdown of the formula is as follows. Starting from the right side, the ALE method calculates the differences in predictions by replacing the feature of interest with the grid values $z_{i,S}$. This difference in prediction represents the *effect* the feature has for an individual observation within a specific interval. The first sum calculates the cumulative effect by summing up the effects for all instances within the interval defined as the neighborhood $N_S(k)$. Then, we divide this sum by the number of observations in that interval, yielding the average difference in predictions for that interval. This average within the interval is referred to as the *local* effect in the ALE terminology. Finally, the left sum signifies the accumulation of average effects across all intervals up to the specific point of interest. For example, if the feature value lies within the third interval, we sum the effects of the first, second, and third intervals together. This accumulation reflects the concept of *accumulated* in ALE, as we progressively accumulate the average effects. To center the effect, we subtract a constant value, which is the average of the uncentered ALE curves. This constant represents the overall average effect of the feature across the data. The formula for the centered estimation of ALE can be found in Equation (4.53).

$$\tilde{f}_{S,ALE}(x) = \tilde{f}_{S,ALE_{\text{unc}}}(x) - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{S,ALE_{\text{unc}}}(x_S^{(i)}). \tag{4.53}$$

Overall, the ALE method combines the local effects within intervals and accumulates them to capture the cumulative effect of the feature while providing a centered perspective by subtracting the average uncentered ALE curves. The value of the ALE can be interpreted as the main effect of the feature of interest at a specific value, relative to the average prediction of the entire dataset. For instance, consider an ALE estimate of -2 at $x_S = 1500$. This indicates that when the feature of interest in the set S takes a value of 1500, the prediction is lower by 2 units compared to the average prediction. Moreover, to define the intervals, the quantiles of the distribution of the feature of interest are used as a grid. The choice of quantiles ensures that each interval contains an equal number of data points. This is important for maintaining consistency in the analysis. However, using quantiles can have a drawback in situations where the feature of interest is highly skewed, with a large number of low values and only a few high values. In such cases, the intervals may have unequal lengths, which can lead to unusual or unexpected patterns in the ALE plots.

Example:
Suppose we have two features, X_1 and X_2, and the feature of interest is X_1, an example dataset is visualized in Figure 4.13. To begin, we select a partition of the range of values of X_1 into intervals. In this example, in order to make it visually easy to interpret, we divide the range into four intervals, indicated by the vertical lines, see Figure 4.13(b). Each interval

contains a set of observations represented by black points. Within each interval, we calculate the difference in predictions by replacing the feature value of $X_1 = x_1$ with the lower and upper bounds of the interval, denoted by $z_{i,1}, i = 1, 2, 3, 4$, while keeping the other feature values unchanged. This calculation provides an approximation of the local effect using the first order finite difference method (see Appendix D). The resulting differences are denoted by the blue points in Figure 4.13(b). Subsequently, these finite differences, which approximate the local effects, are accumulated and centered to obtain the final ALE plot. The accumulation involves summing up the effects from each interval to capture the overall effect of the feature of interest across its range and centering is performed by subtracting the average of the uncentered ALE curves.

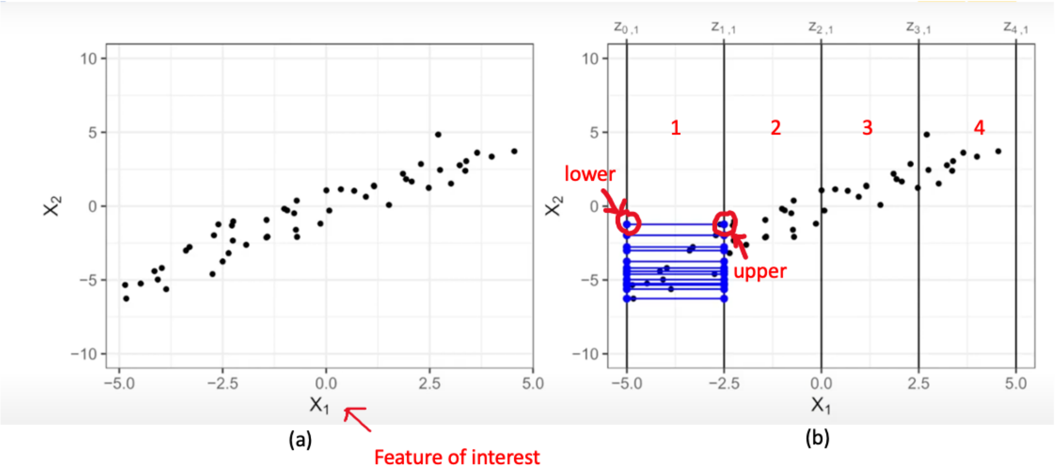


Figure 4.13: (a) Scatterplot of the data with features X_1 and X_2 as the x-axis and y-axis respectively and X_1 the feature of interest. (b) Same scatterplot as (a) but with vertical lines dividing the range of X_1 values into four intervals. Blue points are the lower and upper interval bounds used to compute the prediction difference.

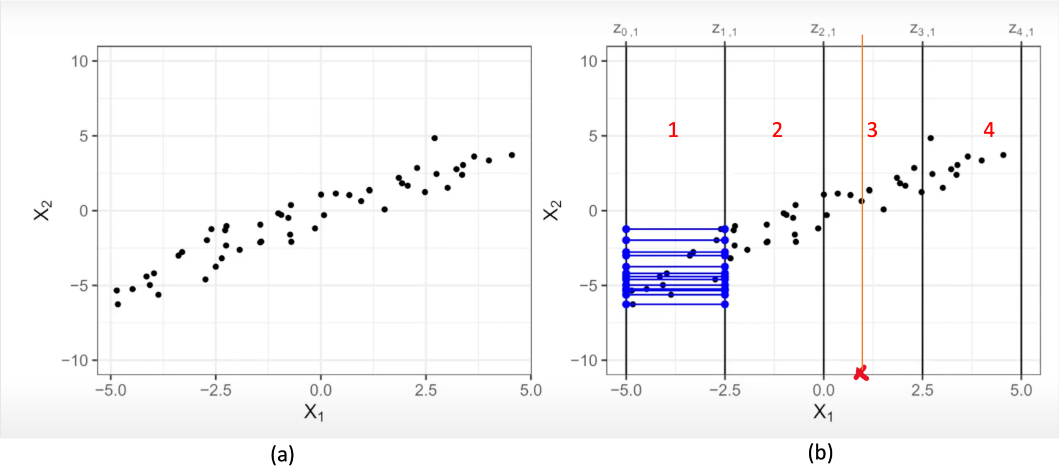


Figure 4.14: (a) Scatterplot of the data with features X_1 and X_2 as the x-axis and y-axis respectively and X_1 the feature of interest. (b) Same scatterplot but with the value of $X_1 = x_1$ of interest indicted by the red cross. To compute the ALE value of x_1, we will sum up all the local effects up to the third interval.

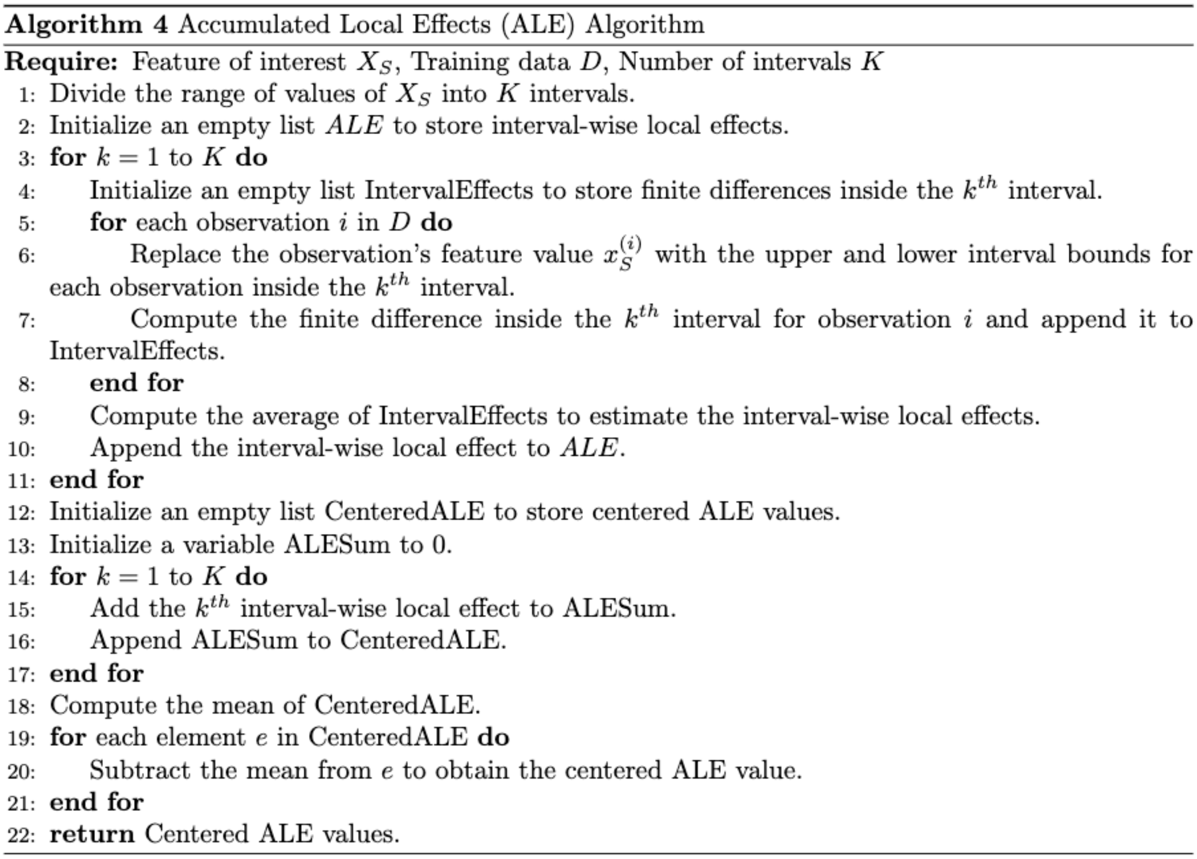
Mathematically, for the feature of interest in the set S, this process can be represented as follows:

1. For the data point $X^{(i)} = (x_S^{(i)}, X_{SC}^{(i)})$, with the value of $x_S^{(i)}$ located within the k^{th} interval of X_S, that is $x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]$, replace $x_S^{(i)}$ by the lower and upper interval bounds while keeping all the other feature values $X_S^{(i)}$ constant.
2. The finite difference corresponds to computing: $\hat{f}(z_{k,S}, X_{S^c}^{(i)}) - \hat{f}(z_{k-1,S}, X_{S^c}^{(i)})$. We do this for all the points within this interval, resulting in a set of finite difference values for this specific interval.
3. Estimate the local effect of X_S within each interval by averaging all the calculated finite differences. This is an approximation of the inner integral (expectation) that integrates over the local effects with respect to the $\mathbb{P}_{X_{S^c}}(\cdot|z_S)$ (conditional distribution).
4. To estimate the outer integral, sum up the local effects of all the intervals up to the point of interest.

By following these steps, we can obtain the ALE plot, which provides insights into the main effect of the feature of interest, taking into account the conditional distribution and accumulation of local effects. To go back to the example described earlier, if for instance the value of $X_1 = x$ as can be seen by the red cross in Figure 4.14(b), to obtain the ALE estimate, we would sum up all the local effects of intervals 1, 2 and 3 as x is located in the third interval.

4.2.3.4 ALE estimation algorithm

The procedure for computing ALE plots using the approximation explained earlier is described in words in Algorithm 4.



Algorithm 4 Accumulated Local Effects (ALE) Algorithm

```
Require: Feature of interest $X_S$, Training data $D$, Number of intervals $K$
Divide the range of values of $X_S$ into $K$ intervals.
Initialize an empty list $ALE$ to store interval-wise local effects.
for $k = 1$ to $K$ do
    Initialize an empty list IntervalEffects to store finite differences inside the $k^{th}$ interval.
    for each observation $i$ in $D$ do
        Replace the observation's feature value $x_S^{(i)}$ with the upper and lower interval bounds for each observation inside the $k^{th}$ interval.
        Compute the finite difference inside the $k^{th}$ interval for observation $i$ and append it to IntervalEffects.
    end for
    Compute the average of IntervalEffects to estimate the interval-wise local effects.
    Append the interval-wise local effect to $ALE$.
end for
Initialize an empty list CenteredALE to store centered ALE values.
Initialize a variable ALESum to 0.
for $k = 1$ to $K$ do
    Add the $k^{th}$ interval-wise local effect to ALESum.
    Append ALESum to CenteredALE.
end for
Compute the mean of CenteredALE.
for each element $e$ in CenteredALE do
    Subtract the mean from $e$ to obtain the centered ALE value.
end for
return Centered ALE values.
```

4.2.3.5 Advantages and Disadvantages of ALE

Advantages of ALE

- **No extrapolation issue and omitted variable bias:** ALE 'solves' these downsides of PDPs and M-plots, respectively, using both the conditional distribution and integration of partial derivatives, as explained previously.
- **Clear interpretation:** ALE plots provide a clear interpretation of the effects of changing a feature on the prediction, conditional on a given feature value. In addition, since ALE plots are centered at zero, the value at each point of the ALE curve is the difference to the mean prediction.
- **Fast computation:** ALE plots are computationally efficient. They can be computed quickly due to the use of grid point estimations. To be more specific, the computational complexity of ALE plots scales linearly with the number of observations. They are $O(n)$, as the largest possible number of intervals is equal to the number of observations, which corresponds to placing each observation in its own interval. As a consequence, ALE plots are faster to compute compared to other methods such as PDPs, which require n times the number of grid point estimations.

Disadvantages of ALE

- **Choice of number of intervals:** ALE plots can exhibit instability and fluctuations when a high number of intervals is used. In such cases, reducing the number of intervals improves the stability of the estimates but also obscures some of the true complexity of the prediction model. Choosing the ideal number of intervals is challenging since a small number may lead to less accurate estimates, while a high number may introduce excessive fluctuation. Finding the right balance between stability and capturing the complexity of the prediction model can be difficult.

- **No Individual Conditional Expectation (ICE) curves:** Unlike PDPs, ALE plots do not include Individual Conditional Expectation (ICE) curves. ICE curves are useful for assessing the heterogeneity of the feature effect, indicating how the effect varies across different subsets of data. In ALE plots, it is only possible to evaluate whether the effect differs between instances within each interval, but since each interval comprises different instances, it does not provide the same level of information as ICE curves.
- **Misinterpretation with highly correlated features:** while ALE plots are unbiased in handling correlated features, interpreting the effects of individual features becomes challenging when there are strong correlations. In such cases, it may be more meaningful to analyze the combined effect of changing multiple features together rather than isolating the effect of a single feature.

4.2.4 Summary PDP, M-plots and ALE plots

In summary, PDPs average the predictions over the marginal distribution at a value $X_S = x_S$,

$$\hat{f}_{S,PDP}(x) = \mathbb{E}_{X_{S^c}}[\hat{f}(x_S, X_{S^c})] = \int_{X_{S^c}} \hat{f}(x_S, X_{S^c}) d\mathbb{P}(X_{S^c}). \tag{4.54}$$

This value corresponds to the prediction function $\hat{f}$ evaluated at the feature value $X_S = x_S$ of interest in the feature set S, averaged across all the features in X_{S^c}. The averaging process involves taking the marginal expectation over the features in the set S^c, which entails integrating the predictions weighted by the probability distribution. In the case of M-plots, the predictions are averaged over the conditional distribution, which focuses on the specific values of the features in S while keeping the remaining features in S^c constant.

$$\hat{f}_{S,M}(x_S) = \mathbb{E}[\hat{f}(X_S, X_{S^c} | X_S = x_S)] = \int_{X_{S^c}} \hat{f}(x_S, X_{S^c}) d\mathbb{P}(X_{S^c} | X_S = x_S). \tag{4.55}$$

ALE plots, on the other hand, involve averaging the changes in predictions conditional on each grid value of the feature of interest. ALE plots calculate the local effects within specific intervals of the feature of interest and accumulate them over the range of feature values. The averaging is performed over the conditional distribution $\mathbb{P}(X_{S^c} | X_S = x_S)$, which takes into account the dependencies between the feature of interest and the complement features. Finally, ALE plots are centered to have a mean of zero.

$$\begin{aligned} \hat{f}_{S,ALE}(x_S) &= \int_{z_{0,S}}^{x_S} \mathbb{E}[\frac{\partial \hat{f}(z_S, X_{S^c})}{\partial X_S} | X_S = z_S] dz_S - constant \\ &= \int_{z_{0,S}}^{x_S} \Big(\int_{x_{S^c}} \frac{\partial \hat{f}(z_S, X_{S^c})}{\partial z_S} d\mathbb{P}(X_{S^c} | X_S = z_S) \Big) dz_S - constant. \end{aligned} \tag{4.56}$$

4.2.5 Interpretation of PDP and ALE plots

In section 5.2, the PDP and ALE plots for the most significant predictor variables will be presented. Interpreting these plots can provide valuable insights into the relationship between predictor variables and the response variable as learned by the model. A general framework for comprehending and deriving insights from these plots can be done as stated below. Overall, PDP and ALE plots help identify important variables and reveal how they influence the model's predictions. They provide valuable insights into the direction, strength, and shape of the relationships between predictors and the outcome variable. However, it's crucial to interpret these plots in conjunction with domain knowledge and consider potential confounding factors or interactions among variables to ensure accurate and meaningful conclusions.

PDP interpretation:

1. Each PDP represents the marginal effect of a single predictor variable on the predicted outcome, while holding other variables constant.
2. The y-axis of the PDP represents the predicted outcome, such as the probability of an event or the mean response. The x-axis represents the range of values for the specific predictor variable being examined.

3. A flat line at a certain y-value indicates no effect of that variable on the predicted outcome, while a non-linear line suggests a more complex relationship.
4. Positive slopes indicate a positive association between the predictor variable and the outcome, while negative slopes indicate a negative association.
5. The magnitude of the slope represents the strength of the relationship, with steeper slopes indicating stronger effects.
6. The intercept represents the baseline level of the predicted outcome when all predictor variables are held constant at their reference or average values. It signifies the predicted outcome value in the absence of any influence or changes in the specific variable being examined.

ALE interpretation

1. ALE plots capture the cumulative effect of a variable on the model's predictions by integrating the partial effects over the range of values.
2. The y-axis represents the change in the predicted outcome caused by variations in the specific predictor variable. The x-axis represents the range of values for the variable.
3. A flat line at zero indicates no effect of the variable on the outcome.
4. Positive values suggest a positive impact, while negative values suggest a negative impact.
5. The slope of the ALE plot represents the average change in the predicted outcome for each unit change in the predictor variable, considering its cumulative effect over its entire range. This slope captures the average effect of the variable on the outcome across its varying values.
6. The magnitude of the slope indicates the strength of the effect, with steeper slopes suggesting larger impacts.
7. The intercept in an ALE plot corresponds to the average predicted outcome when all other variables are held constant. It represents the baseline level of the outcome, typically observed when the predictor variable of interest is at its reference value or average level.

5 | Results

5.1. Model building using Random Forest

This section aims to explain the model building process using random forest and uses the techniques, evaluation metrics and graphs explained in section 4.1. It is worth mentioning that the missing values of the variables in consideration are imputed using the random forest itself. First, the complete dataset is used for model building followed by the dataset with removed outliers based on Bilirubin and Glomular Filtration Rate.

The random seed generator is used during the model building and training process. Setting a specific seed value ensures reproducibility of the results. In this case, the random seed is set to 123 for reproducibility, but it can be adjusted to any desired value or omitted.

5.1.1 Model building process

In order to build a random forest model the following steps have been taken:

1. Split the original dataset into 70% training and 30% testing data.
2. Add all the patients who have PD as final response to chemotherapy to the training data, while leaving the test set untouched. From now on, this training set will be referred to as the *(enhanced) original training dataset.* By doing this, we guarantee the independence of the test set from the training data used in subsequent predictions.
3. Based on this (enhanced) original training dataset, make oversampled, undersampled, both over- and undersampled and ROSE training datasets.
4. Train a random forest model on the different training datasets with or without the addition of misclassification weights.
5. Tune the 'mtry' parameter in the random forest model to optimize its predictive accuracy. 'mtry' determines the number of randomly selected features at each split in the decision tree building process. By exploring different values, the optimal 'mtry' can be identified with the default value being the square root of the total number of features. By selecting the 'mtry' value that yields the best performance, one can optimize the random forest and improve its predictive power. This 'tuned' random forest will then be used in the next steps.
6. Generate a confusion matrix to evaluate the model performance.
7. Create a variance importance plot with two subplots. The left subplot displays the mean decrease in accuracy when removing each variable, indicating the importance of each variable based on how much the model's performance is affected. Higher mean decrease in accuracy implies greater importance. The right subplot shows the mean decrease in Gini index, which measures the purity of the leaves at the end of each tree without each variable. Higher values indicate greater importance of the variable in the model. Note that the random forest algorithm grows fully expanded trees without pruning, where each tree's terminal nodes represent pure subsets of the data.
8. Plot the error rate of the random forest. This includes the OOB-error rate as a function of the number of trees in the forest as well as the error rate on the training dataset for both the DC as well as the PD patient groups. Based on when the OOB-error stabilizes, a maximum number of trees in the forest can be determined to reduce computation time or to add more trees to have a stable OOB-error.
9. Retune the random forest based on the optimal value of mtry as well as the selected number of trees in the forest based on the OOB-error rate.
10. Make predictions on the test dataset to evaluate the model performance.

5.1.2 Datasets used in the random forest

In order to deal with the class imbalance that is present in the provided PREOPANC2-iKnowIT dataset, we will use different training datasets to find the optimal random forest model. These will be created as follows, after splitting the entire dataset (either the full dataset or the dataset after removal of outliers based on GFR and BR) in 70% training and 30% testing as mentioned before. The test dataset is left untouched and the entire minority class (PD) is added to the training dataset using complete minority class inclusion. Then based on this 'enhanced' training dataset the following training datasets are created:

1. **Oversampled training dataset:** The oversampled training dataset will have an equal number of DC as well as PD patients. In order to do this, samples in the PD group will be replicated using oversampling until it has the same number of samples as in the DC group. Suppose we have $n_{DC} = 120$ in the original training dataset and $n_{PD} = 74$, then we *oversample* until $n_{over_{PD}} = 120 = n_{DC} = n_{over_{DC}}$, making a total of $n_{over} = 240$.
2. **Undersampled training dataset:** Similar to the oversampled training dataset, the undersampled training dataset is created. However, the opposite is now done. In this case samples in the DC group will be randomly removed until the DC group has the same number of samples as the PD group. Thus, suppose we have $n_{PD} = 74$ and $n_{DC} = 120$, then 46 samples will be randomly removed from the DC group in order to have $n_{under_{DC}} = 74 = n_{PD} = n_{under_{PD}}$.
3. **Both over- under sampling training dataset:** In this case a combination of both under- and oversampling is used using a *sampling ratio p* that determines the split between over- and undersampling. However, we want the total number of samples in this training dataset to be the same as the original training dataset. Thus, suppose we have $n_{training} = 194$ then $n_{combi} = 194$ and then depending on the sampling ratio (usually this is taken to be 0.5) the number of DC and PD patients is determined.
4. **ROSE:** Since in ROSE we synthetically generate samples, in this case any number of samples can be used.

In the subsequent sections, during the model building process, the number of observations in each dataset will be provided in a table including the split between DC and PD groups.

5.1.3 Decision trees (full dataset)

Prior to delving into the training of random forest, it is worth considering decision trees to have a visual intuition. The decision trees generated in this context are derived from the following methodologies in order to deal with the class imbalance present in the dataset using the previously explained training datasets: original, oversampled, undersampled, both over- and undersampled and ROSE training dataset. An examples of a decision trees (seed = 123) created using the original training dataset is provided in Figure 5.1. Other possible decision trees generated using the other training datasets can be found in Appendix D.2. The Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values depicted in Figure 5.2 demonstrate that the decision tree created using the undersampled training dataset yields the most accurate classification. Interestingly, the application of the ROSE method does not outperform random classification, neither does oversampling or the combined utilization of oversampling and undersampling techniques yield significantly higher AUC values or ROC curves either compared to the original training dataset. One explanation why the undersampled tree performs best might be due to the fact that undersampling the majority class can reduce the dominance of it and allow the decision tree to focus more on learning the patterns and characteristics of the minority class. By default the stopping criterion is to attempt a split if the current node has at least a minimum number of 20 observations ($minsplit = 20$) and only accept a split if the resulting nodes have at least round($\frac{minsplit}{3}$) observations and the resulting overall fit improves by $\delta Gini = 0.01$ (if Gini index is chosen to be the splitting criterion).

The decision trees can be interpreted as follows: Each box represents a node classification based on a specific variable (e.g., CRdiff < -9.5). The number within the box, ranging from 0 to 1, indicates the proportion of observations that are not classified as 0 or 1 in that particular node. For example, in Figure 5.1, the top box shows a '0' and the value underneath is 0.38, indicating that 38% of the observations in this node are classified as '1' rather than '0'. The subsequent number, 100%, represents the total percentage of observations included in that node. Since this is the root node, all 100% of the observations are present.

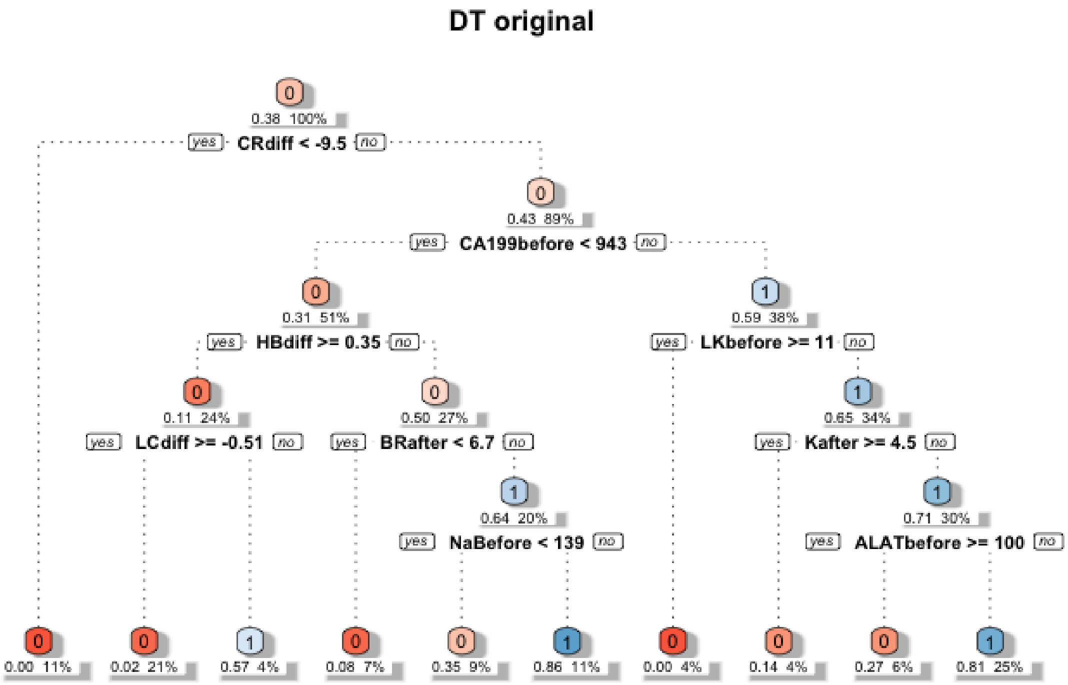


Figure 5.1: Decision tree (DT) created using the original training dataset, n=194, 0=Disease Control (n=120), 1=Progressive disease (n=74).

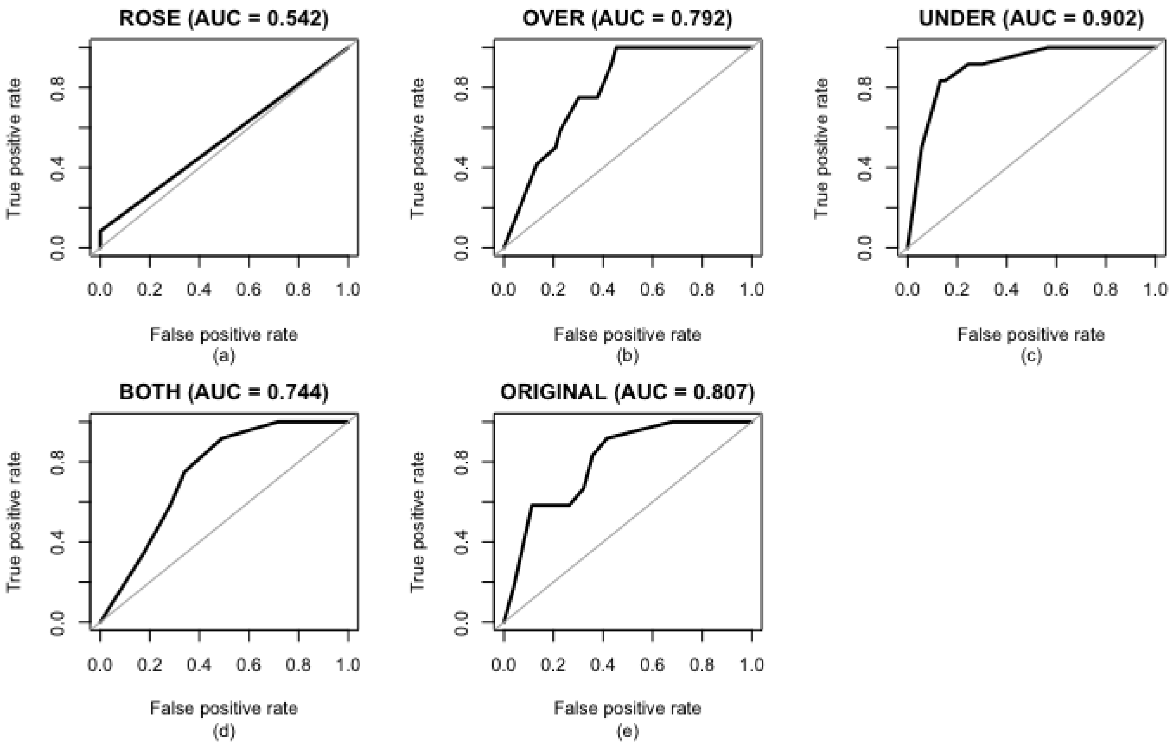


Figure 5.2: Receiver Operating Characteristic Curves (ROC) and Area Under the Curve (AUC) for the decision trees created using the (a) ROSE training dataset (n=500, Disease Control (n=265), Progressive disease (n=235)), (b) Oversampled training dataset (n=240, Disease Control (n=120), Progressive disease (n=120)), (c) Undersampled training dataset (n=148, Disease Control (n=74), Progressive disease (n=74)), (d) Both over- and undersampled training dataset (n=194, Disease Control (n=100), Progressive disease (n=94)), (e) Original training dataset (n=194, Disease Control (n=120), Progressive disease (n=74)).

5.1.4 Random Forest (full dataset)

This section aims to explain the exploration conducted on different random forests created using various datasets: original, oversampled, undersampled, both over- and undersampled and ROSE training dataset. For each random forest model, the optimal value for the “mtry" parameter is determined by tuning it based on the Out-of-Bag (OOB) error estimate. Subsequently, a plot is presented, showcasing the OOB error and training error for the two final response groups. Furthermore, a variance importance plot is generated, using the Mean Decrease in Accuracy and Mean Decrease in Gini index as metrics. The Mean Decrease in Accuracy reflects the reduction in model accuracy when excluding a particular variable from the analysis. On the other hand, the Mean Decrease in Gini index pertains to the decrease in impurity within the leaf nodes of the model when a specific variable is omitted.

Full dataset	**Disease Control (0)**	**Progressive Disease (1)**	**Total number of patients (n)**
Full dataset	173	43	216
Training dataset	120	74	194
Test dataset	53	12	65
Oversampled data	120	120	240
Undersampled data	74	74	148
Both over- and undersampled	100	94	194
ROSE	265	235	500

Table 5.1: Number of patients in each dataset used in the training of the random forest models.

Recall that random forest is an ensemble learning method where each tree is trained on approximately 63.2% of the total training data using random sampling with replacement (bootstrapping). Additionally, a random subset of variables is selected from the total variables for each tree's construction. Unlike standard decision trees that consider all variables for node splitting, random forest uses a subset $m < p$ of predictor variables. The remaining 36% of the data, known as the OOB data, is used to calculate the OOB error rate for each tree. Aggregating the OOB error rates provides an overall estimate of the model's performance on unseen data. Moreover, random forest models are inherently stochastic, meaning that they produce different outcomes with each run due to the randomness involved in the algorithm. The construction of each tree involves bootstrapping and random variable selection, resulting in unique training sets and variable subsets for each tree. As a result, the OOB-error estimate plots and variable importance plots generated for random forests can vary between runs. The plots presented in this context serve as examples to illustrate the potential output of a random forest model. It is important to note that due to the inherent randomness, different runs of the algorithm can produce distinct plots with varying OOB error rates and variable importance rankings. Examples of such plots are provided in Appendix C.2.1.

The figures provided in Figure C.21 and Figure C.22 in Appendix C.2 show the confusion matrix based on the OOB-samples as well as the evaluation metrics. It can be seen from the confusion matrices that the original random forest has the best accuracy and predicts every patient correctly. The second best model is the random forest model created using the oversampled training dataset. However, this can be expected as oversampling tends to lead to overfitted models. After this model, the random forest created using both the under- and oversampled training dataset performs best. Therefore, it has been chosen to continue evaluating the random forest model on the original training dataset and the random forest model created using the under- and oversampled dataset.

Model	**Original**	**Over**	**Under**	**Both**	**ROSE**
Number of Trees	500	500	500	500	500
Number of variables per split (optimal)	4	2	4	4	7
OOB estimate	7.73%	4.58%	18.24%	8.76%	5.2%
Accuracy	1.00	0.98	0.91	0.88	0.85
Balanced Accuracy	1.00	0.96	0.94	0.86	0.58

Table 5.2: Random forest output of the models trained using the original, oversampled, undersampled, both over- and undersampled and ROSE training datasets. The number of variables per split is optimized per model (mtry) and the OOB estimate refers to the OOB estimate of the error rate. The accuracy and balanced accuracy are based on the predictions made using the trained model and compared to the reference test set.

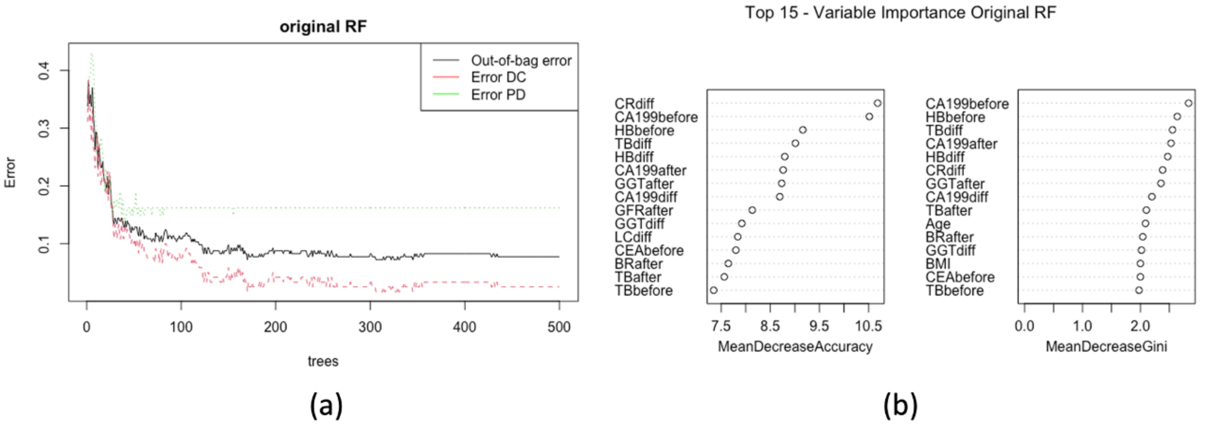


Figure 5.3: (a) Out-Of-Bag Error and Training error (b) Top 15 variable importance plot of the random forest created using the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

To optimize the random forest model trained on the over- and undersampled training dataset, a grid search is performed to fine-tune the sampling ratio parameter denoted as ‘p’. The study sets the sampling ratio to approximately 50% ($p = 0.5$) to retain a balanced representation of the minority class. Small adjustments to the sampling ratio, such as $p = 0.51$ or $p = 0.49$, are explored to assess performance improvements. The final grid search includes ratios of $p = 0.5, 0.51, 0.49, 0.48$, and 0.52. Due to model randomness, evaluation metrics varied across runs but yielded consistent outcomes. Ultimately, the original sampling ratio of $p = 0.5$ was selected for its consistently desirable results.

5.1.4.1 Misclassification Costs (full dataset)

Another approach to address imbalanced datasets is to incorporate misclassification costs or weights, where the misclassification of certain classes is considered more costly than others. This is particularly important when misclassifying instances from the minority class can have severe consequences. Optimal weights can be determined through a grid search using evaluation metrics like accuracy, F1-score, and balanced accuracy score, which are informative for imbalanced datasets. In this study, two random forest models will be trained: one on the original dataset and another on an over- and undersampled dataset. Assigning clear costs to misclassifications in a medical context is challenging, but misclassification can have significant implications. Accurate classification is crucial as misclassifying a patient with PD as DC could lead to missed treatment opportunities, while misclassifying a patient with DC as having PD could result in unnecessary chemotherapy.

Misclassification Costs for the Original Random Forest
To begin, consider the random forest trained on the original (enhanced) training dataset. A grid search across the values 1-1000 are used for the weight values where we assign a class weight of ‘1’ for misclassifying a DC patient as PD and a value between 1-1000 as well as 0-1 (by the grid search) to misclassifying a PD patient as DC. Then the optimal random forest is trained based on either the best accuracy score, f1-score and balanced accuracy. The results of the grid search on the random forest model trained using the original training dataset is provided in Table 5.3. Surprisingly, the misclassification costs for both groups remain at 1, indicating that the best performance is achieved when the weights assigned to misclassifications are equal across the classes.

Misclassification Costs for the Over- and undersampled Random Forest
A comparable grid search is conducted to determine the optimal misclassification weights for the random forest models trained on the original and over-undersampled training datasets. The results show that the misclassification weights for both groups remain at 1. This consistent pattern suggests that assigning higher weights to the minority class did not improve performance. Instead, the best results were achieved when costs for both classes were equal. This finding reinforces the importance of equal misclassification costs for DC and PD patients, more details can be found in Table 5.4.

In conclusion, the analyses of the random forest models on the original and over-undersampled training datasets consistently indicated that assigning higher misclassification weights to the minority class did not improve performance. The best results were consistently achieved when equal misclassification costs were assigned to both DC and PD patients. These findings highlight the balanced significance of accurate classification across the dataset and emphasize the importance of equal misclassification costs for both classes. However, it should be noted that the full dataset contains outliers, as discussed in Section B.2. Consequently, the construction of the “optimal model" using random forest and imbalanced dataset methods will be based on the dataset after outlier removal in subsequent sections. Nevertheless, considering the entire dataset could still provide valuable insights.

Original Random Forest Model	**Best Accuracy**	**Best f1-score**	**Best Balanced Accuracy**
OOB estimate of error rate	9.28%	10.82%	10.31%
Accuracy	0.953	1.000	0.985
Sensitivity	0.941	1.000	0.980
Specificity	1.000	1.000	1.000
Balanced Accuracy	0.971	1.000	1.000
Weights (DC, PD)	(1,1)	(1,1)	(1,1)

Table 5.3: Output of the Random forest trained on the original training dataset (ntree = 500, mtry = 7) with a grid search (1-1000) performed for the misclassification costs based on optimizing the accuracy, f1-score and balanced accuracy scores, n=194, Disease Control (n=120), Progressive disease (n=74).

Over- and Under sampled Random Forest Model	**Best Accuracy**	**Best f1-score**	**Best Balanced Accuracy**
OOB estimate of error rate	7.5%	8.5%	10.5%
Accuracy	0.939	0.908	0.923
Sensitivity	0.941	0.902	0.922
Specificity	0.929	0.929	0.929
Balanced Accuracy	0.935	0.915	0.925
Weights (DC, PD)	(1,1)	(1,1)	(1,1)

Table 5.4: Output of the Random forest trained on the over- and undersampled training dataset (ntree = 500, mtry = 7) with a grid search (1-1000) performed for the misclassification costs based on optimizing the accuracy, f1-score and balanced accuracy scores, n=200, Disease Control (n=90), Progressive disease (110).

5.1.4.2 Optimal model (full dataset)

Based on previous analyses of various random forest models, the “original" random forest model consistently outperformed other models and the inclusion of misclassification costs did not improve performance. Therefore, the optimal model is determined to be the random forest model trained on the original dataset without costs. To gain a comprehensive understanding of this model, 100 runs were conducted, each training a new random forest model optimized by adjusting the “mtry" parameter as the random forest model’s inherent randomness may result in slight variations between runs. The results presented here are based on the average outcomes derived from these 100 random forests.

Previous analyses revealed that both the OOB error and classification errors for the two groups stabilize around 300 trees. Therefore, training the random forest with 300 trees is deemed sufficient for achieving the desired accuracy. After fine-tuning the “mtry" parameter to optimize its value, the resulting optimal “mtry" value will be reported in each run. Performance will be accessed using accuracy, f1-score, and balanced accuracy calculated for each random forest model and averaged across the 100 iterations. This approach provides a more reliable estimation of overall performance compared to relying on a single score from a single random forest model. Additionally, the 15 most important variables, based on mean decrease in Gini index, will be identified in each iteration. The Gini index is chosen as the measure of importance because it considers class frequencies and node purity within decision trees, making it more sensitive to class imbalances and better suited for capturing the quality of splits compared to accuracy.

Table 5.5 presents statistics for the “mtry" variable and performance scores. From this, it can be concluded that the median optimal “mtry" value is 5, and the median accuracy, f1-score, and balanced accuracy is 1.00. All together suggesting a high level of performance. The histograms and boxplots in Figure 5.4 provide a more detailed distribution. However, the perfect classification of the test set might be caused by the CII used in the training of the model as well as the small test set size of n_{test} = 65 observations with 53 DC and 12 PD. Furthermore, the rank sum of the top 15 variables, determined based on their mean decrease in Gini index, is calculated using Equation (4.30) and visualized in Figure 5.5. From this it can be deduced that the variable “CA19-9 before" holds the highest importance across all random forest models, followed by “Hemoglobin difference" and “Thrombocyte difference." This suggests that these variables play a crucial role in determining the final response to chemotherapy.

Optimal Random Forest Full Dataset	**Mtry**	**Accuracy**	**F1-score**	**Balanced Accuracy**
Min	2	0.95	0.97	0.97
1st quartile	4	0.98	0.99	0.99
Median	5	1.00	1.00	1.00
Mean	5.3	0.99	1.00	1.00
3rd quartile	7	1.00	1.00	1.00
Max	10	1.00	1.00	1.00

Table 5.5: Statistics of the number of variables per split (mtry), and performance scores: accuracy, f1-score and balanced accuracy of the random forests (ntree = 300) across 100 runs trained on the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

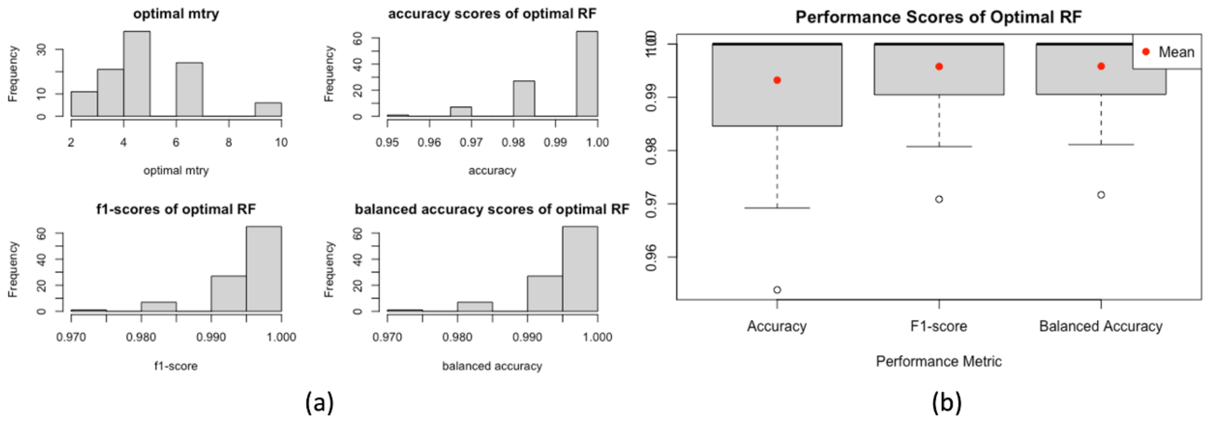


Figure 5.4: (a) Histograms of the optimal tuned number of variables per split ('mtry'), accuracy, f1-score and balanced accuracy (b) Boxplot of accuracy, f1-score and balanced accuracy of the optimal random forest (ntree = 300) across 100 runs trained on the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

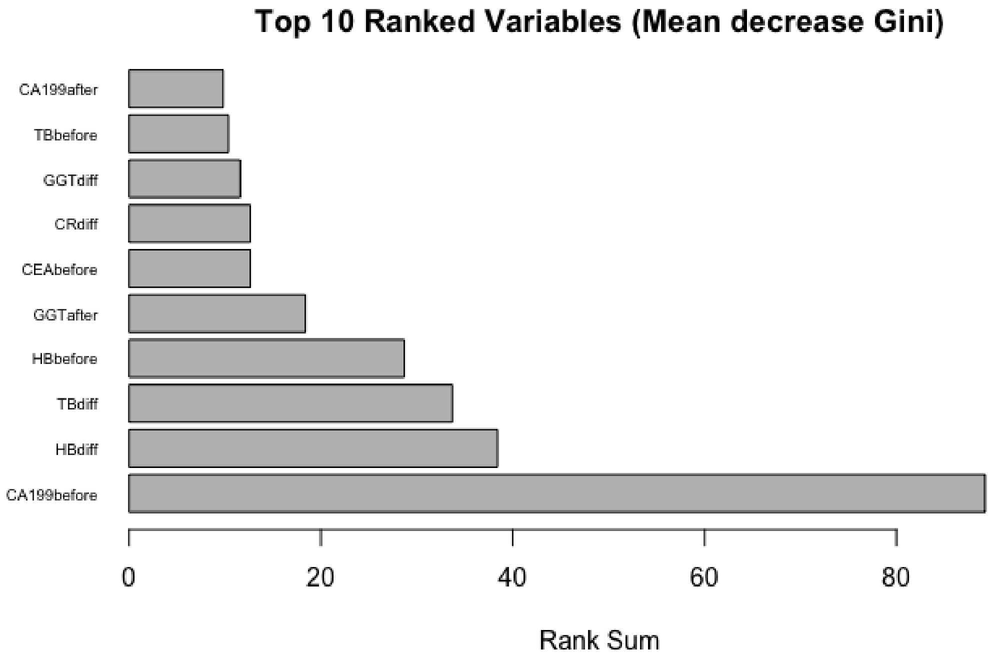


Figure 5.5: Top 10 ranked variables based on their corresponding ranksum (calculated using (4.30)) based on the mean decrease in Gini index in each of the optimal random forest (ntree = 300) tuned to their respective optimal mtry value across 100 runs trained on the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

5.1.5 Decision Trees (after removal of outliers based on GFR and BR)

Despite the comprehensive outlier analysis conducted in Section 3.2, the outcomes were not incorporated into the modeling process. This decision stems from the utilization of the IQR-method to identify outliers, which could have categorized certain patients as outliers for specific variables that might not be deemed outliers from a medical standpoint. Additionally, removing all patients flagged as outliers would result in an empty dataset, rendering the modeling infeasible. Consequently, the identification of outliers has been guided by advice of medical experts from the Erasmus MC. They suggest that patients with a Glomular Filtration Rate (GFR) value below 30 mL/min and patients with a Bilirubin value exceeding 50 $\mu mol/L$ prior to chemotherapy should be considered as outlier values and thus excluded from the dataset. These patients are deemed ineligible for chemotherapy treatment due to a failing kidney function and gallbladder obstruction. Following the removal of these individuals, the dataset consists of a total of 203 observations, with 163 classified as DC and 40 as PD. The subsequent analysis employs the same random forest methodology as previously described to evaluate this refined dataset.

After partitioning the dataset into 70% training and 30% testing data and enhancing the training dataset with the minority class, the split between DC and PD are stated in Table 5.6. Examples of possible decision trees created using the various training datasets are depicted in Figures C.6, C.7, C.8, C.9 and C.10 in Appendix C.1.2. Their corresponding ROC curves and AUC values are shown in Figure 5.6. The tumor marker CA19-9 before emerged as the most important variable in all decision trees except for the tree generated using the ROSE method, which identified CA19-9 difference as the most important variable. The former is consistent with previous literature indicating that CA19-9 is prognostic for determining chemotherapy response as well as earlier findings in the full dataset model. Interestingly, the second most important variable for the PD group is the PLR difference before and after the first chemotherapy in both the 'original', 'under' and 'over' trees. In contrast, no clear dominant parameter was observed for the DC group. Finally, the decision tree generated using the both the over- and undersampled training set performed best in terms of AUC as indicated by the ROC curves.

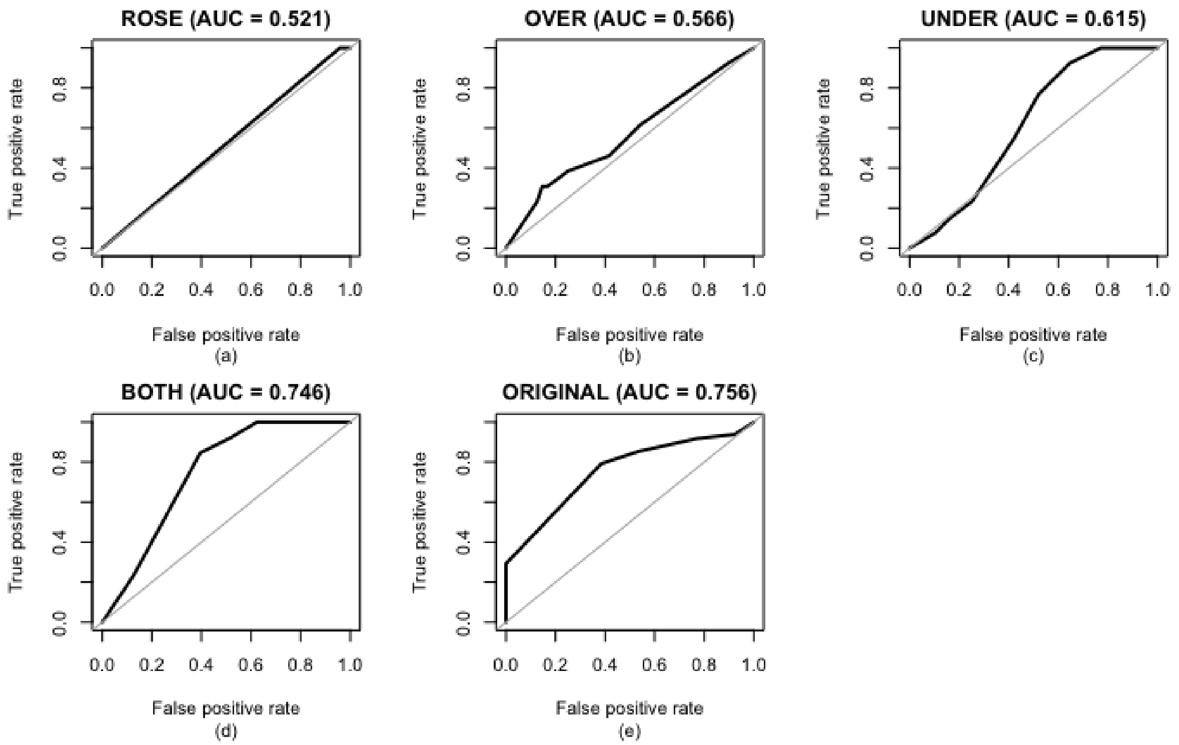


Figure 5.6: Receiver Operating Characteristic Curves (ROC) and Area Under the Curve (AUC) for the decision trees created after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values using the (a) ROSE training dataset (n=500, DC(n=265), PD(n=235)), (b) Oversampled training dataset (n=230, DC(n=115), PD(n=115)), (c) Undersampled training dataset (n=134, DC(n=67), PD (n=67)), (d) Both over- and undersampled training dataset (n=182, DC (n=95), PD (n=87)) (e) Original training dataset (n=182, DC(n=115), PD(n=67)).

After removal of outlier values based on GFR < 30 and BR > 50	**Disease Control (0)**	**Progressive Disease (1)**	**Total number of Patients (n)**
Full dataset	163	40	203
Training dataset	115	67	182
Test dataset	48	13	61
Oversampled data	115	115	230
Undersampled data	67	67	134
Both over- and undersampled	95	87	182
ROSE	265	235	500

Table 5.6: Number of patients in each dataset used in the training of the different decision tree and random forest models after removal of the outlier values based on a Glomular Filtration Rate value < 30 mL/min and a Bilirubin value of $> 50\mu mol/L$.

5.1.6 Random Forest Model (after removal of outlier values based on GFR and BR)

After removing outliers, the datasets are subjected to a similar random forest analysis as conducted on the full dataset. Each training dataset is used to train a random forest model and ‘mtry’ is tuned in the process as well. The reported OOB error estimate, as well as the (balanced) accuracy and f1-score are reported in Table 5.7. The resulting confusion matrices of a random run of the trained random forests on the test set for the original, ROSE, oversampled, undersampled, and both over- and undersampled training datasets are displayed in Appendix C.2.3. From the confusion matrices it can be seen that the two best models are the original random forest and the both over- and undersampled random forest models. The ROSE approach remains the least effective, indicating that the usage of random sampling via ROSE is not beneficial for this dataset. Consequently, only the random forest models trained on the original and both over- and undersampled training datasets will be used in further analyses.

The figures presented in Figure 5.7 depict the OOB error estimate and the error made on two final response groups to chemotherapy during the training of the random forest model. The OOB error for the PD group in the original random forest stabilizes at around 60 trees, while the errors for the DC group only stabilize after around 200-300 trees. This suggests that using 300 trees in the original random forest would be sufficient to achieve the same error. However, the errors in the random forest trained on both the over- and undersampled training data continue to fluctuate without any clear stabilization. Figure 5.7(b) indicates that this error stabilizes around 400-500 trees. Thus, using 500 trees in the forest would be sufficient in general for this model.

Model	**Original**	**Over**	**Under**	**Both**	**ROSE**
Number of Trees	500	500	500	500	500
Number of variables per split (optimal)	7	5	5	7	7
OOB estimate	10.44%	6.09%	18.66%	10.44%	1.40%
Accuracy	0.95	0.92	0.79	0.74	0.66
Balanced Accuracy	0.97	0.89	0.86	0.78	0.50

Table 5.7: Random forest output of the models trained using the original, oversampled, undersampled, both over- and undersampled training datasets after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values. The number of variables per split is optimized per model (mtry) and the OOB estimate refers to the OOB estimate of the error rate. The accuracy and balanced accuracy are based on the predictions made using the trained model and compared to the reference test set.

Furthermore, the variable importance plots based on the mean decrease in accuracy Gini index of both models are presented in Figure 5.8. The plots show that the most important variable remains CA19-9 before the first chemotherapy cycle, consistent with previous findings. In the random forest trained on the original training dataset, AF and CA19-9 after the first cycle are the second most important variables based on the mean decrease in accuracy and Gini, respectively. The subsequent other important variables differ per metric as well. In contrast, the random forest trained on both the over- and undersampled training dataset shows that the second most important variable is GGT before the first chemotherapy cycle, followed by the GGT after the first cycle. After that, depending on the metric used, the variables differs. Nevertheless, it is noteworthy that as explained earlier, accuracy does not take the minority class into account in the way that the Gini index is able to. Therefore, it is wise, when dealing with an imbalance in classes to value the variable importance based on the mean decrease in Gini index over the mean decrease in accuracy. These plots show clearly that depending on the metric used, the results differ.

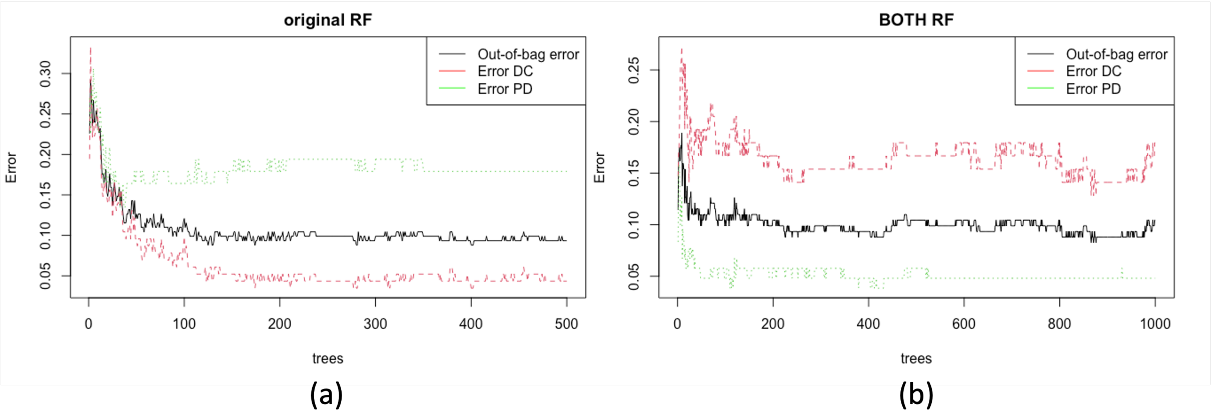


Figure 5.7: Out-Of-Bag Error and Training error of the random forest created using the (a) original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=115), Progressive disease (n=67) (b) both the over- and undersampled training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values trained using 1000 trees, n=182, Disease Control (n=95), Progressive disease (n=87).

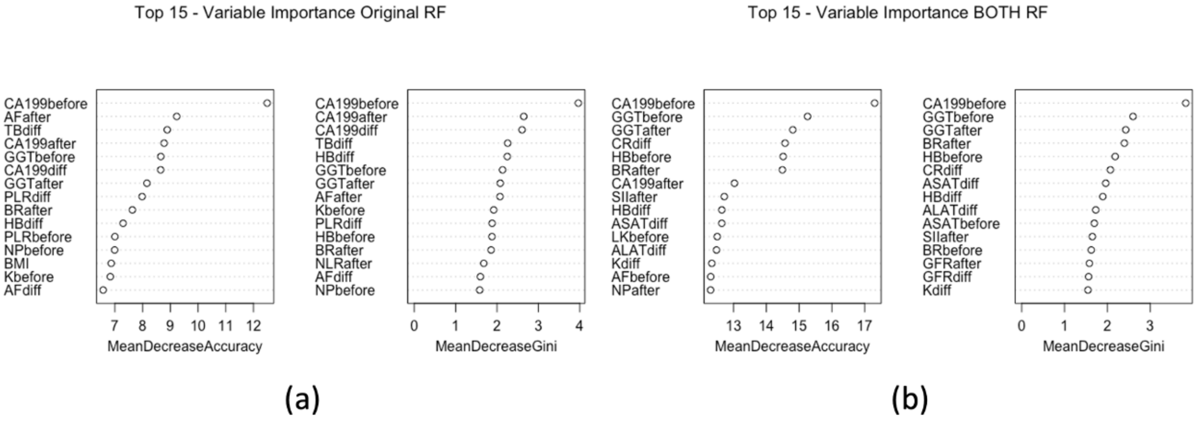


Figure 5.8: Top 15 variable importance plot of the random forest created using (a) the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=115), Progressive disease (n=67), (b) the both the over- and undersampled training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=95), Progressive disease (n=87)

5.1.6.1 Misclassification costs after removal of outliers

Similar to the random forest build on the original full dataset, misclassification costs are introduced in both the original well as the model trained on the over- and undersampled training dataset. Based on the metrics accuracy, f1-score and balanced accuracy a grid search is performed to find the optimal model with misclassification costs. Surprisingly, regardless which metric is used, the optimal misclassification weights remain one for both final response groups in the random forest trained on the original training dataset without outliers. On the contrary, the random forest trained on both the under- and oversampled training dataset does weigh misclassifying a PD patient as DC twice as costly. However, the results are pretty similar with only some small differences due to the inherent randomness in the random forest model itself. Nevertheless, despite having a lower OOB error estimate compared to the original, the both over- and undersampled random forest model does not perform better in terms of accuracy and other evaluation metrics. Therefore, the original random forest model is chosen to be the ‘optimal model’ after removal of outliers based on GFR and BR, which is consistent with the previous ‘optimal model’ found on the full dataset.

Original Random Forest without outliers	**Best Accuracy**	**Best f1-score**	**Best Balanced Accuracy**
OOB estimate of error rate	10.44%	10.99%	11.54%
Accuracy	0.95	1.00	0.98
Sensitivity	0.94	1.00	0.98
Specificity	1.00	1.00	1.00
Balanced Accuracy	0.97	1.00	0.99
Weights (DC, PD)	(1,1)	(1,1)	(1,1)

Table 5.8: Output of the Random forest trained on the original training dataset (ntree = 500, mtry = 7) with a grid search (1-1000) performed for the misclassification costs based on optimizing the accuracy, f1-score and balanced accuracy scores, n=182, Disease Control (n=115), Progressive disease (n=67).

Both over- and undersampled Random Forest without outliers	**Best Accuracy**	**Best f1-score**	**Best Balanced Accuracy**
OOB estimate of error rate	7.69%	9.34%	8.24%
Accuracy	0.79	0.79	0.75
Sensitivity	0.77	0.77	0.73
Specificity	0.85	0.85	0.85
Balanced Accuracy	0.81	0.81	0.78
Weights (DC, PD)	(1,2)	(1,2)	(1,2)

Table 5.9: Output of the Random forest trained on the both over- and undersampled training dataset (ntree = 500, mtry = 7) with a grid search (1-1000) performed for the misclassification costs based on optimizing the accuracy, f1-score and balanced accuracy scores, n=182, Disease Control (n=95), Progressive disease (n=87).

5.1.6.2 Optimal Model after removal of outliers based on GFR and BR

After exploring various random forest models with different training datasets and considering misclassification costs, the optimal model remains the original random forest model trained on the dataset without any misclassification costs. This model consistently outperformed other models in previous analyses and is therefore chosen as the 'optimal model' for this dataset. The histograms and boxplots in Figure 5.9 display the distribution of the optimal 'mtry' parameter and the performance metrics (accuracy, f1-score, and balanced accuracy) across 100 optimized random forest model. The results show that the optimal 'mtry' value is predominantly 5, indicating the use of five variables per split in most models. F1-score and (balanced) accuracy values are also consistently high across the 100 models, averaging at 0.98 for both metrics f1-score and balanced accuracy and 0.94 for accuracy. This indicates that the model has good classification performance overall. Note that the values of these metrics can range between 0 and 1, where 1 represents perfect classification performance, and 0 represents no predictive power. Note that the model's classification performance is consistent across all the random forests, which suggests that the model is robust and not overfitting to the data. More detailed statistical analysis of the performance scores for can be found in Table 5.10. Moreover, from the 15 most important variables based on the mean decrease in Gini determined during the 100 runs of the random forests, the most important variable that dominates is remains CA19-9 value before the first chemotherapy cycle. The second most important variable was either with CA19-9 after the first cycle or CA19-9 difference. In general, it is clear that CA19-9 is very important in determining the final response to chemotherapy as it dominates the top 3 in all models.

The ranksum of the top 15 variables selected based on the mean decrease in Gini of the 100 optimal random forests is calculated and used to generate visual representations of the top 10 variables with the highest ranksum, displayed Figure 5.10. CA19-9 before consistently holds the top rank followed by CA19-9 after and CA19-9 difference. Other important variables include: HB difference, TB difference, GGT before and after the first chemotherapy cycle. As previously explained, Gini index is favored over accuracy as it accounts for class frequencies and prioritizes node purity, thereby providing a more accurate reflection of the significance of correctly classifying the minority class. Table C.1 in Appendix C provide a more detailed ranking of the variables as well as their respective counts.

Optimal random forest No outliers based on GFR and BR	**Mtry**	**Accuracy**	**F1-score**	**Balanced Accuracy**
Min	2	0.92	0.95	0.95
1st quartile	4	0.96	0.98	0.98
Median	5	0.95	0.98	0.98
Mean	5.6	0.94	0.98	0.98
3rd quartile	7	0.95	0.99	0.99
Max	10	0.98	1.00	1.00

Table 5.10: Statistics of the number of variables per split (mtry), accuracy, f1-score and balanced accuracy of the random forests (ntree = 300) across 100 runs trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=115), Progressive disease (n=67).

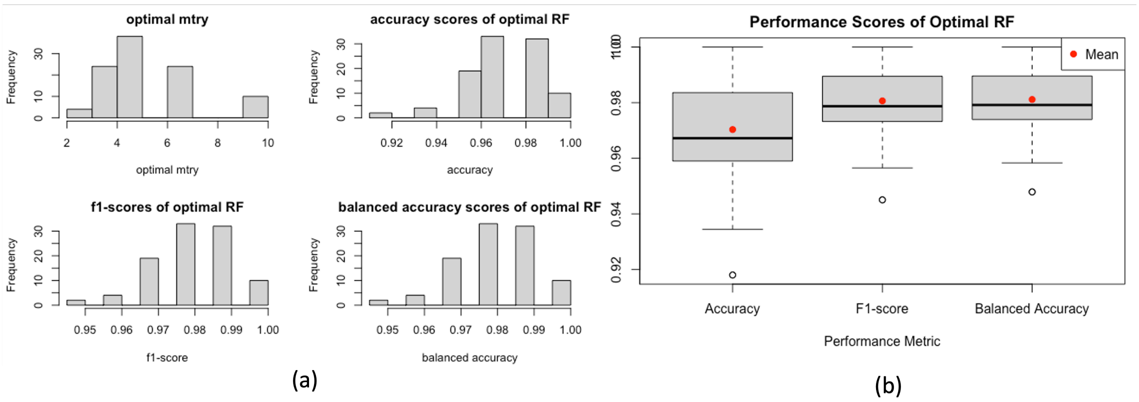


Figure 5.9: (a) Histograms of the optimal tuned number of variables per split ('mtry'), accuracy, f1-score and balanced accuracy (b) Boxplot of accuracy, f1-score and balanced accuracy of the optimal random forest (ntree = 300) across 100 runs trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=115), Progressive disease (n=67).

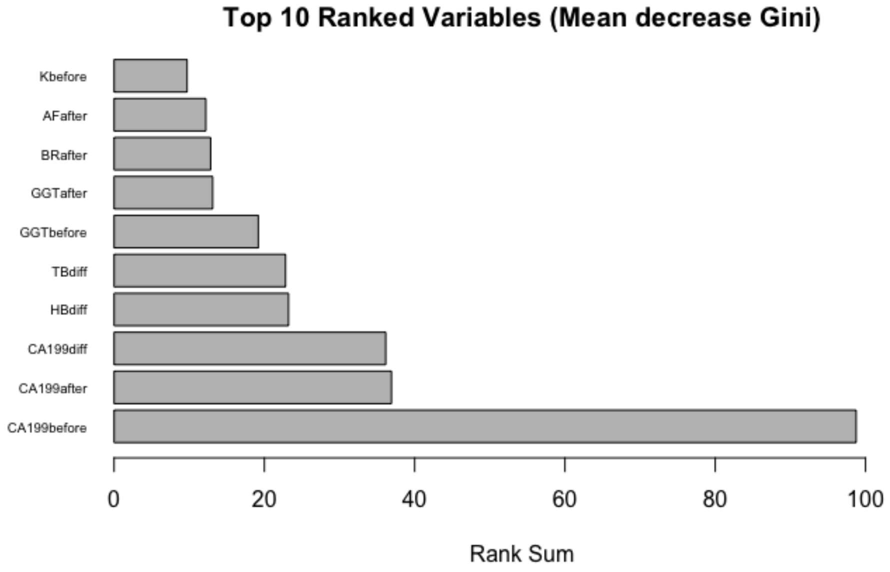


Figure 5.10: Top 10 ranked variables based on their corresponding ranksum based on the mean decrease in Gini index in each of the optimal random forest (ntree = 300) tuned to their respective optimal mtry value across 100 runs trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=115), Progressive disease (n=67).

5.1.7 Random Forest (optimal model with only top 10 variables)

To gain further insights, it is worthwhile to evaluate the performance of the original random forest model on the top 10 variables selected based on the ranksum derived from the 100 runs in the preceding sections. These variables include CA19-9 before, CA19-9 after, CA19-9 difference, HB difference, TB difference, GGT before, GGT after, BR after, AF after, and K before, where "before" and "after" refer to the measurements taken before and after the first chemotherapy cycle. Similar to the previous approach, the training dataset is augmented by including the full minority class to ensure comprehensive model training. The distribution of patients across the datasets is the same as before, given in Table 5.6. The OOB-error plot in Figure 5.11(a) illustrates that the OOB error, as well as the classification errors for both response groups, converge to around 0.12 after approximately 150 trees. Furthermore, the top 10 variables are re-evaluated based on these 10 variables alone. Interestingly, the order of importance differs from the previous analysis, as depicted in Figure 5.11(b). However, CA19-9 before remains the most significant variable, followed by either the HB or TB difference, depending on whether the mean decrease in accuracy or the Gini index is utilized as the criterion.

A similar approach to the previous analysis is employed to obtain a robust assessment of the model's performance using only the top 10 variables. The resulting histograms and box plots can be found in Figure 5.12. In this case the optimal "mtry" on average, the f1-score and balanced accuracy are approximately 0.89 and 0.90, respectively, compared to 0.98 when considering all variables in the dataset. This implies that by using only the top 10 variables, the model is still able to successfully capture a significant portion of the dataset's information, suggesting their crucial role in determining the final response to chemotherapy. Moreover, the rankings of the top 10 variables based on their mean decrease in the Gini index is visualized in Figure 5.13. This shows that CA19-9 before and CA19-9 after the first chemotherapy cycle remain the most important variables. However, the relative importance of the TB difference and GGT before the first cycle has increased, and the CA19-9 difference now ranks last. Overall, these findings highlight the significance of the top 10 variables in predicting the response to chemotherapy in PDAC patients.

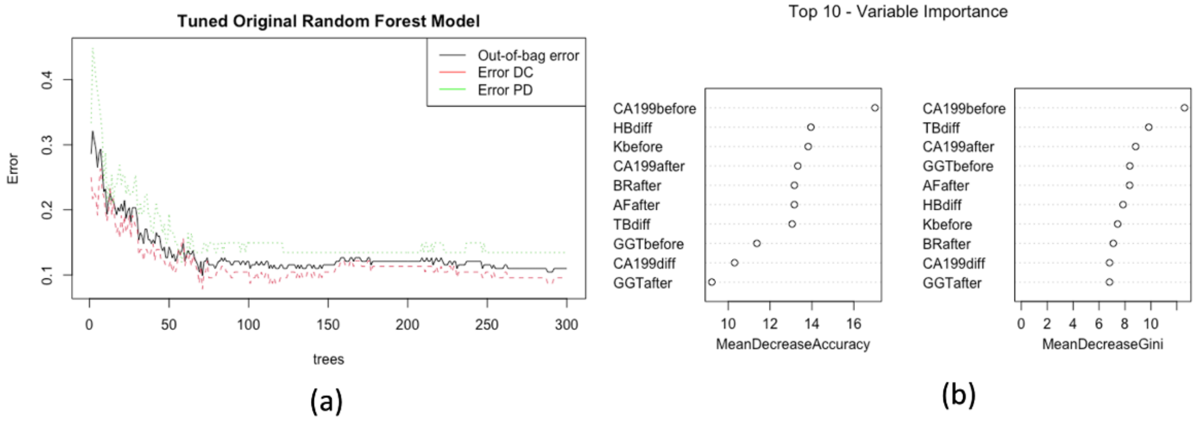


Figure 5.11: (a) Out-Of-Bag Error and Training error (b) Top 10 variable importance plot of the random forest created using the original training dataset defined by the top 10 most important variables after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

Optimal Random Forest trained only top 10 after removal of outliers based on GFR and BR	**Mtry**	**Accuracy**	**F1-score**	**Balanced Accuracy**
Min	2	0.79	0.84	0.86
1st quartile	2	0.84	0.88	0.90
Median	2	0.84	0.88	0.90
Mean	2.3	0.84	0.89	0.90
3rd quartile	2	0.85	0.90	0.91
Max	4	0.89	0.92	0.93

Table 5.11: Statistics of the number of variables per split (mtry), and performance scores: accuracy, f1-score and balanced accuracy of the optimal random forest trained on the original training dataset defined by the top 10 most important variables after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

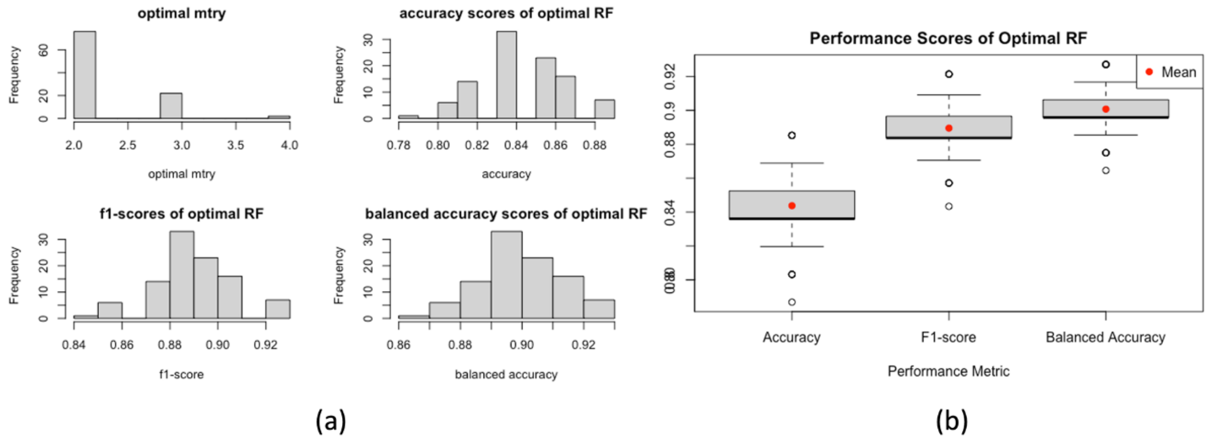


Figure 5.12: (a) Histograms of the optimal tuned number of variables per split (‘mtry’), accuracy, f1-score and balanced accuracy (b) Boxplot of accuracy, f1-score and balanced accuracy of the optimal random forest (ntree = 300) across 100 runs trained on the original training dataset defined by the top 10 most important variables after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

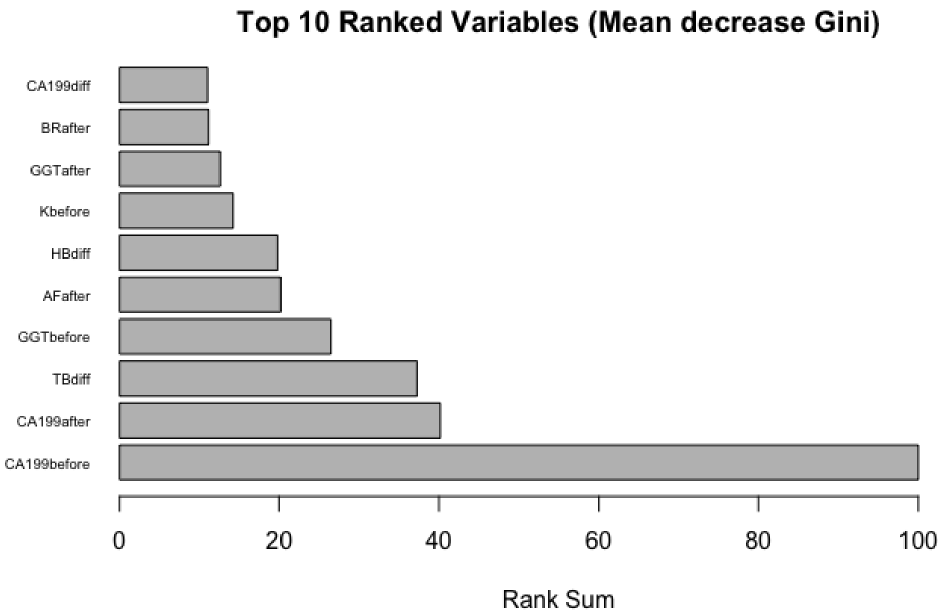


Figure 5.13: Top 10 ranked variables based on their corresponding ranksum (calculated using Equation (4.30)) based on the mean decrease in Gini index in each of the optimal random forest (ntree =300) tuned to their respective optimal mtry value across 100 runs trained on the original training dataset defined by the top 10 most important variables after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

5.1.8 Random forest (only pre-chemotherapy values)

In addition to optimizing the model using both pre- and post-chemotherapy values, it is important to establish a model using only pre-treatment values. These values, measured before any intervention, can assist in deciding whether to initiate chemotherapy, considering its toxicity in PDAC. Similar to the previous approach, various techniques including oversampling, undersampling, a combination of both, and the ROSE technique were tested, but the original random forest model remained the ‘best’ model among them. Introducing misclassification costs did not improve the model’s performance either. Therefore, the optimal model remains the original random forest.

The pre-treatment dataset consists of 23 variables including the dichotomized final response. The distribution of the final response groups in the complete dataset, enhanced training dataset, and test dataset can be found in Table 5.6. After optimizing the number of variables per split (“mtry"), we determined the number of trees to use to be 300, based on the stabilization of the OOB error and classification errors of the training set. The variable importance plot in Figure 5.14(b) confirms the top three significant variables found in the previous models, namely: CA19-9 before, GGT before, and HB before. Next, in the simulation consisting of 100 newly optimized random forest models, each composed of 300 trees, the variable importance is determined using the mean decrease in Gini index, following the same methodology as before. The results of the 100 random forests, constructed with the optimized number of trees and variables per split, are summarized in Table 5.12. The most frequently observed optimal number of variables per split is 3, as indicated by the histogram in Figure 5.15(a). The model performs exceptionally well, with an average accuracy of 0.95, an average f1-score of 0.97, and an average balanced accuracy of 0.97 on the test set. This demonstrates the predictive power of utilizing only the pre-chemotherapy values in determining the final response. The variable importance plot in Figure 5.16 reveals that CA19-9 before is the most informative variable, followed by the GGT value. Surprisingly, neutrophils and other factors such as age and BMI also exhibit notable importance in predicting the final response to chemotherapy.

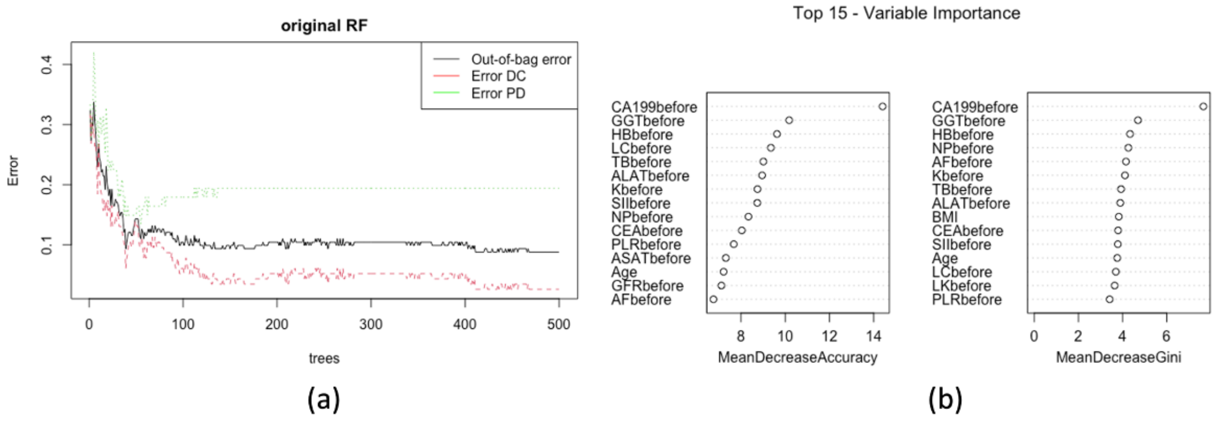


Figure 5.14: (a) Out-Of-Bag Error and Training error (b) Top 15 variable importance plot of the random forest created using the original training dataset with only the before chemotherapy values after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

Optimal Random Forest trained only on before chemotherapy variables after removal of outliers based on GFR and BR	**Mtry**	**Accuracy**	**F1-score**	**Balanced Accuracy**
Min	2	0.90	0.93	0.94
1st quartile	2	0.93	0.96	0.96
Median	3	0.95	0.97	0.97
Mean	2.9	0.95	0.97	0.97
3rd quartile	4	0.97	0.98	0.98
Max	6	1.00	1.00	1.00

Table 5.12: Statistics of the number of variables per split (mtry), and performance scores: accuracy, f1-score and balanced accuracy of the optimal random forests (ntree = 300) across 100 runs trained on the dataset with only the before chemotherapy values after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values.

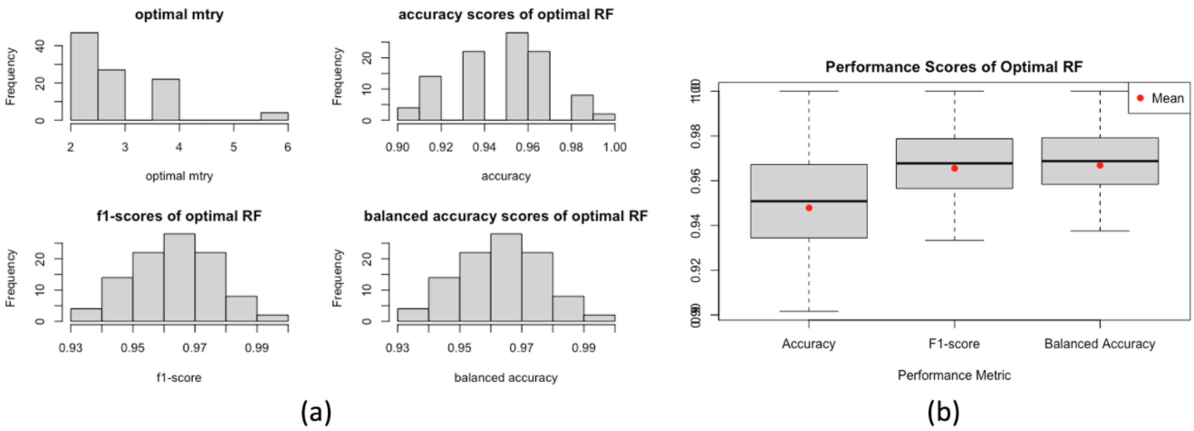


Figure 5.15: (a) Histograms of the optimal tuned number of variables per split (‘mtry’), accuracy, f1-score, balanced accuracy)(b) Boxplot of the accuracy, f1-score and balanced accuracy of the optimal random forest (ntree = 300) across 100 runs trained on the original training dataset with only pre-chemotherapy values after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

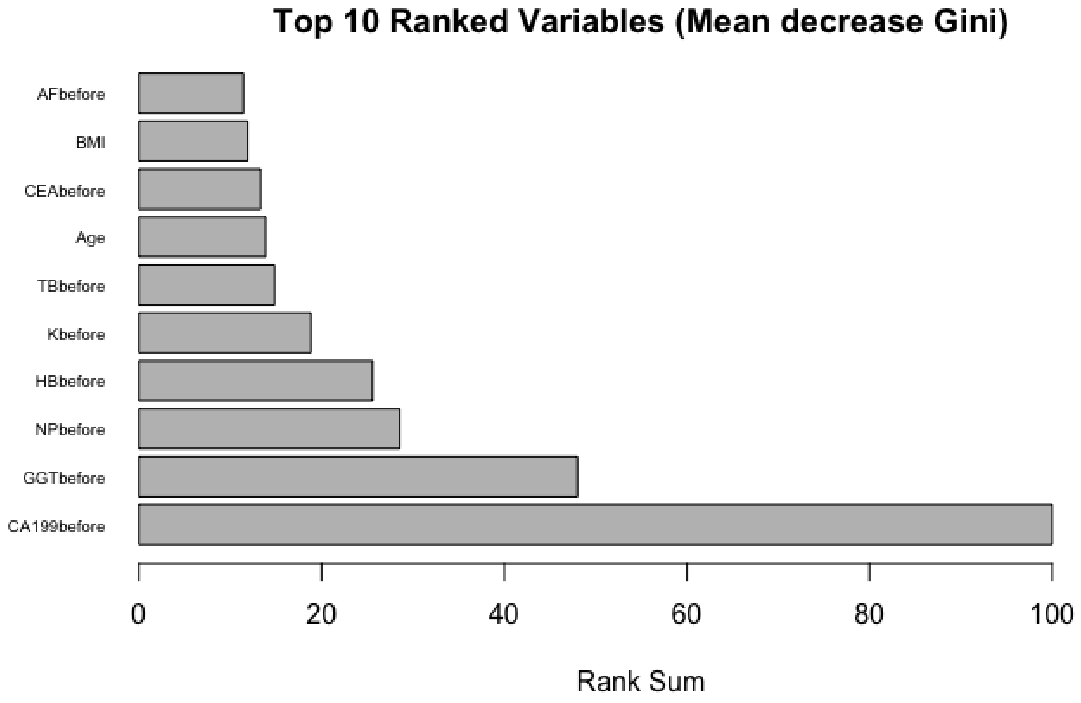


Figure 5.16: Top 10 ranked variables based on their corresponding ranksum (calculated using (4.30)) based on the mean decrease in Gini index in each of the optimal random forest (ntree =300) tuned to their respective optimal mtry value across 100 runs trained on the original training dataset with only the pre-chemotherapy values after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values n=182, Disease Control (n=115), Progressive disease (n=67).

5.1.9 Overview of the top 10 most important variables

After optimizing four distinct random forest models using four different datasets, namely: (1) full dataset, (2) dataset with outliers removed based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, (3) dataset based on the top 10 variables identified by (2), (4) dataset after removal of outliers based on GFR and BR values containing only pre-chemotherapy values, the top 10 most important variables were identified. This is done using the mean decrease in Gini index in each random forest, and then their corresponding ranksum across the 100 runs was calculated. An overview of the rankings for each model is presented in Table 5.13. The results highlight that “CA19-9 before" consistently emerges as the most important variable across all models. Besides “CA19-9 before," variables such as HB, TB and GGT consistently emerge as important factors in determining the response to chemotherapy across all models. These variables provide valuable insights into a patient’s overall health and can impact treatment outcomes. Lower hemoglobin levels may indicate anemia or compromised oxygen-carrying capacity, affecting treatment tolerance and efficacy. Abnormalities in thrombocyte counts can indicate issues with clotting and immune function, while elevated GGT levels may suggest liver dysfunction, which can affect drug metabolism and clearance.

In addition, experts from Erasmus MC Rotterdam have identified a set of ten crucial variables for predicting the response to chemotherapy. In alignment with the random forest models they propose CA19-9 to be the most important variable. In contrast to the predictions made by the models, the experts propose that CEA should be considered the second most important variable. This recommendation is supported by existing literature, which establishes CEA as another noteworthy tumor marker with strong prognostic value [40] [41]. Subsequent to CA19-9 and CEA, they suggest that GGT before values and the difference in SII should be taken into account, along with differences in NP and TB. However, it should be noted that the latter two variables are already encompassed within the SII calculation, as defined by Equation (B.4). Additionally, it might be valuable to know what variables would constitute the top 10 when considering them individually, independent of whether these are ‘before, after or difference’. In that case, in the random forest without outliers based on BR and GFR, the unique top 10 variables identified include NP, NLR, and PLR.

Top 10 most important Variables	**Random Forest Model full dataset**	**Random Forest Model No outliers**	**Random Forest Model top 10**	**Random Forest Only before**	**Advice from medical experts Erasmus MC**	**Random Forest No Outliers Unique top 10**
1	CA19-9 before	CA19-9 before	CA19-9 before	CA19-9 before	CA19-9 before	CA19-9
2	HB diff	CA19-9 after	CA19-9 after	GGT before	CA19-9 after	HB
3	TB diff	CA19-9 diff	TB diff	NP before	CA19-9 diff	TB
4	HB before	HB diff	GGT before	HB before	CEA before	GGT
5	GGT after	TB diff	AF after	K before	CEA after	BR
6	CEA before	GGT before	HB diff	TB before	CEA diff	AF
7	CR diff	GGT after	K before	Age	GGT before	K
8	GGT diff	BR after	GGT after	CEA before	SII diff	NP
9	TB before	AF after	BR after	BMI	NP diff	NLR
10	CA19-9 after	K before	CA19-9 diff	AF before	TB diff	PLR

Table 5.13: Overview of the top 10 ranked variables based on the ranksum calculated after a running 100 optimized random forests for each of the four datasets. These include the full dataset (n=194, Disease Control (n=120), Progressive disease (n=74)), the dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values (n=182, Disease Control (n=95), Progressive disease (n=87)) , the dataset containing only the top 10 variables (n=182, Disease Control (n=115), Progressive disease (n=67)), the dataset with only pre-chemotherapy values (n=182, Disease Control (n=115), Progressive disease (n=67)), the top 10 variable rank according to medical experts from the Erasmus Medical Centre Rotterdam in the Netherlands and the top 10 ‘unique’ variables in the dataset after removal of outliers based on GFR and BR.

One possible explanation why CEA is not present in the top 10 identified variables by the random forest model could be its strong correlation with CA19-9, as identified in the PCA analysis (see section 3.3 and Appendix B.3). Notably, CA19-9 has already demonstrated good predictive performance in literature studies and exhibits a stronger correlation with the final response compared to CEA, see Table 5.14. Previous research by Nicodemus and Malley [60] suggests that random forest tends to favor uncorrelated predictors during the construction of decision trees within the forest when using the Gini index. Additionally, Meng et al. [61] found that the strength of association between a predictor and the response significantly influences the impact of predictor correlation on random forest’s performance. In a simulation study conducted by Nicodemus et al. [62], a simple linear model consisting of 12 predictors was utilized. Among these predictors, four were strongly correlated (with a correlation of $r = 0.9$), and eight were uncorrelated. Three of the four correlated predictors and three of the eight uncorrelated predictors possessed non-zero coefficients. The findings of this study revealed that at the first split in the tree, correlated predictors were more frequently selected. However, as the parameter ‘mtry’ increased, the preference for uncorrelated predictors became more pronounced when considering all splits. Uncorrelated but strongly associated predictors tended to be selected more frequently across all trees in the random forest when the pool of potential predictor variables (mtry) was set to be larger. The increase in mtry allows for a broader range of candidate predictors, making it more likely for uncorrelated but relevant predictors to be chosen in the tree construction process. Specifically, the preference for correlated predictors as the first splitting variable was observed only when these predictors exhibited an

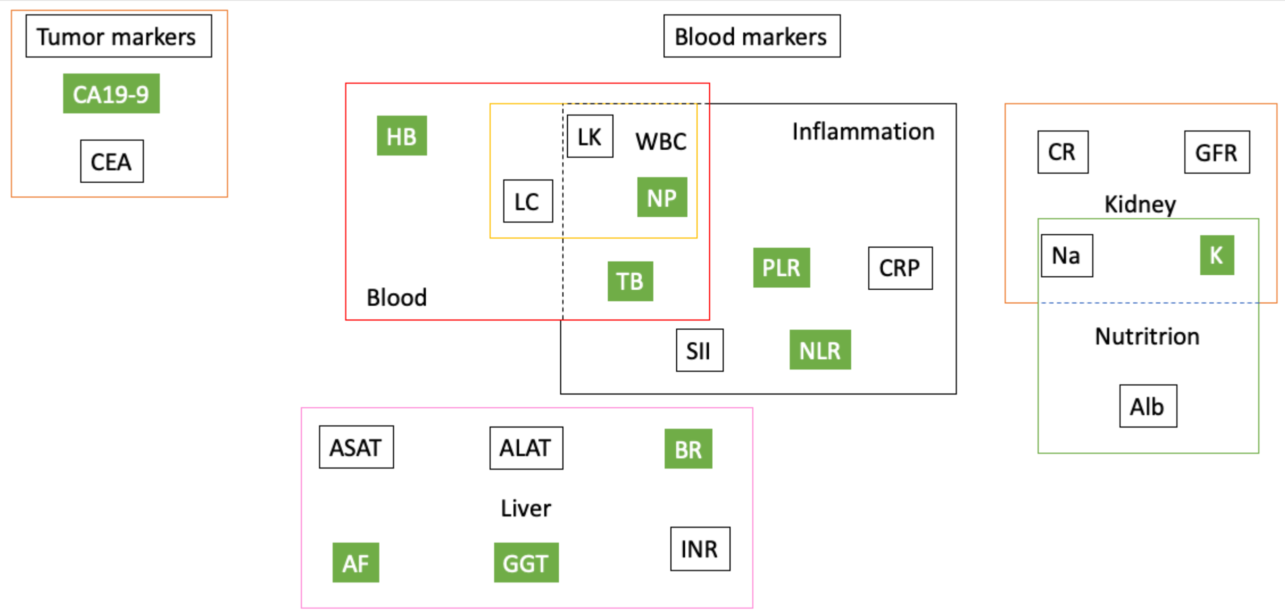


Figure 5.17: Overview of the groupings of the variables considered in the analyses with the top 10 identified variables in green, when considering only the 'unique' variables.

association with the outcome.

Furthermore, other potential reasons for CEA's lower variable importance in the random forest model might be attributed to the influence of other features. In the random forest algorithm, variable importance is not solely dependent on individual features but also on their interactions with other features in the dataset. Despite the high correlation between CA19-9 and CEA, they might interact differently with other influential features in the dataset. It is possible that CA19-9 has stronger interactions with certain influential features, leading to a higher importance score, while CEA's interactions might not be as impactful. Additionally, the proven prognostic and predictive value of CA19-9, particularly in the Lewis (+) population, could further contribute to its higher importance in the random forest model. As a result of its superior performance in predicting the final response, CA19-9 might be consistently favored over CEA during the tree construction process. Moreover, a potential reason why SII has not been included in the top 10 variables in the random forest might be due to the large number of missing values (98) compared to the other variables. Nevertheless, the random forest imputes these missing values based on the other data.

Variable	**Pearson's Correlation with Final response**	**p-value**
CA19-9 before	0.204	0.0035
CEA before	0.125	0.0756
CA19-9 after	0.207	0.0031
CEA after	0.138	0.0495
CA19-9 diff	-0.143	0.0419
CEA diff	-0.009	0.8986
HB diff	-0.163	0.0205
TB diff	-0.078	0.2659

Table 5.14: Estimated Pearson's correlation and corresponding p-value with the final response outcome.

5.2. PDP and ALE plots optimal random forest model without outliers

In the preceding section, we have successfully derived the optimal random forest models for predicting the final response to chemotherapy based on the measured blood variables present in the provided dataset. In this section, we intend to investigate the marginal impact of the most significant features on the predicted outcome of the random forest machine learning model. One approach to accomplish this is by the use of partial dependence plots (PDPs). However, due to the strong assumption of independence in PDPs, and taking into account earlier observations in Appendix B.1 and Appendix B.3, where we discovered substantial correlations among multiple variables, we will also construct accumulated local effect plots (ALE) plots for a more comprehensive analysis. As emphasized earlier in section 5.2, the selection of an appropriate grid size is of importance when constructing ALE plots. The rationale for choosing the grid size is demonstrated through Figure 5.18, which displays three ALE plots for the HB difference (mmol/L) using grid sizes of (a) 13, (b) 30, and (c) 182, respectively. These choices are based on the general rule of thumb of taking the grid size to be the square root of the number of observations. In this case we have 182 observations, hence $\sqrt{182} \approx 13.49 \approx 13$. However, since we want to see the impact of a smaller grid size, we have chosen 30 and 182 as well. Note that an excessively coarse grid size may mask potential complexities in the model, while an excessive number of intervals can lead to instability and fluctuations in the plots. Nevertheless, implementing a finer grid, such as 200, would be illogical as certain intervals would be devoid of observations. To illustrate this distinction, the plots in Figure 5.18 can serve as exemplars. Specifically, Figure 5.18(a) exhibits a comparatively smoother curve, while Figure 5.18(b) displays a greater degree of fluctuations, and Figure 5.18(c) exhibits the most pronounced fluctuations. Consequently, for the purpose of this analysis, a grid size of 30 is deemed appropriate for generating the ALE plots, striking a balance between capturing essential details and minimizing excessive instability. Additionally, given that the PDP and ALE plots for the DC and PD groups are opposites of each other, it suffices to present the plots for only one of the groups. However, to demonstrate this contrast, both the PDP and ALE plots will be shown for the most important variable, CA19-9 before. These plots can be seen in Figure 5.19.

Furthermore, PDP and ALE plots are primarily used to analyze and interpret the behavior of a model based on its learned patterns from the training dataset. These plots help understand the relationship between individual predictor variables and the response variable to gain insights into the effects and importance of different features on the model's predictions. Therefore, the PDP and ALE plots in Figures 5.19, 5.20, 5.21 and 5.22 below are generated using the training dataset to provide a visual representation of the final response to changes in specific variables. These plots are not intended to evaluate the model's generalization performance on unseen data, but rather to understand how it behaves on the data it was trained on.

The PDP and ALE plots in Figure 5.19 show the relationship between the variable CA19-9 before and the final response. The grey areas depicted in the PDPs represent 95% confidence intervals for the PDP curve. From the plots, it is evident that values below $\approx 1000 - 1500$ kU/L indicate a high probability of PD, while higher values suggest DC. The ALE plot closely aligns with the PDP curves, indicating a consistent effect of CA19-9 before on the final response across its range. In Figure 5.20, similar patterns emerge for CA19-9 difference and CA19-9 after. A negative CA19-9 difference implies a low probability of DC and a high probability of PD. Unlike the CA19-9 tumor marker plots, the HB and TB difference plots lack a clear boundary to effectively distinguish the two groups. The plots for the HB difference shows that a positive value is associated with DC, while a negative value indicates PD (see Figure 5.21(a) and (c)). For the TB difference, values within 0-200 $\times 10^9/L$ are linked to DC, compared to values outside this range, as seen in Figure 5.21(b) and (d). The difference in shape in the PDP and ALE plots for both HB and TB difference might be due to the underlying assumptions of PDPs discussed in section 4.2, therefore the ALE plot is preferred for interpretation purposes.

Lastly, the PDP and ALE plots in Figure 5.22 reveal insights into the classification probabilities for the DC and PD groups based on SII difference and CEA difference. For the SII difference, values between approximately [-500, 1800] indicate a high probability of DC, while negative values below -500 suggest a significantly higher likelihood of PD and values above 1800 do not show a clear distinction between the groups. These patterns are consistent in both the PDP and ALE plots. In contrast, for the CEA difference, negative values correspond to a high probability of DC, while positive values indicate a higher likelihood of PD. This relationship is consistent in both the PDP and ALE plots. The PDP and ALE plots for the other identified important variables are presented in Appendix D.2. Finally, it is important to note that in this case we are only considering the marginal probabilities of one variable on the final response and keep all the other variables constant and these plots are only intended to have some insights into the relationship between important predictors and classification probabilities for DC and PD. As a side note, the PDP and ALE plots presented in this section and in Appendix D.2 clearly show the non-linear relationships present between the variables and the final response. This might be why PCA was not able to distinguish between the two responses.

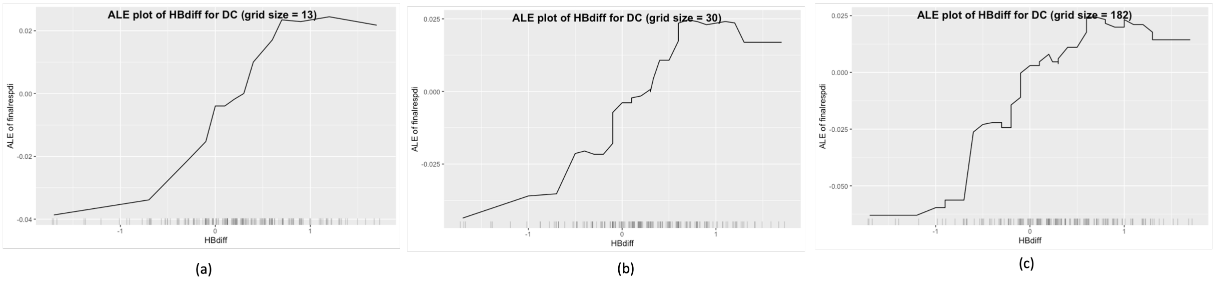


Figure 5.18: Accumulated Local Effect Plots of the Hemoglobin difference values ($mmol/L$) based on the random forest trained using the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)) using three different grid-sizes: (a) Grid-size of 13 (b) Grid-size of 30 (c) Grid-size of 182.

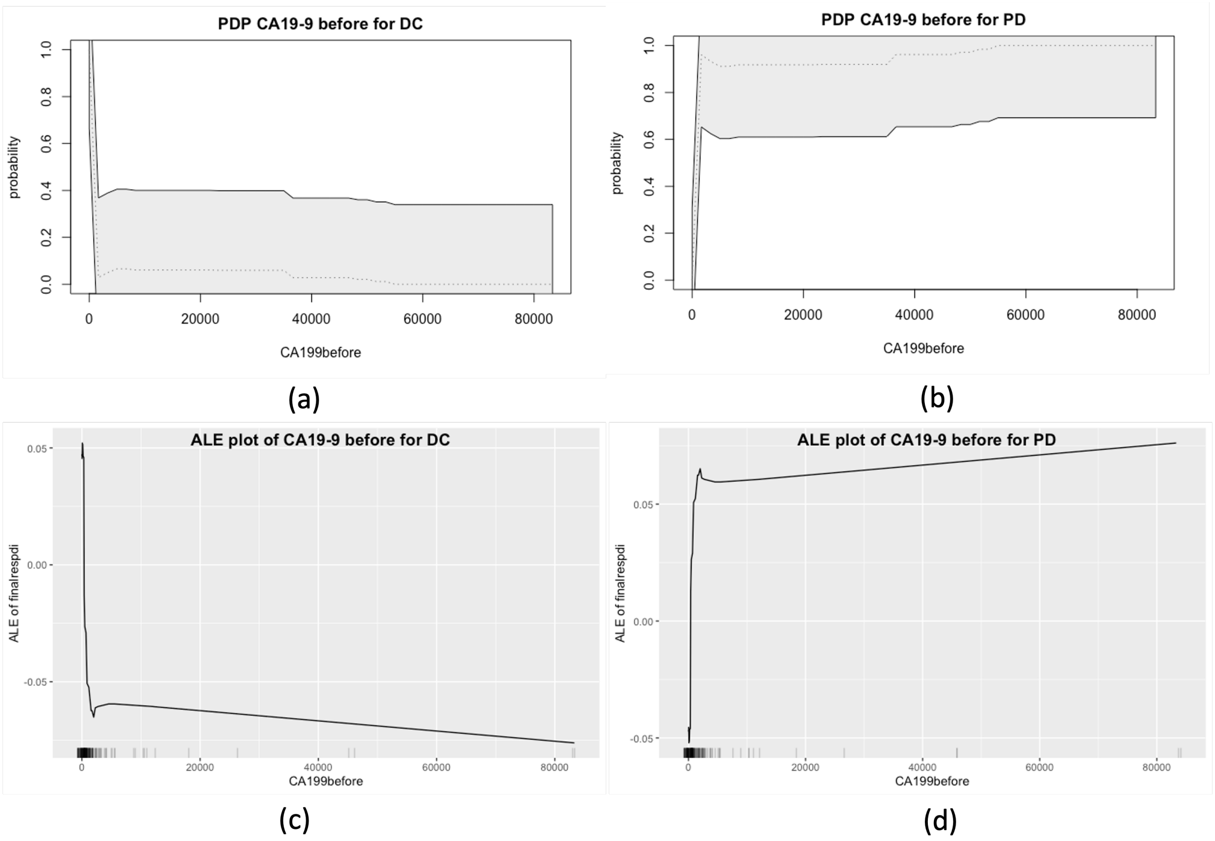


Figure 5.19: Partial Dependence Plots and Accumulated Local Effect plots of the CA19-9 before (kU/L) values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of CA19-9 before (kU/L) for DC as response (b) PDP plot with 95% confidence interval of CA19-9 before (kU/L) for PD as response (c) ALE plot of CA19-9 before (kU/L) for DC as response (grid size = 30) (d) ALE plot of CA19-9 before (kU/L) for PD as response (grid size = 30).

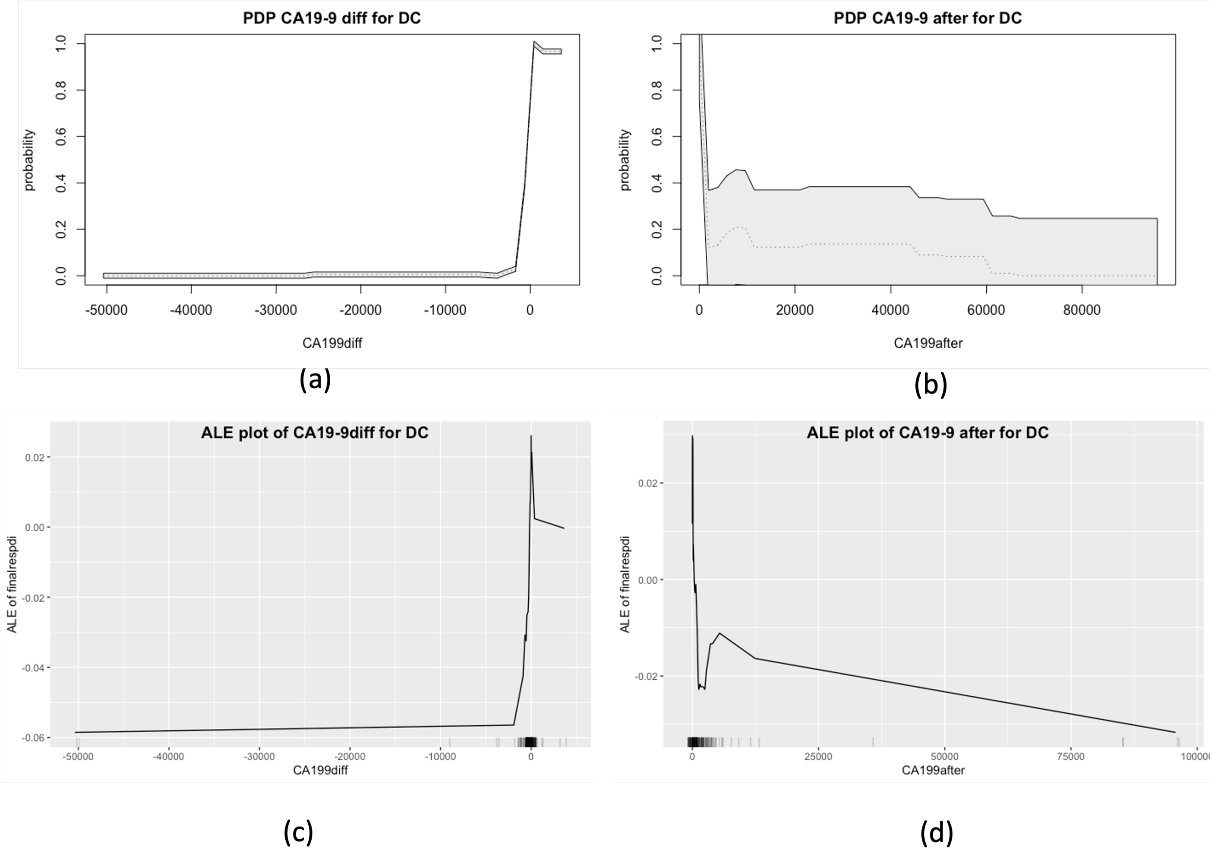
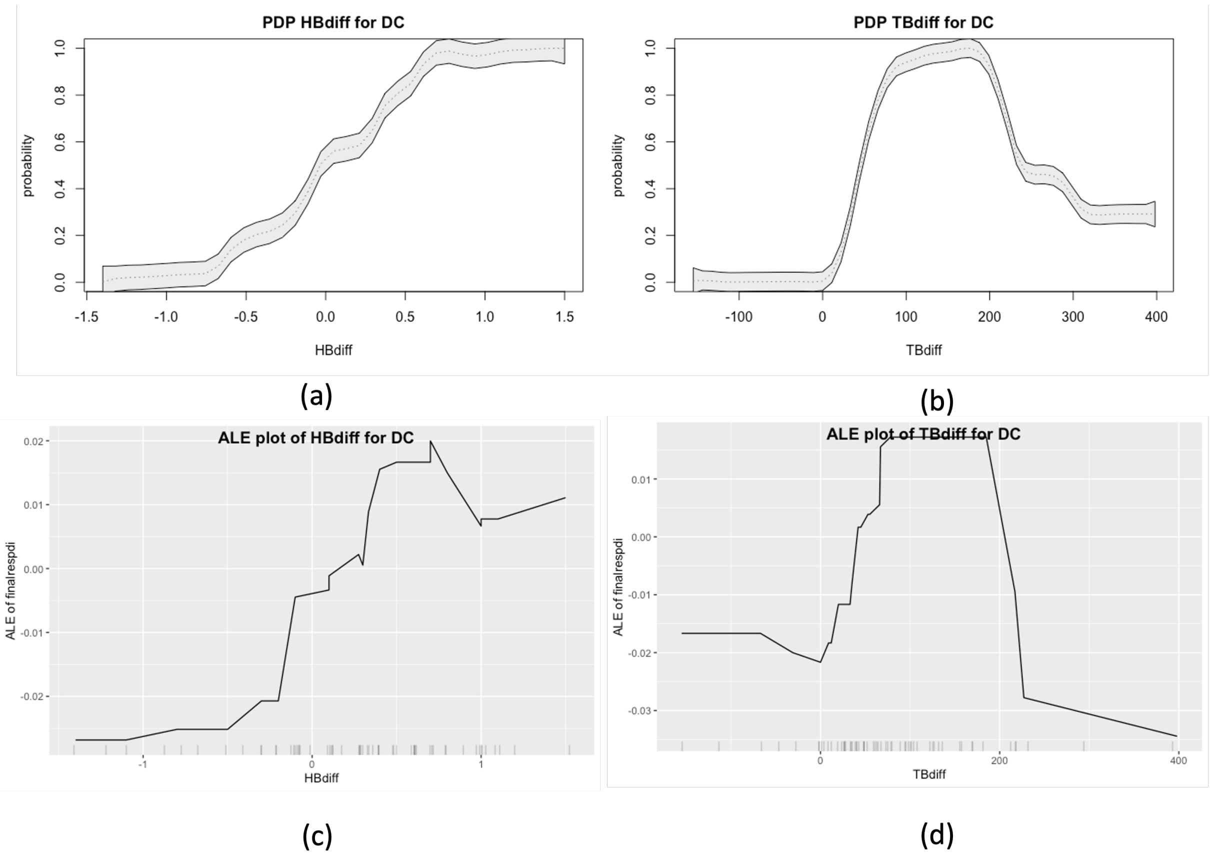


Figure 5.20: Partial Dependence Plots and Accumulated Local Effect plots of the CA19-9 difference (kU/L) and CA19-9 after the first chemotherapy cycle values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of CA19-9 difference (kU/L) for DC as response (b) PDP plot with 95% confidence interval of CA19-9 after (kU/L) for DC as response (c) ALE plot of CA19-9 difference (kU/L) for DC as response (grid size = 30) (d) ALE plot of CA19-9 after (kU/L) for DC as response (grid size = 30).



(a)

(b)

(c)

(d)

Figure 5.21: Partial Dependence Plots and Accumulated Local Effect plots of the Hemoglobin and Thrombocyte differences before and after the first chemotherapy cycle based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of Hemoglobin difference (mmol/L) for DC as response (b) PDP plot with 95% confidence interval of Thrombocyte count difference ($10^9/L$) for DC as response (c) ALE plot of Hemoglobin difference (mmol/L) for DC as response (grid size = 30) (d) ALE plot of Thrombocyte count difference ($10^9/L$) for DC as response (grid size = 30).

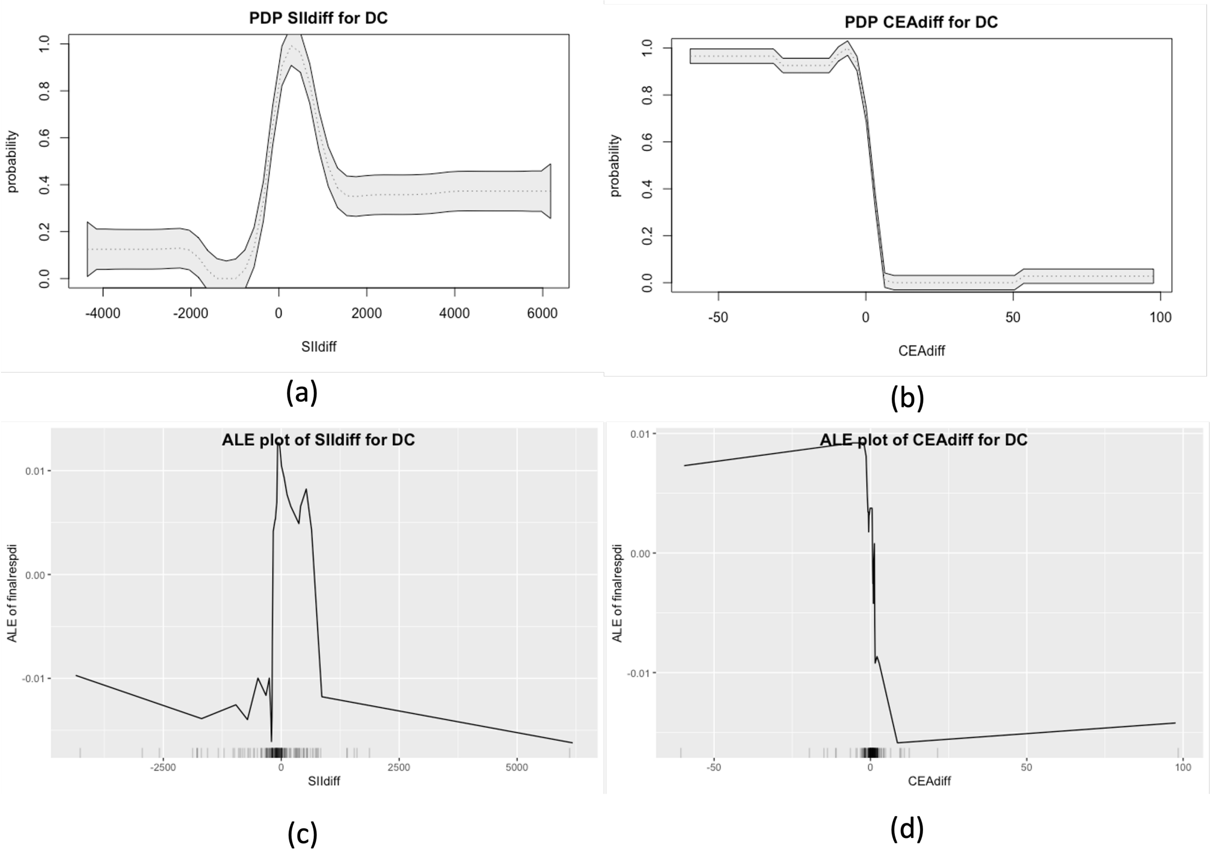


Figure 5.22: Partial Dependence Plots and Accumulated Local Effect plots of the Systemic inflammation index and CEA differences before and after the first chemotherapy cycle based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of Systemic Inflammation Index difference for DC as response (b) PDP plot with 95% confidence interval of CEA difference ($\mu g/L$) for DC as response (c) ALE plot of Systemic Inflammation Index difference for DC as response (grid size = 30) (d) ALE plot of CEA difference ($\mu g/L$) for DC as response (grid size = 30).

5.3. Interpretation of the findings for top 10 variables

In this section we want to combine all the findings of the exploratory data analysis and random forest modelling together. The focus will be on the top 10 most important variables identified using the random forest based on the dataset after removal of outlier values using GFR and BR. In addition, relevant studies will be mentioned to deepen the understanding from a medical perspective.

CA19-9 has been widely established in the literature as a diagnostic, prognostic, and predictive biomarker for patients with PDAC [24]. This research aligns with existing evidence, as initial data analysis already indicated that PD patients exhibited the highest CA19-9 values, followed by the SD group, with no discernible difference between genders. The Wilcoxon Ranksum test also showed a significant difference between the CA19-9 values between the DC and PD groups for the CA19-9 values measured before and after the first chemotherapy cycle (CA19-9 before $p = 0.002$ see Table 3.7, CA19-9 after $p = 0.001$ see Table 3.8, respectively in section 3.1), confirming the observations in several studies like [35] [36] [24]. Higher CA19-9 levels may indicate an increased risk of having PD as final response. During PCA, strong correlations were observed between CA19-9 and CEA values at diagnosis, before and after the first chemotherapy cycle. However, solely relying on CA19-9 and CEA was insufficient to have a distinct differentiation between the two groups. Surprisingly, when the covariance matrix was used, CA19-9 before and after emerged as the primary variables that accounted for 97% of the overall variance within the dataset when considering all the measured markers together. When considering only the Age, BMI, and tumor and blood marker difference values before and after the first chemotherapy cycle, CA19-9 difference stood out as the most important marker capturing 96% of the total variance in the data. This finding signifies that CA19-9 levels, both before and after diagnosis as well as the difference, play a crucial role in explaining the majority of the variability in the dataset.

Moreover, after training and optimizing random forest models, and conducting variable importance analysis using the mean decrease in the Gini index, CA19-9 before consistently remained the most significant variable across all optimal models. After removal of outliers based on GFR and BR values, CA19-9 after emerged as the second most important variable after the CA19-9 before variable. This suggests that the CA19-9 value determined before initiating chemotherapy can serve as an early indicator of the patient's ultimate response to chemotherapy, while after one cycle, both before and after values are important. Additionally, the analysis of PDP and ALE plots indicated that a “cut-off" value of approximately 1000-1500 kU/L could distinctly differentiate DC from PD, with higher CA19-9 levels before treatment correlating to a greater probability of being classified as PD. Similarly, a negative CA19-9 difference after the first cycle of chemotherapy, implying an increase in CA19-9 levels, was also prognostic for a final response of PD. Lee et al. also found that reduced CA19-9 concentrations after neoadjuvant chemotherapy has comparable prognostic performance in PDAC patients with borderline resectable and LAPC [63].

A major limitation of CA19-9 as biomarker for PDAC is the Lewis (-) population, whose serum CA19-9 tends to be consistently low because they do not produce CA19-9. To address these challenges, several studies have explored potential adjustments to enhance the usage of CA19-9 measurements. A commonly used approach involves normalizing the CA19-9 value by the bilirubin level in patients with jaundice like the study by Liu et al. [64]. Huang et al. [65] sought to refine the preoperative serum CA19-9 value by dividing it by liver enzyme levels of ASAT and GGT in pancreatic cancer patients with jaundice. They compared the adjusted CA19-9 values with previously reported studies to evaluate their prognostic value in predicting patient survival. The study revealed that combining the scoring of CA19-9/ASAT and CA19-9/GGT emerged as an independent predictor of overall survival in patients with PDAC whose preoperative serum total bilirubin exceeded 102.6 μmol/L [65]. Additionally, the study by Li et al. [66] showed that the preoperative γ-glutamyltransferase-to-albumin ratio (GAR) is an independent prognostic factor for prediction of surgical outcomes in PDAC patients. Future investigation is necessary to be able to draw definitive conclusions on the prognostic and predictive power of these ratios. Moreover, Ermiah et al. [67] also found a strong correlation between CA19-9 and CEA values. Higher CEA and CA19-9 serum levels were significantly associated with a malignant phenotype and shorter overall survival rates as well as lower disease-free survival rates. Nevertheless, CEA did not show to be as prognostic as CA19-9 in this research as in the optimal random forest after removal of outliers based on BR and GFR values, no CEA variable was present in the top 10 variables based on the variable ranksum. However, CEA before does seem to be important in the optimal model based on the full dataset as well as the optimal model using only the pre-chemotherapy values, ranking 6^{th} and 8^{th} respectively.

The analysis revealed that hemoglobin difference (HBdiff) and thrombocyte difference (TBdiff) were ranked as the second and third most influential variables in the model, following CA19-9. However, the variables alone were insufficient in predicting response to chemotherapy, which supports findings by for instance Wu et al. [68], which found that solely using the HbA_{1C} level was insufficient to be prognostic. However, this study looked at the association of elevation in glycated hemoglobin (HbA_{1c}) (see Appendix E) instead of the normal hemoglobin values. After examination of the hemoglobin levels before and after the initial chemotherapy cycle in the initial data analysis in this study, no evident differences were observed between the final response groups, and HBdiff did not exhibit a distinct separation either. However, the Wilcoxon ranksum test did show a significant difference between the DC and PD groups with a p-value of $p = 0.032$ for the HB before and $p = 0.041$ for the HB difference values, but not for the HB after value, which only had a p-value of 0.430 (see Table 3.7, Table 3.8, Table 3.9, respectively). Additionally, it is noteworthy that hemoglobin was one of the few variables that showed a normal distribution. Moreover, PCA indicated that HBdiff displayed weak correlations with other blood cell variables and did not significantly contribute to the primary principal components. The significance of the HB before and after the first cycle of chemotherapy indicates its relevance in the context of PDAC and its potential implications for treatment outcomes. Changes in hemoglobin levels can indicate various physiological conditions including anemia, which is characterized by a decrease in red blood cell count or hemoglobin concentration. Significant changes in hemoglobin levels might indicate the presence of anemia or other underlying health conditions that can impact treatment outcomes. Cancer patients often experience anemia as it occurs in more than 40% of the cases [69]. Especially in patients treated with chemotherapy, the incidence of anemia may rise to 90% [70]. Caro et al.[71] found that in cancer patients, anemia is identified as an adverse prognostic factor. Anemia can be caused by various factors such as hemolysis, herediatry diseases, nutritional inefficiencies or other underlying inflammatory diseases [72]. It can contribute to fatigue, reduced oxygen supply to tissues and impaired overall health [1] which may influence the patient's response to chemotherapy. Dicato et al. [73] demonstrated that inflammatory cytokines such as tumor necrosis factor-α (TNF-α) and interleukin-6 (IL-6), among others, play a major role in the pathophysiology of anemia in cancer patients.

On the other hand, thrombocytes (TB) emerged as the most crucial variable in PCA when using the covariance matrix, while contributing solely to PC4 in PCA. Data analysis showed a slight decreasing trend in TB count after the first chemotherapy cycle, without clear differentiation among the final response groups. Contrastingly, the most substantial differences in TB levels were observed in the PD and SD groups. This is reflected in a $p = 0.013$ in the initial data analysis, see Table 3.9 in the TB difference variable. Nevertheless, unlike CA19-9, no definitive cut-off value to distinguish the PD group from the DC group was evident in the PDP and ALE plots for both HBdiff and TBdiff. However, regarding HBdiff, in general an increase in hemoglobin levels after chemotherapy likely indicates an inclination towards PD. Likewise, elevated TB levels after chemotherapy suggests a similar tendency. Yet, a substantial increase in TB may also serve as an indicator of PD. Recent research has indicated that TB have broader implications in maintaining a balance between health and disease. The presence of tumor cells can influence TB counts through various mediators, cytokines, and tumor cell-induced platelet activation, which represents the initial step towards cancer-induced thrombosis [74]. Conversely, emerging evidence suggests that complex interactions between tumor cells and circulating TB may significantly impact tumor growth, dissemination, and angiogenesis. This reciprocal relationship between tumors and TB appears to promote tumor-induced thrombosis while facilitating the development and metastasis of malignant tumors [75] [76]. Therefore, a bidirectional relationship between tumors and thrombocytes seems to exist, promoting tumor-induced thrombosis on one hand and the establishment of malignant tumors and their metastasis on the other.

Another influential variable in the analysis was γ-Glutamyl Transferase (GGT) before and after values. In the random forest model, GGT before ranked as the second most significant variable following CA19-9 before. GGT is a liver enzyme involved in glutathione metabolism, as well as the breakdown of food, drinks, and waste materials and reflects the liver functionality. The initial data analysis did not reveal any evident differences across the final response groups or between genders based on GGT levels before or after chemotherapy. However, a notable observation was that almost all GGT values before and after the first chemotherapy cycle exceeded the predetermined healthy range of 40 U/L for males and 25 U/L for females. Although most GGT before values were higher compared to GGT after values, there was no clear distinguishable pattern as it varied among each patient. Nevertheless, GGT before and after values exhibited strong correlations with each other, as well as with Alkaline Phosphatase (AF) before and after values, as demonstrated by the correlation matrix during PCA. However, their contributions to the first and second principal components (PC1 and PC2) were not as prominent compared to Alanine Aminotransferase (ALAT) and Aspartate Aminotransferase (ASAT) values. From literature it is becoming more and more clear that inflammation is a critical component in tumor initiation and progression [77]. Especially oxidative stress can activate a series of transcription factors, leading to expressions of pro-inflammatory molecules and therefore promoting the transformation of normal cells into tumor cells, tumor cell survival, proliferation and invasion [78]. GGT plays a pivotal role in maintaining sufficient levels of Glutathione, which protects the cells from oxidative damage. GGT has been shown to be elevated under oxidative stress [79]and a common variation in the GGT1 gene was found to be involved in pancreatic carcinogenesis [80]. Based on the observations in this research as well as the other literature mentioned, GGT does not only inform about liver functionality, but is also reflects internal inflammation.

In addition to GGT, Bilirubin (BR) and Alkaline Phosphatase (AF) values after the first chemotherapy cycle emerged as the most significant variables identified by the optimal random forest model. Similarly to GGT, BR and AF serve as indicators of liver functionality. Data analysis in this research revealed a significant decrease in BR values after the first chemotherapy cycle, transitioning from a high, unhealthy range to a lower, healthier range. Yet, no distinct differences were observed across the final response groups. Considering that the outliers were also partially identified based on BR values, it would be prudent to ensure the presence of healthy BR levels prior to initiating chemotherapy. As for AF levels, there was considerable variation among patients, with some showing an increase and others experiencing a decrease after the first chemotherapy cycle. No clear patterns emerged among the final response groups. The only noticeable trend was the largest differences in AF levels before and after the first chemotherapy cycle were observed in the PD group. Still, since this pattern was only evident in three patients, drawing conclusive interpretations is not feasible. Notably, AF values exhibited a strong positive correlation with GGT values, while BR showed correlations only with its own measurements (before, after, and difference). Consequently, in the PCA using the correlation matrix, AF after and GGT after values predominantly contributed to the PC2, while BR after became apparent in the fourth or fifth dimension. Conversely, in the PCA conducted using the covariance matrix, BR after was primarily prominent in the tenth dimension.

[1] This might be a sign of bone marrow depression according to the medical experts from Erasmus MC

From a medical perspective, the importance of GGT, BR, and AF values lies in their association with liver functionality. The decrease in BR values within a healthier range after the first chemotherapy cycle suggests potential improvements in liver function. Yet, the absence of clear distinctions among the final response groups implies that BR alone may not serve as a reliable predictor of chemotherapy response. Similarly, the variability in AF levels and the lack of consistent patterns among response groups hinder its utility as a definitive indicator. The strong correlations observed between GGT and AF highlight their interconnectedness in reflecting liver function. This suggests that liver enzyme levels after the first cycle of chemotherapy may provide valuable insights into how patients are likely to respond to subsequent cycles of treatment. Consequently, this information can aid clinicians in making informed decisions regarding the continuation or discontinuation of chemotherapy. The association between higher liver enzyme levels and PD may be attributed to the impact of pancreatic cancer on liver function or the overall disease progression itself. Additionally, liver enzyme levels can be influenced by factors such as inflammation and cholestasis, which commonly occur in pancreatic cancer patients. Early detection of elevated liver enzyme levels may prompt further investigations and adjustments to the treatment plan, potentially leading to improved patient outcomes. A properly functioning liver is crucial for the processing and elimination of drugs, including chemotherapeutic agents. Impaired liver function can affect the metabolism and excretion of drugs, thereby potentially influencing the response to chemotherapy.

Potassium (K) values before chemotherapy were found to be the tenth most important variable in the optimal random forest model. Although this variable did not appear in the optimal random forest model trained on the full dataset, it ranked fifth in importance when considering only the before chemotherapy values. The presence of electrolyte abnormalities, including hyperkalemia, hyponatremia, hypophosphatemia, and hypercalcemia, is commonly observed in cancer patients, which can be attributed to the cancer itself or its treatment [81]. Such electrolyte disorders in cancer patients have been associated with unfavorable prognoses. Hyperkalemia, characterised by higher levels of serum or plasma K levels than normal, is in cancer patients often linked to conditions such as acute kidney injury, rhabdomyolysis, or tumor lysis syndrome [81]. Furthermore, the data analysis revealed a slight decrease in K levels after the first chemotherapy cycle. However, drawing definitive conclusions from these findings is challenging due to the considerable variation in K levels among patients, which can be influenced by factors such as the occurrence of side effects (e.g., vomiting, diarrhea) and the use of medications. These factors can either decrease or increase K levels in individuals. Nevertheless, no clear distinction was observed between the final response groups based on this measured blood marker. The correlation matrix used in the PCA provided insights into the relationships between K values and other variables. Specifically, K after values displayed positive correlation with K before values and a strong negative correlation with K difference. However, minimal correlations were observed between K values and other measured variables. The PDP en ALE plots in Figure D.3(a) and Figure D.3(c) provide additional support to this finding, indicating that higher K levels prior to chemotherapy are indicative of PD. Recent research by scientists at the National Cancer Institute's Center for Cancer Research revealed that harnessing T-cell "stemness" could enhance cancer immunotherapy [82]. Surprisingly, pre-chemotherapy K levels, associated with electrolyte imbalances in cancer patients, align with treatment response and may impact the effectiveness of cancer immunotherapies. Further studies are required to establish potassium's role in cancer and chemotherapy response.

Furthermore, the initial data analysis revealed an increase in Systemic Inflammation Index (SII) after the first chemotherapy cycle (detailed analysis provided in Appendix B.1). However, it should be noted that SII did not consistently increase after the first cycle; in some patients, it exhibited a slight decrease. Nevertheless, for these patients, the SII levels were already close to the considered healthy range of 900. After the first chemotherapy cycle, a significant number of SII values exceeded the healthy range. Furthermore, the difference in SII before and after the first chemotherapy cycle indicated that the largest differences occurred in the SD group, while the smallest differences were observed in the PR group. Nonetheless, no significant differences were seen in the SII variables between the DC and PD groups. Since SII is calculated according to Equation (B.4), it exhibits strong positive correlations with neutrophil-to-lymphocyte ratio (NLR) and platelet-to-lymphocyte ratio (PLR), as demonstrated in the correlation matrix obtained from the PCA (refer to Appendix B.4 for detailed information). SII difference emerged as the most important variable in PC1, followed by NLR difference and PLR difference. Outlier analysis revealed that after removing outliers based on the interquartile range, SII difference exhibited a normal distribution (p-value = 0.558). These findings are partly confirmed by the PDP and ALE plots presented in Section 5.2. Specifically, in Figure 5.22(a) and Figure 5.22(c) it can be observed that a increase in SII values strongly predict a patient as having PD, while values between 0-900 indicate a high probability of classifying a patient as having DC. After an SII difference value of approximately 1000, the classification probability becomes evenly distributed between the two groups. On the contrary, Murthy et al. [53] demonstrated that a high SII was associated with unfavorable outcomes in PDAC. Specifically, they found that an $SII > 900$ before resection in patients treated with neoadjuvant therapy was linked to a worse prognosis compared to an $SII \leq 900$ [53]. Interestingly, when considering the variable importance ranking based on the random forest model, SII did not emerge among the top 10 ranking factors, which might be due to the large number (n=98) of missing values. In the 100 runs conducted it only appeared once at rank 8 and 9, respectively.

Moreover, the data analysis revealed a slight decrease in TB count and a significant increase in NP after the first chemotherapy cycle, while LC were minimally affected, only exhibiting a small increase. Consequently, the rise in SII is primarily attributed to the substantial increase in neutrophil levels. Several studies support the prognostic significance of the NLR in pancreatic cancer patients. For instance, Iwai et al. [83] demonstrated that NLR independently predicts outcomes in patients with unresectable pancreatic cancer. Additionally, Yang et al. [84] reported that elevated NLR in peripheral blood is associated with poor prognosis in PDAC patients, as increased neutrophil infiltration creates a favorable tumor micro-environment for cancer progression. Consequently, it may be advisable to consider NLR as a prognostic factor instead of SII. However, when assessing the top 10 variables, it was observed that the neutrophil count only ranked among the top 10 factors in the optimized random forest model based on pre-chemotherapy values. This suggests that the neutrophil count before initiating chemotherapy may already possess strong prognostic value in determining the patients' ultimate response to treatment. Nevertheless, referring back to the data analysis conducted in Appendix B.1.3, it was observed that neutrophil levels for almost all patients were within the healthy range before chemotherapy, and they rapidly increased after the first cycle, without clear differentiation between the final response groups supported by p-values > 0.05. Conversely, when considering the NLR value, rather than the neutrophil count alone, the analysis in Appendix B.1.3 indicated that the NLR value remained relatively stable before and after the first chemotherapy cycle in the PR group, while in the PD and SD groups, the NLR significantly increased for many patients following the first cycle. Hence a clinical implication is that stable NLR levels have higher probability of DC.

Lastly, the variables 'Age' and 'BMI' were found to be significantly important in the random forest model based on pre-chemotherapy values. Despite Age and BMI not acting as independent predictive factors found in the initial data analysis, their significance becomes evident when considered in combination with other markers in the random forest model. This was also seen in the PCA analysis in Section 3.3 with the full analysis present in Appendix B.3, Age determined one principal component and BMI another, indicating that they are both uncorrelated and capture different structures in the dataset. However, BMI and Age alone were not sufficient to make a distinguishment between the two final response groups. In literature a prognostic effect of BMI remains unclear for patients under therapy for pancreatic cancer as published results are contradictory [85] [86]. Nevertheless, the study by Sclick et al. [23] did identify a BMI>25 to be a negative prognostic indicator for OS and low BMI as predictor of early treatment related toxicity. These contradictions may arise due to the fact that studies do not always analyze homogeneous patient populations in regard to inclusion of obese or overweight patients. Therefore, the only conclusion that can be drawn is that BMI and Age could be taken into account when deciding to give a patient chemotherapy or not, but these are not the most important factors. However, it can be said that generally obese patients have a higher probability of having other diseases such as high blood pressure (hypertension), high LDL cholesterol, and coronary heart disease [87]. These other health-related conditions could impact their response to chemotherapy as well. On the other hand, the older the patient, the more likely it is that they may have other health-related issues, such as osteoarthritis, chronic obstructive pulmonary disease, and diabetes, according to the World Health Organization [88]. All these other health conditions can also affect their response to chemotherapy.

5.4. Clinical Implications

From the findings of all the analyses in this research, clinical implications can be deduced. It is important to take these clinical implications as *suggestions* and not as rigorous rules. In addition, the analysis in this study only focused on PDAC patients treated with FOLFIRINOX and took into account only the patients who have reported CA19-9 values. The following key recommendations for clinical practice, based on the analysis of tumor and blood markers, can be made:

1. **Comprehensive measurement of tumor and blood markers:** It is advisable to measure tumor and blood markers prior to initiating chemotherapy. This provides essential baseline information to assess the patient's condition and potential treatment response. Additionally, measuring these markers after the first cycle of chemotherapy can offer valuable insights into the early treatment effects and guide subsequent therapeutic decisions. Especially the top 10 important tumor and blood markers should be taken into rigorous consideration. These are: CA19-9 before, CA19-9 after, CA19-9 difference, HB difference, TB difference, GGT before, GGT after, BR after, AF after and K before.
2. **Integration of marker values into predictive models:** The measured marker values should be incorporated into the optimized predictive random forest models to estimate the patient's (potential) response to chemotherapy. These models can assist doctors and other healthcare professionals in evaluating the likelihood of treatment success and tailoring the treatment plan accordingly. However, it is crucial to remember that these predictions are based on historical data and should be considered as *indications* rather than definitive decisions of individual patient outcomes.
3. **Consideration of baseline characteristics:** In addition to tumor and blood markers, it is important to consider various patient-specific factors, including age, BMI, history of chemotherapy, history of malignancy, family history, disease stage, smoking history, alcohol usage, diabetes, pancreatitis, and other relevant health conditions. These baseline characteristics provide a broader context for treatment decision-making and can help to personalize the therapeutic approach.
4. **Individualized approach:** Each patient is unique, and treatment decisions should be tailored to their specific circumstances. While predictive models and marker values offer valuable insights, clinical decisions should always be made in conjunction with patient preferences and individual considerations. The discussion between healthcare professionals and patients regarding the initiation and continuation of chemotherapy should involve a comprehensive evaluation of all available information and together with the patient him or herself.
5. **Management of side effects:** Once chemotherapy is initiated, it is essential to proactively manage potential side effects. Medications such as granulocyte-colony stimulating factor (G-CSF) to stimulate the production of neutrophiles, granulocytes and blood cells and erythropoietin-stimulating agents to boost red blood cell levels can be considered to mitigate adverse effects and improve patient well-being.

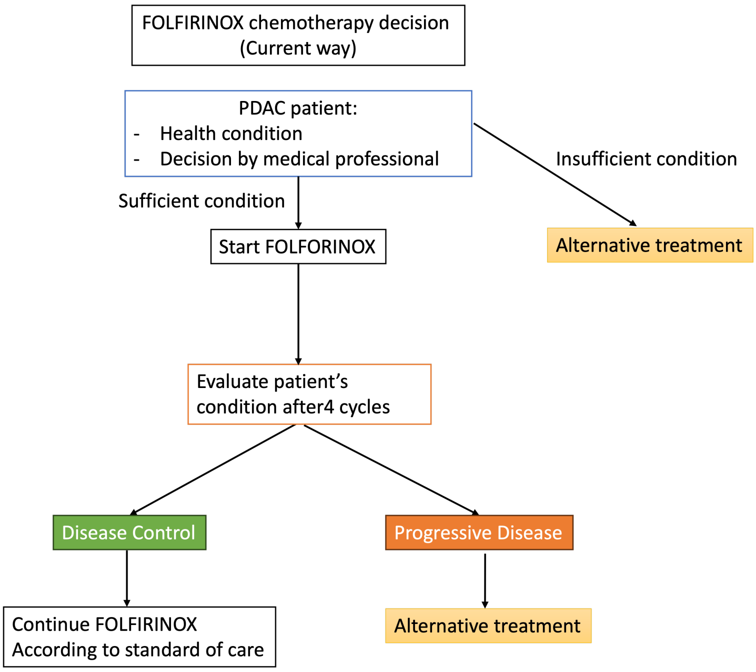


Figure 5.23: Current way of deciding FOLFIRINOX chemotherapy, the standard of care in the Netherlands is provided in the Richtlijnendatabase [89] [90].

The administration of FOLFIRINOX chemotherapy is currently determined based on the patient's health condition, as illustrated in Figure 5.23. When the medical professional assesses that the patient is in a suitable health state to undergo chemotherapy, four cycles of FOLFORINOX are administered. Following this, treatment continuation or alternative therapies are decided based on the patient's response to the treatment, whether it shows disease control or progression of disease based on radiologic imaging and circulating tumormarkers [89] [90]. To enhance the treatment decision-making process, an alternative approach is proposed, depicted in Figure 5.24. This approach involves employing the random forest model before initiating chemotherapy, utilizing measurements of blood and tumor values. Based on the model's predictions of DC or PD, chemotherapy is either initiated or a discussion is held with the patient to evaluate whether to proceed with chemotherapy or consider other options. After the first cycle of FOLFORINOX, tumor and blood values are measured again and fed into the random forest model to predict the patient's response (DC or PD). Treatment continuation is determined for cases of DC, while for those showing PD the decision is made considering the patient's overall health condition, potentially leading to alternative therapies.

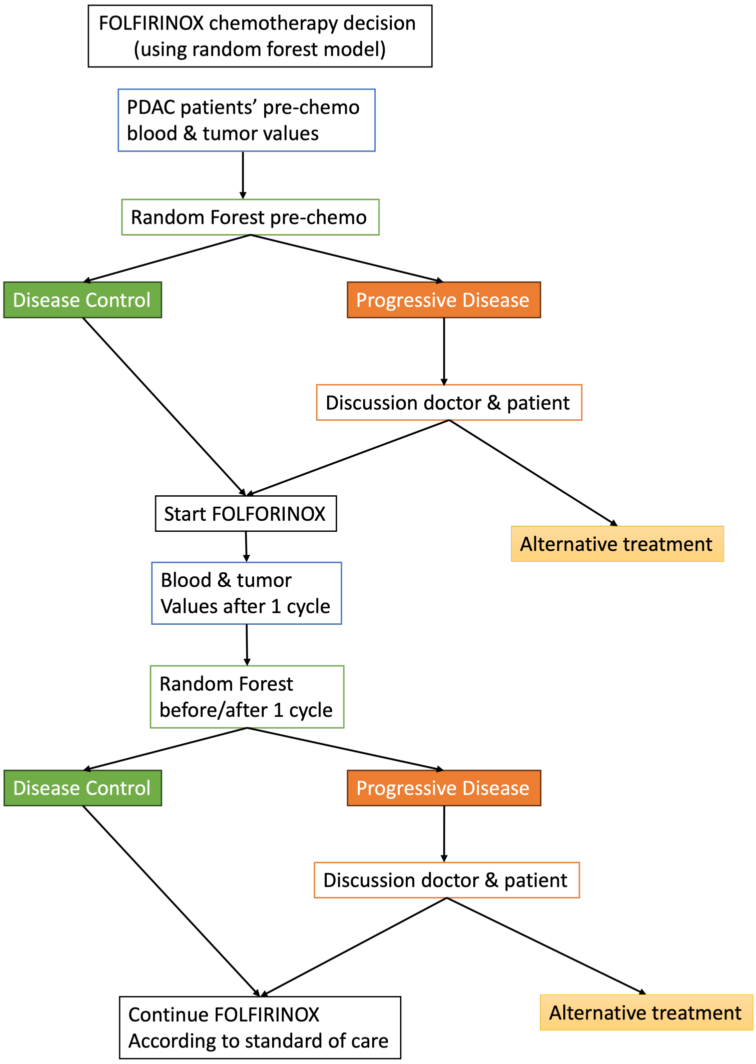

FOLFIRINOX chemotherapy decision (using random forest model)
PDAC patients' pre-chemo blood & tumor values
Random Forest pre-chemo
Disease Control
Progressive Disease
Discussion doctor & patient
Start FOLFORINOX
Alternative treatment
Blood & tumor Values after 1 cycle
Random Forest before/after 1 cycle
Disease Control
Progressive Disease
Discussion doctor & patient
Continue FOLFIRINOX According to standard of care
Alternative treatment

Figure 5.24: Proposed way of decision making for FOLFIRINOX chemotherapy based on the optimized random forest models. The standard of care in the Netherlands is provided in the Richtlijnendatabase [89] [90].

6 | Discussion

This Master thesis aimed to conduct a thorough data analysis in order to build a robust prediction model to classify PDAC patients based on measured tumor and blood markers for their response to FOLFIRINOX chemotherapy. Stratifying PDAC patients for chemotherapy is crucial. Not only for individual patient benefit by reducing treatment-associated morbidity and mortality, but also for addressing the socioeconomic challenge posed by increasing cancer incidence worldwide and rising healthcare costs [1]. Ineffective chemotherapy treatments and associated toxicities represent significant financial burdens in the management of PDAC patients and require urgent improvement. Machine learning models have shown potential in breast cancer prediction, like the random forest model, with superior predictive performance among various other classification models that Meti et al. considered [91]. Similarly, Guo et al. [92] proposed a classification model for cervical cancer patients. However, robust machine learning models that use blood and tumor markers for predicting chemotherapy response in PDAC are scarce. Nasief et al. did propose a delta-radiomics-based machine learning approach using cross-validated support vector machines to predict distant metastasis after chemo-radiation therapy in pancreatic cancer. Yet, they only showed that the combination of delta-radiomics, CA19-9 and PanIN was a potentially predictive biomarker [93].

Currently, about 60% of all PDAC patients in the Netherlands does not receive treatment but primarily receive best supportive care to manage symptoms [2]. These treatment decisions rely predominantly on CT-scans, CA19-9 and clinical experience of physicians. However, non-metastatic and metastatic patients receiving only supportive care have a median survival rate of 1.4 and 3 months, respectively, whereas FOLFORINOX chemotherapy demonstrated a median survival of 8.0 months [2]. Therefore, it is interesting to predict whether a patient would benefit from receiving FOLFIRINOX based chemotherapy. The proposed models in this study complement the existing approaches by providing these treatment outcome predictions. To demonstrate the predictive capability, if the entire dataset ($n = 203$) is run in the optimized model, it would predict $\frac{44}{203} \approx 22\%$ of patients as PD and $\frac{159}{203} \approx 78\%$ as DC [1]. Of these 44 patients, only 4 were misclassified as PD instead of DC. However, the model is trained using 70% of the same dataset, but it still shows a good classification performance of the model. In practice, it means we would 'save' 40 patients from ineffective chemotherapy. Nevertheless, every patient who is classified as PD would undergo a discussion with a medical team to decide whether to start chemotherapy or not, hence the misclassified patients might still be offered chemotherapy. It is important to note that the random forest model's inherent randomness can lead to varying numbers of misclassifications in different runs. For instance, in one run, only one patient may be misclassified, while in another run, two patients may be misclassified. This variability arises from the random selection of samples and features during the model's construction, which can influence the final outcomes across multiple runs [2].

In terms of cost savings, consider a hypothetical scenario with an estimated 3200 new PDAC patients in 2023 [3]. Assume that approximately 40% of these patients receive treatment and among them, 30% receive FOLFIRINOX. This means that about 384 patients[4] would be treated with FOLFIRINOX chemotherapy [2]. Suppose that based on the current data, around 22% of patients exhibit PD. That means that approximately 85 patients could be spared unnecessary chemotherapy treatment. Attard et al.[94] reported the cost of one cycle of FOLFIRINOX to be ≈ \$1633.21 ≈[5] €1480.51. A typical treatment consists of 8 cycles for (borderline) resectable or LAPC and 12 cycles for metastatic PDAC, respectively in the Netherlands [2]. Thus, a total cost savings of approximately €11,844.08 to €17,766.12 per patient exhibiting PD could be achieved. According to the NKR [6] in the Netherlands, approximately 43% of patients are diagnosed with borderline resectable or LAPC and 57% metastatic PDAC [2]. In total, under the given assumptions, potential annual cost savings of approximately $0.43 \times 85\times$ €11,844,08 + $0.57 \times 85\times$ €17,766.12 ≈€1,293,670 ≈ €1,3 million could be realized. This calculated figure concerns specifically to the cost of FOLFIRINOX chemotherapy and does not include additional expenses like supplementary medication, hospital admission, or other interventions required to manage side effects.

Despite FOLFIRINOX being the current standard treatment for borderline resectable, LAPC and metastatic PDAC patients [16], for PDAC patients with resectable disease, alternative therapies like chemo-radiotherapy or surgery are currently the standard. Consequently, the developed random models can be adapted by training them on datasets comprising exclusively (borderline) resectable PDAC patients or locally advanced and metastatic PDAC patients. Subsequently, predictions can be made for this subgroup of PDAC patients. In addition, the model can be applied to various chemo(radio)therapy treatments, not limited to FOLFIRINOX alone. It requires training on a dataset specific to the chosen treatment, using the same or a greater number of measured tumor and blood markers, to predict the final response to therapy. Then, the model's performance can be validated using an independent test set.

[1] The DCR falls within the 95% CI established in literature [4]

[2] Despite the randomness, the model achieved an average accuracy of 0.94 across 100 runs, misclassifying 0-5 patients.

[3] This assumption is based on the statistics from [25].

[4] in 2019, 368 patients received FOLFIRINOX as first-line treatment, while 458 patients received it as second-line treatment[2].

[5] Based on current state of \$1.00 ≈ €0.91 [95].

[6] Dutch Cancer Registry

Strengths of the random forest models developed in this study are that it uses easily obtainable data, namely tumor and blood biopsies, an interpretable approach (forest of decision trees), provides variable importance rankings and can capture nonlinear relationships. As a result, the models are able to help with personalized treatment decisions, optimize resource allocation, reduce treatment burden and contribute to research and development efforts in this field. In addition to the biomarkers investigated in this study, integrating other genetic and immunological markers with potential predictive value into the dataset could further enhance the model's predictive capabilities, such as investigated by Van der Sijde et al [96]. These markers may include cytokines, macrophages, specific T- or B-cells, tumor DNA, or immune checkpoints. Incorporating such specific markers in future studies would provide a more comprehensive understanding of the underlying mechanisms and improve the accuracy of the predictive model, ultimately facilitating the identification of personalized treatment approaches for pancreatic cancer patients.

Furthermore, the thorough data analysis conducted on the measured tumor and blood markers used in this study showed that the measured markers and groups based on similar functions alone, are insufficient to predict a patient's response to chemotherapy. It is necessary to consider all variables together, especially CA19-9, HB, TB and GGT. Of these, CA19-9 measured before and after the first cycle emerged as the most important variables, aligning with findings from other studies [24] [35] [36]. Nevertheless, approximately 20% of pancreatic cancer patients exhibit no or low secretion of CA19-9 due to the Lewis (-) blood group phenotype. Consequently, the developed model cannot be directly applied to patients with the Lewis (-) phenotype. CEA and CA125 have shown potential as alternative biomarkers for Lewis (-) PDAC patients, as these markers have been associated with tumor metastasis and therapeutic response [33]. Future research endeavors should focus on collecting more clinical data specifically for Lewis (-) patients and incorporating CEA and CA125 as biomarkers, enabling the construction of a separate model tailored to this specific patient population or adjusting the proposed model, as the proposed models in this study are not based on Lewis (-) patients.

Nevertheless, the present study also has certain limitations. Firstly, the uneven distribution of patients causes an imbalance in the dataset. This imbalance was addressed by including the entire PD group in the training dataset, which may have contributed to the high model performance. However, other techniques to address dataset imbalance yielded little results. Nonetheless, not using complete class inclusion of the minority class, regardless of whether over- or undersampling or any other method to deal with imbalanced datasets was used, resulted in the model consistently classifying patients as DC, corresponding to the current practice. Therefore, it is suggested to conduct measurements of all the tumor and blood variables used in this study before starting chemotherapy and running the optimal random forest model based on only pre-chemotherapy values. Then based on the output, consider starting chemotherapy using FOLFIRINOX or not. Once started, keep tracking the patient's tumor and blood markers and use the optimized random forest model using both before and after the first cycle values to reconsider the continuation of treatment or not after the first cycle. To gain a more comprehensive understanding of the relationships between important variables, future studies are encouraged to explore advanced techniques that incorporate both marginal and interaction effects in PDP and ALE plots to gain further insights into their (combined) impact on a patient's response to chemotherapy.

Moreover, outliers were initially identified based on the defined criteria of glomerular filtration rate (GFR) $< 30mL/min$ and bilirubin (BR) $> 50\mu mol/L$, as advised by medical experts at the Erasmus MC Rotterdam. However, further analysis revealed the presence of additional potential outliers, indicating the importance of establishing consensus among medical experts to define outlier classification in PDAC patients. Additionally, PCA did not reveal distinct patterns or structures within the dataset. However, this might be because PCA captures linear relationships. The generated PDP and ALE plots clearly show that many variables exhibit non-linear relationships, explaining why PCA failed to distinguish between the two response groups. Non-linear dimensionality reduction techniques should be considered in further analyses. Furthermore, it is important to note that individual patient responses to chemotherapy within the DC group can vary. While the DC category generally indicates favorable outcomes, it encompasses a heterogeneous population with diverse treatment responses, including CR, PR and SD. Variations in certain variables were observed among patients with SD, emphasizing the need to consider the nuanced nature of treatment responses within the DC group. Consequently, the classification of DC should be interpreted cautiously, considering the potential variations in individual treatment responses. This recognition underscores the importance of personalized and patient-specific considerations in the clinical decision-making process for chemotherapy management. Additionally, future studies are encouraged to use a more comprehensive dataset based on a larger-scale multi-center prospective study, preferably characterized by a balanced distribution among the two classes, to reevaluate the model's classification performance prior to chemotherapy and after one cycle and further refine it based on this expanded dataset.

Finally, while the random forest models demonstrated strong predictive performance, alternative modeling approaches could be explored to validate further and compare the results. For instance, support vector machines or other gradient boosting models like XGBoost or neural networks could be implemented to assess their ability to classify PDAC patients based on measured tumor and blood markers. Especially the more advanced models like XGboost or neural networks might be able to capture more complex relationships in the data that have not been identified by the random forest. Also, regularization techniques like L_1 or L_2 regularization can help control model complexity and prevent overfitting. Additionally, hyperparameters can be optimized using for instance Bayesian optimization to achieve potentially even better performance. However, more complex models need not to be better, but it is worth trying and comparing the various models performance to come to an ultimate model that can be used in practice.

7 | Conclusion

Pancreatic ductal adenocarcinoma (PDAC) poses significant challenges due to late-stage diagnosis and limited treatment options with low response rates. Accurate patient stratification is crucial for enhancing patient quality of life, reducing health risks, and minimizing healthcare costs. Despite the increasing use of chemotherapy, a robust classification model based on predictive blood and tumor markers for FOLFIRINOX chemotherapy response in PDAC is currently lacking. From the extensive data analysis conducted, it was evident that individual markers and functional groupings alone were insufficient to distinguish between chemotherapy responses. The two main proposed random forest models are based on (a) pre-chemotherapy and (b) before and after one chemotherapy cycle variables. They exhibit outstanding performances with high accuracy, f1-score, and balanced accuracy scores of (a) 0.95, 0.97 and 0.97 (b) 0.94, 0.98 and 0.98, respectively. Key variables identified included CA19-9, Hemoglobin, Thrombocytes, and γ-Glutamyl Transferase. Nonetheless, validation with a separate cohort is essential to ensure accuracy and reliability, considering the imbalanced and small dataset used for testing. For clinical practice, the recommendation is to use the pre-chemotherapy random forest model for initial prediction of chemotherapy response. Based on the outcome, decisions can be made regarding chemotherapy initiation or alternative options. Measuring relevant tumor and blood markers before and after the first chemotherapy session can further aid in decision-making using the other random forest model. This model can further help to predict treatment outcomes based on one cycle. However, patient-specific factors, individual responses, and potential side effects should also be considered in the decision-making process, promoting a patient-centered and personalized approach to optimize chemotherapy management. By adopting such an approach, not only can costs be saved, but also unnecessary patient suffering can be minimized. Finally, further clinical trials are necessary, involving more comprehensive data collection and analysis to identify potential new or better predictive markers, to improve and validate upon the proposed models.

Acknowledgments: the author thanks Casper van Eijck and Gaby Strijk for the help during the thesis from a medical perspective as well as Geurt Jongbloed from a statistical viewpoint.

Bibliography

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[2] W. Tolsma, M. van Vliet, O. Busch, J. W. de Groot, B. G. Koerkamp, I. de Hingh, J. van Hooft, F. Kohler, and H. Wilmink, "Alvleesklierkanker in nederland," *IKNL*, november 2021.

[3] Q. P. Janssen, E. M. O'Reilly, C. H. Van Eijck, and B. Groot Koerkamp, "Neoadjuvant treatment in patients with resectable and borderline resectable pancreatic cancer," *Frontiers in oncology*, vol. 10, p. 41, 2020.

[4] S. Thibodeau and I. A. Voutsadakis, "Folfirinox chemotherapy in metastatic pancreatic cancer: a systematic review and meta-analysis of retrospective and phase ii studies," *Journal of clinical medicine*, vol. 7, no. 1, p. 7, 2018.

[5] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney *et al.*, "New response evaluation criteria in solid tumours: revised recist guideline (version 1.1)," *European journal of cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[6] World Cancer Research Fund International, "Pancreatic cancer statistics," https://www.wcrf.org/cancer-trends/pancreatic-cancer-statistics/, Accessed 2023.

[7] A. E. Latenstein, L. G. van der Geest, B. A. Bonsing, B. G. Koerkamp, N. H. Mohammad, I. H. de Hingh, V. E. de Meijer, I. Q. Molenaar, H. C. van Santvoort, G. van Tienhoven *et al.*, "Nationwide trends in incidence, treatment and survival of pancreatic ductal adenocarcinoma," *European Journal of Cancer*, vol. 125, pp. 83–93, 2020.

[8] L. Rahib, B. D. Smith, R. Aizenberg, A. B. Rosenzweig, J. M. Fleshman, and L. M. Matrisian, "Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the united states," *Cancer research*, vol. 74, no. 11, pp. 2913–2921, 2014.

[9] S. C. Lau and W. Y. Cheung, "Evolving treatment landscape for early and advanced pancreatic cancer," *World journal of gastrointestinal oncology*, vol. 9, no. 7, p. 281, 2017.

[10] "Pancreatic cancer," https://my.clevelandclinic.org/health/diseases/15806-pancreatic-cancer.

[11] F. M. Specialisten, "Richtlijnendatabase - pancreascarcinoom," Online, June 2019, available at https://richtlijnendatabase.nl/richtlijn/pancreascarcinoom/startpagina.html.

[12] Q. Janssen, J. van Dam, B. Bonsing, H. Bos, K. Bosscha, P. Coene, C. van Eijck, I. de Hingh, T. Karsten, M. van der Kolk *et al.*, "Total neoadjuvant folfirinox versus neoadjuvant gemcitabine-based chemoradiotherapy and adjuvant gemcitabine for resectable and borderline resectable pancreatic cancer (preopanc-2 trial): study protocol for a nationwide multicenter randomized controlled trial," *BMC cancer*, vol. 21, pp. 1–8, 2021.

[13] S. Thibodeau and I. A. Voutsadakis, "Folfirinox chemotherapy in metastatic pancreatic cancer: a systematic review and meta-analysis of retrospective and phase ii studies," *Journal of clinical medicine*, vol. 7, no. 1, p. 7, 2018.

[14] F. van der Sijde, J. L. van Dam, B. Groot Koerkamp, B. Haberkorn, M. Y. Homs, D. Mathijssen, M. G. Besselink, J. W. Wilmink, and C. H. van Eijck, "Treatment response and conditional survival in advanced pancreatic cancer patients treated with folfirinox: a multicenter cohort study," *Journal of oncology*, vol. 2022, 2022.

[15] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, and P. P. Carbone, "Toxicity and response criteria of the eastern cooperative oncology group," *American journal of clinical oncology*, vol. 5, no. 6, pp. 649–656, 1982.

[16] T. Conroy, F. Desseigne, M. Ychou, O. Bouché, R. Guimbaud, Y. Bécouarn, A. Adenis, J.-L. Raoul, S. Gourgou-Bourgade, C. de la Fouchardière *et al.*, "Folfirinox versus gemcitabine for metastatic pancreatic cancer," *New England journal of medicine*, vol. 364, no. 19, pp. 1817–1825, 2011.

[17] M. Suker, B. R. Beumer, E. Sadot, L. Marthey, J. E. Faris, E. A. Mellon, B. F. El-Rayes, A. Wang-Gillam, J. Lacy, P. J. Hosein *et al.*, "Folfirinox for locally advanced pancreatic cancer: a systematic review and patient-level meta-analysis," *The Lancet Oncology*, vol. 17, no. 6, pp. 801–810, 2016.

[18] X. Qiu, "Discussion on pancreatic cancer and its targeted drug development," June 2018. [Online]. Available: https://xueqiu.com/6815845163/109657186

[19] F. Safi, R. Roscher, and H. Beger, “Tumor markers in pancreatic cancer. sensitivity and specificity of ca 19-9.” *Hepato-gastroenterology*, vol. 36, no. 6, pp. 419–423, 1989.

[20] G. Luo, C. Liu, M. Guo, J. Long, Z. Liu, Z. Xiao, K. Jin, H. Cheng, Y. Lu, Q. Ni *et al.*, “Ca19-9-low&lewis (+) pancreatic cancer: A unique subtype,” *Cancer Letters*, vol. 385, pp. 46–50, 2017.

[21] C. Haglund, P. J. Roberts, P. Kuusela, T. M. Scheinin, O. Mäkelä, and H. Jalanko, “Evaluation of ca 19-9 as a serum tumour marker in pancreatic cancer,” *British journal of cancer*, vol. 53, no. 2, pp. 197–202, 1986.

[22] A. Azzariti, O. Brunetti, L. Porcelli, G. Graziano, R. M. Iacobazzi, M. Signorile, A. Scarpa, V. Lorusso, and N. Silvestris, “Potential predictive role of chemotherapy-induced changes of soluble cd40 ligand in untreated advanced pancreatic ductal adenocarcinoma,” *OncoTargets and therapy*, pp. 4681–4686, 2016.

[23] K. Schlick, T. Magnes, L. Ratzinger, B. Jaud, L. Weiss, T. Melchardt, R. Greil, and A. Egle, “Novel models for prediction of benefit and toxicity with folfirinox treatment of pancreatic cancer using clinically available parameters,” *PLoS One*, vol. 13, no. 11, p. e0206688, 2018.

[24] F. van der Sijde, E. E. Vietsch, D. A. Mustafa, M. G. Besselink, B. Groot Koerkamp, and C. H. van Eijck, “Circulating biomarkers for prediction of objective response to chemotherapy in pancreatic cancer patients,” *Cancers*, vol. 11, no. 1, p. 93, 2019.

[25] I. K. Nederland. [Online]. Available: https://nkr-cijfers.iknl.nl/#/viewer/95a3d55f-5284-49ec-a501-986989730799

[26] K. Strimbu and J. A. Tavel, “What are biomarkers?” *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010.

[27] J. L. Humphris, D. K. Chang, A. L. Johns, C. J. Scarlett, M. Pajic, M. D. Jones, E. K. Colvin, A. Nagrial, V. T. Chin, L. A. Chantrill *et al.*, “The prognostic and predictive value of serum ca19. 9 in pancreatic cancer,” *Annals of oncology*, vol. 23, no. 7, pp. 1713–1722, 2012.

[28] A. C. Berger, M. Garcia Jr, J. P. Hoffman, W. F. Regine, R. A. Abrams, H. Safran, A. Konski, A. B. Benson III, J. MacDonald, and C. G. Willett, “Postresection ca 19-9 predicts overall survival in patients with pancreatic cancer treated with adjuvant chemoradiation: a prospective validation by rtog 9704,” *Journal of clinical oncology*, vol. 26, no. 36, p. 5918, 2008.

[29] F. Safi, H. G. Beger, R. Bittner, M. Büchler, and W. Krautzberger, “Ca19-9 and pancreatic adenocarcinoma,” *Cancer*, vol. 57, no. 4, pp. 779–783, 1986.

[30] M. A. Tempero, E. Uchida, H. Takasaki, D. A. Burnett, Z. Steplewski, and P. M. Pour, “Relationship of carbohydrate antigen 19-9 and lewis antigens in pancreatic cancer,” *Cancer research*, vol. 47, no. 20, pp. 5501–5503, 1987.

[31] P. M. Pour, M. M. Tempero, H. Takasaki, E. Uchida, Y. Takiyama, D. A. Burnett, and Z. Steplewski, “Expression of blood group-related antigens abh, lewis a, lewis b, lewis x, lewis y, and ca 19-9 in pancreatic cancer cells in comparison with the patient’s blood group type,” *Cancer research*, vol. 48, no. 19, pp. 5422–5426, 1988.

[32] G. Luo, C. Liu, M. Guo, H. Cheng, Y. Lu, K. Jin, L. Liu, J. Long, J. Xu, R. Lu *et al.*, “Potential biomarkers in lewis negative patients with pancreatic cancer,” *Annals of surgery*, vol. 265, no. 4, pp. 800–805, 2017.

[33] ——, “Potential biomarkers in lewis negative patients with pancreatic cancer,” *Annals of surgery*, vol. 265, no. 4, pp. 800–805, 2017.

[34] W. Hartwig, O. Strobel, U. Hinz, S. Fritz, T. Hackert, C. Roth, M. W. Büchler, and J. Werner, “Ca19-9 in potentially resectable pancreatic cancer: perspective to adjust surgical and perioperative therapy,” *Annals of surgical oncology*, vol. 20, pp. 2188–2196, 2013.

[35] A. C. Berger, I. M. Meszoely, E. A. Ross, J. C. Watson, and J. P. Hoffman, “Undetectable preoperative levels of serum ca 19-9 correlate with improved survival for patients with resectable pancreatic adenocarcinoma,” *Annals of surgical oncology*, vol. 11, pp. 644–649, 2004.

[36] L. K. Martin, L. Wei, E. Trolli, and T. Bekaii-Saab, “Elevated baseline ca19-9 levels correlate with adverse prognosis in patients with early-or advanced-stage pancreas cancer,” *Medical Oncology*, vol. 29, pp. 3101–3107, 2012.

[37] R. Molina, V. Barak, A. van Dalen, M. J. Duffy, R. Einarsson, M. Gion, H. Goike, R. Lamerz, M. Nap, G. Sölétormos *et al.*, “Tumor markers in breast cancer–european group on tumor markers recommendations,” *Tumor Biology*, vol. 26, no. 6, pp. 281–293, 2005.

[38] M. Grunnet and J. Sorensen, “Carcinoembryonic antigen (cea) as tumor marker in lung cancer,” *Lung cancer*, vol. 76, no. 2, pp. 138–143, 2012.

[39] M. Juweid, R. M. Sharkey, T. Behr, L. C. Swayne, A. D. Rubin, T. Herskovic, D. Hanley, A. Markowitz, R. Dunn, J. Siegel *et al.*, “Improved detection of medullary thyroid cancer with radiolabeled antibodies to carcinoembryonic antigen.” *Journal of clinical oncology*, vol. 14, no. 4, pp. 1209–1217, 1996.

[40] H. Imaoka, N. Mizuno, K. Hara, S. Hijioka, M. Tajika, T. Tanaka, M. Ishihara, Y. Hirayama, N. Hieda, T. Yoshida *et al.*, "Prognostic impact of carcinoembryonic antigen (cea) on patients with metastatic pancreatic cancer: A retrospective cohort study," *Pancreatology*, vol. 16, no. 5, pp. 859–864, 2016.

[41] O. Nazli, A. D. Bozdag, T. Tansug, R. Kir, and E. Kaymak, "The diagnostic importance of cea and ca 19-9 for the early diagnosis of pancreatic carcinoma." *Hepato-gastroenterology*, vol. 47, no. 36, pp. 1750–1752, 2000.

[42] K. Satake, Y.-S. Chung, H. Yokomatsu, B. Nakata, H. Tanaka, T. Sawada, H. Nishiwaki, and K. Umeyama, "A clinical evaluation of various tumor markers for the diagnosis of pancreatic cancer," *International journal of pancreatology*, vol. 7, pp. 25–36, 1990.

[43] A. Otake, D. Tsuji, K. Taku, Y. Kawasaki, M. Yokoi, H. Nakamori, M. Osada, M. Matsumoto, K. Inoue, K. Hirai *et al.*, "Chemotherapy-induced neutropenia as a prognostic factor in patients with metastatic pancreatic cancer treated with gemcitabine," *European Journal of Clinical Pharmacology*, vol. 73, pp. 1033–1039, 2017.

[44] R. Ahmad, M. Awais, N. Kausar, and T. Akram, "White blood cells classification using entropy-controlled deep features optimization," *Diagnostics*, vol. 13, no. 3, p. 352, 2023.

[45] I. Park, S. J. Choi, Y. S. Kim, H. K. Ahn, J. Hong, S. J. Sym, J. Park, E. K. Cho, J. H. Lee, Y. J. Shin *et al.*, "Prognostic factors for risk stratification of patients with recurrent or metastatic pancreatic adenocarcinoma who were treated with gemcitabine-based chemotherapy," *Cancer research and treatment: official journal of Korean Cancer Association*, vol. 48, no. 4, pp. 1264–1273, 2016.

[46] M. Hoshikawa, T. Aoki, H. Matsushita, T. Karasaki, A. Hosoi, K. Odaira, N. Fujieda, Y. Kobayashi, K. Kambara, O. Ohara *et al.*, "Nk cell and ifn signatures are positive prognostic biomarkers for resectable pancreatic cancer," *Biochemical and biophysical research communications*, vol. 495, no. 2, pp. 2058–2065, 2018.

[47] D. E. Sanford, B. A. Belt, R. Z. Panni, A. Mayer, A. D. Deshpande, D. Carpenter, J. B. Mitchem, S. M. Plambeck-Suess, L. A. Worley, B. D. Goetz *et al.*, "Inflammatory monocyte mobilization decreases patient survival in pancreatic cancer: A role for targeting the ccl2/ccr2 axisrole of inflammatory monocytes in pancreatic cancer," *Clinical cancer research*, vol. 19, no. 13, pp. 3404–3415, 2013.

[48] C. Liu, H. Cheng, G. Luo, Y. Lu, K. Jin, M. Guo, Q. Ni, and X. Yu, "Circulating regulatory t cell subsets predict overall survival of patients with unresectable pancreatic cancer," *International journal of oncology*, vol. 51, no. 2, pp. 686–694, 2017.

[49] A. Balmanoukian, X. Ye, J. Herman, D. Laheru, and S. A. Grossman, "The association between treatment-related lymphopenia and survival in newly diagnosed patients with resected adenocarcinoma of the pancreas," *Cancer investigation*, vol. 30, no. 8, pp. 571–576, 2012.

[50] E. Clark, S. Connor, M. Taylor, K. Madhavan, O. Garden, and R. Parks, "Preoperative lymphocyte count as a prognostic factor in resected pancreatic ductal adenocarcinoma," *Hpb*, vol. 9, no. 6, pp. 456–460, 2007.

[51] P. Fogar, C. Sperti, D. Basso, M. C. Sanzari, E. Greco, C. Davoli, F. Navaglia, C.-F. Zambon, C. Pasquali, E. Venza *et al.*, "Decreased total lymphocyte counts in pancreatic cancer: an index of adverse outcome," *Pancreas*, vol. 32, no. 1, pp. 22–28, 2006.

[52] M. H. Aziz, K. Sideras, N. A. Aziz, K. Mauff, R. Haen, D. Roos, L. Saida, M. Suker, E. van der Harst, J. S. Mieog *et al.*, "The systemic-immune-inflammation index independently predicts survival and recurrence in resectable pancreatic cancer and its prognostic value depends on bilirubin levels: a retrospective multicenter cohort study," *Annals of surgery*, vol. 270, no. 1, pp. 139–146, 2019.

[53] P. Murthy, M. S. Zenati, A. I. Al Abbas, C. J. Rieser, N. Bahary, M. T. Lotze, H. J. Zeh, A. H. Zureikat, and B. A. Boone, "Prognostic value of the systemic immune-inflammation index (sii) after neoadjuvant therapy for patients with resected pancreatic cancer," *Annals of surgical oncology*, vol. 27, pp. 898–906, 2020.

[54] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[55] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood, A. Haider, M. Hosseinzadeh, and R. Ali Naqvi, "Machine learning (ml) in medicine: Review, applications, and challenges," *Mathematics*, vol. 9, no. 22, p. 2970, 2021.

[56] N. Lunardon, G. Menardi, and N. Torelli, "Rose: a package for binary imbalanced learning." *R journal*, vol. 6, no. 1, 2014.

[57] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[58] S. Glen, "Roc curve explained in one picture," September 2019. [Online]. Available: https://www.datasciencecentral.com/roc-curve-explained-in-one-picture/

[59] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *biometrics*, pp. 159–174, 1977.

[60] K. K. Nicodemus and J. D. Malley, “Predictor correlation impacts machine learning algorithms: implications for genomic studies,” *Bioinformatics*, vol. 25, no. 15, pp. 1884–1890, 2009.

[61] Y. A. Meng, Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta, “Performance of random forest when snps are in linkage disequilibrium,” *BMC bioinformatics*, vol. 10, no. 1, pp. 1–17, 2009.

[62] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, “The behaviour of random forest permutation-based variable importance measures under predictor correlation,” *BMC bioinformatics*, vol. 11, pp. 1–13, 2010.

[63] W. Lee, Y. Park, J. W. Kwon, E. Jun, K. B. Song, J. H. Lee, D. W. Hwang, C. Yoo, K.-p. Kim, J. H. Jeong *et al.*, “Reduced and normalized carbohydrate antigen 19-9 concentrations after neoadjuvant chemotherapy have comparable prognostic performance in patients with borderline resectable and locally advanced pancreatic cancer,” *Journal of Clinical Medicine*, vol. 9, no. 5, p. 1477, 2020.

[64] W. Liu, Q. Liu, W. Wang, P. Wang, J. Chen, T. Hong, N. Zhang, B. Li, Q. Qu, and X. He, “Differential diagnostic roles of the serum ca19-9, total bilirubin (tbil) and the ratio of ca19-9 to tbil for benign and malignant,” *Journal of Cancer*, vol. 9, no. 10, p. 1804, 2018.

[65] X. Huang, Z. Lu, K. Zhang, G. Wang, B. Cai, P. Wu, J. Yin, Y. Miao, and K. Jiang, “Prognostic impact of the ratio of preoperative ca19-9 to liver enzyme levels in pancreatic cancer patients with jaundice (predictability of combined ca19-9/ast and ca19-9/γ-ggt for jaundiced pdac patients),” *Pancreatology*, vol. 21, no. 6, pp. 1092–1101, 2021.

[66] S. Li, H. Xu, C. Wu, W. Wang, W. Jin, H. Gao, H. Li, S. Zhang, J. Xu, W. Zhang *et al.*, “Prognostic value of γ-glutamyltransferase-to-albumin ratio in patients with pancreatic ductal adenocarcinoma following radical surgery,” *Cancer medicine*, vol. 8, no. 2, pp. 572–584, 2019.

[67] E. Ermiah, M. Eddfair, O. Abdulrahman, M. Elfagieh, A. Jebriel, M. Al-Sharif, M. Assidi, and A. Buhmeida, “Prognostic value of serum cea and ca19-9 levels in pancreatic ductal adenocarcinoma,” *Molecular and Clinical Oncology*, vol. 17, no. 2, pp. 1–10, 2022.

[68] B. U. Wu, R. K. Butler, E. Lustigova, J. M. Lawrence, and W. Chen, “Association of glycated hemoglobin levels with risk of pancreatic cancer,” *JAMA Network Open*, vol. 3, no. 6, pp. e204 945–e204 945, 2020.

[69] K. Knight, S. Wade, and L. Balducci, “Prevalence and outcomes of anemia in cancer: a systematic review of the literature,” *The American journal of medicine*, vol. 116, no. 7, pp. 11–26, 2004.

[70] F. Tas, Y. Eralp, M. Basaran, B. Sakar, S. Alici, A. Argon, G. Bulutlar, H. Camlica, A. Aydiner, and E. Topuz, “Anemia in oncology practice: relation to diseases and their therapies,” *American journal of clinical oncology*, vol. 25, no. 4, pp. 371–379, 2002.

[71] J. J. Caro, M. Salas, A. Ward, and G. Goss, “Anemia as an independent prognostic factor for survival in patients with cancer: a systematic, quantitative review,” *Cancer*, vol. 91, no. 12, pp. 2214–2221, 2001.

[72] D. P. Steensma, “Is anemia of cancer different from chemotherapy-induced anemia?” *Journal of Clinical Oncology*, vol. 26, no. 7, pp. 1022–1024, 2008.

[73] M. Dicato, . L. Plawny, and M. Diederich, “Anemia in cancer,” *Annals of Oncology*, vol. 21, pp. vii167–vii172, 2010.

[74] H. A. Goubran, J. Stakiw, M. Radosevic, and T. Burnouf, “Platelet–cancer interactions,” in *Seminars in thrombosis and hemostasis*, vol. 40, no. 03. Thieme Medical Publishers, 2014, pp. 296–305.

[75] D. Varon and E. Shai, “Role of platelet-derived microparticles in angiogenesis and tumor progression,” *Discovery medicine*, vol. 8, no. 43, pp. 237–241, 2009.

[76] G. Nash, L. Turner, M. Scully, and A. Kakkar, “Platelets and cancer,” *The lancet oncology*, vol. 3, no. 7, pp. 425–430, 2002.

[77] L. M. Coussens and Z. Werb, “Inflammation and cancer,” *Nature*, vol. 420, no. 6917, pp. 860–867, 2002.

[78] S. Reuter, S. C. Gupta, M. M. Chaturvedi, and B. B. Aggarwal, “Oxidative stress, inflammation, and cancer: how are they linked?” *Free radical biology and medicine*, vol. 49, no. 11, pp. 1603–1616, 2010.

[79] J. Whitfield, “Gamma glutamyl transferase,” *Critical reviews in clinical laboratory sciences*, vol. 38, no. 4, pp. 263–355, 2001.

[80] B. Diergaarde, R. Brand, J. Lamb, S. Y. Cheong, K. Stello, M. M. Barmada, E. Feingold, and D. C. Whitcomb, “Pooling-based genome-wide association study implicates gamma-glutamyltransferase 1 (ggt1) gene in pancreatic carcinogenesis,” *Pancreatology*, vol. 10, no. 2-3, pp. 194–200, 2010.

[81] M. H. Rosner and A. C. Dalkin, “Electrolyte disorders associated with cancer,” *Advances in chronic kidney disease*, vol. 21, no. 1, pp. 7–17, 2014.

[82] S. K. Vodnala, R. Eil, R. J. Kishton, M. Sukumar, T. N. Yamamoto, N.-H. Ha, P.-H. Lee, M. Shin, S. J. Patel, Z. Yu *et al.*, “T cell stemness and dysfunction in tumors are triggered by a common mechanism,” *Science*, vol. 363, no. 6434, p. eaau0135, 2019.

[83] N. Iwai, T. Okuda, J. Sakagami, T. Harada, T. Ohara, M. Taniguchi, H. Sakai, K. Oka, T. Hara, T. Tsuji *et al.*, “Neutrophil to lymphocyte ratio predicts prognosis in unresectable pancreatic cancer,” *Scientific reports*, vol. 10, no. 1, p. 18758, 2020.

[84] J.-J. Yang, Z.-G. Hu, W.-X. Shi, T. Deng, S.-Q. He, and S.-G. Yuan, “Prognostic significance of neutrophil to lymphocyte ratio in pancreatic cancer: a meta-analysis,” *World journal of gastroenterology: WJG*, vol. 21, no. 9, p. 2807, 2015.

[85] Y.-Q. Shi, J. Yang, P. Du, T. Xu, X.-H. Zhuang, J.-Q. Shen, and C.-F. Xu, “Effect of body mass index on overall survival of pancreatic cancer: a meta-analysis,” *Medicine*, vol. 95, no. 14, 2016.

[86] Q.-L. Jiang, C.-F. Wang, Y.-T. Tian, H. Huang, S.-S. Zhang, D.-B. Zhao, J. Ma, W. Yuan, Y.-M. Sun, X. Che *et al.*, “Body mass index does not affect the survival of pancreatic cancer patients,” *World Journal of Gastroenterology*, vol. 23, no. 34, p. 6287, 2017.

[87] Centers for Disease Control and Prevention, “The health effects of overweight and obesity,” Website, September 2022.

[88] World Health Organization, “Ageing and health,” Website, October 2022, retrieved June 16, 2023, from https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[89] “Pancreascarcinoom: neo-/adjuvante chemotherapie, radiotherapie of chemoradiotherapie,” https://richtlijnendatabase.nl/richtlijn/pancreascarcinoom/neo-_adjuvante_chemotherapie_radiotherapie_of_chemoradiotherapie.html June 2019.

[90] “Palliatieve zorg bij pancreascarcinoom,” https://richtlijnendatabase.nl/richtlijn/pancreascarcinoom/palliatieve_zorg_bij_pancreascarcinoom.html, June 2019.

[91] N. Meti, K. Saednia, A. Lagree, S. Tabbarah, M. Mohebpour, A. Kiss, F.-I. Lu, E. Slodkowska, S. Gandhi, K. J. Jerzak *et al.*, “Machine learning frameworks to predict neoadjuvant chemotherapy response in breast cancer using clinical and pathological features,” *JCO Clinical Cancer Informatics*, vol. 5, pp. 66–80, 2021.

[92] L. Guo, W. Wang, X. Xie, S. Wang, and Y. Zhang, “Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer,” *Biomedicine & Pharmacotherapy*, vol. 159, p. 114256, 2023.

[93] H. Nasief, C. Zheng, D. Schott, W. Hall, S. Tsai, B. Erickson, and X. Allen Li, “A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer,” *NPJ precision oncology*, vol. 3, no. 1, p. 25, 2019.

[94] C. Attard, S. Brown, K. Alloul, and M. Moore, “Cost-effectiveness of folfirinox for first-line treatment of metastatic pancreatic cancer,” *Current Oncology*, vol. 21, no. 1, p. e41, 2014.

[95] Current date accessed is 29-7-2023. [Online]. Available: https://www.wisselkoers.nl/dollar-euro

[96] F. van der Sijde, Z. Azmani, M. G. Besselink, B. A. Bonsing, J. W. B. de Groot, B. Groot Koerkamp, B. C. Haberkorn, M. Y. Homs, W. F. van IJcken, Q. P. Janssen *et al.*, “Circulating tp53 mutations are associated with early tumor progression and poor survival in pancreatic cancer patients treated with folfirinox,” *Therapeutic Advances in Medical Oncology*, vol. 13, p. 17588359211033704, 2021.

[97] R. V. Hogg and A. T. Craig, “Introduction to mathematical statistics.(5"" edition),” *Englewood Hills, New Jersey*, 1995.

[98] “Fisher’s exact test,” Statology.org. [Online]. Available: https://www.statology.org/fishers-exact-test/

[99] “Mann-whitney u test,” DataTab.net, accessed: June 23, 2023. [Online]. Available: https://datatab.net/tutorial/mann-whitney-u-test

[100] D. Li, “Diabetes and pancreatic cancer,” *Molecular carcinogenesis*, vol. 51, no. 1, pp. 64–74, 2012.

[101] M. Clinic, “Pancreatitis: Symptoms & causes,” september 2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pancreatitis/symptoms-causes/syc-20360227

[102] C. Clinic, “G-csf (granulocyte-colony stimulating factor),” https://my.clevelandclinic.org/health/treatments/24126-g-csf-treatment, August 2022.

[103] C. Haglund, P. J. Roberts, P. Kuusela, T. M. Scheinin, O. Mäkelä, and H. Jalanko, “Evaluation of ca 19-9 as a serum tumour marker in pancreatic cancer,” *British journal of cancer*, vol. 53, no. 2, pp. 197–202, 1986.

[104] C. T. C. of America, “Cea test: Diagnosing cancer,” November 2021. [Online]. Available: https://www.cancercenter.com/diagnosing-cancer/lab-tests/cea-test#:~:text=The%20normal%20range%20for%20CEA,the%20cancer%20may%20be%20spreading.

[105] B. Y. merritt, “Hemoglobin concentration (hb),” November 2019. [Online]. Available: https://emedicine.medscape.com/article/2085614-overview#:~:text=The%20reference%20ranges%20for%20hemoglobin,mmol%2FL%20(SI%20units)

[106] N. K. Foundation, “Gfr,” 2023. [Online]. Available: https://www.kidney.org/atoz/content/gfr

[107] Y. Matsuda, M. Tsuchishima, Y. Ueshima, S. Takase, and A. Takada, “The relationship between the development of alcoholic liver and pancreatic diseases and the induction of gamma glutamyl transferase,” *Alcohol and Alcoholism*, vol. 28, no. Supplement_1B, pp. 27–33, 1993.

[108] P. Forget, C. Khalifa, J.-P. Defour, D. Latinne, M.-C. Van Pel, and M. De Kock, “What is the normal value of the neutrophil-to-lymphocyte ratio?” *BMC research notes*, vol. 10, no. 1, pp. 1–4, 2017.

[109] C. Oldenhuis, S. Oosting, J. Gietema, and E. De Vries, “Prognostic versus predictive value of biomarkers in oncology,” *European journal of cancer*, vol. 44, no. 7, pp. 946–953, 2008.

[110] Diagram Research, “Phases of clinical trials,” https://diagramresearch.com/phases-of-clinical-trials/#:~:text=Phase%20III%20clinical%20trials%20compare,a%20large%20group%20of%20subjects., Accessed 2023.

[111] “Granulocytes,” Cleveland Clinic, Accessed: June 14, 2023, retrieved from https://my.clevelandclinic.org/health/body/22016-granulocytes.

[112] “Neutrophils,” Cleveland Clinic, Accessed: June 14, 2023, retrieved from https://my.clevelandclinic.org/health/body/22313-neutrophils.

[113] “Lymphocytes,” Cleveland Clinic, Accessed: June 14, 2023, retrieved from https://my.clevelandclinic.org/health/body/23342-lymphocytes.

[114] “Monocytes,” Cleveland Clinic, Accessed: June 14, 2023, retrieved from https://my.clevelandclinic.org/health/body/22110-monocytes.

[115] R. Lugano, M. Ramachandran, and A. Dimberg, “Tumor angiogenesis: causes, consequences, challenges and opportunities,” *Cellular and Molecular Life Sciences*, vol. 77, pp. 1745–1770, 2020.

[116] K. P. Peterson, J. G. Pavlovich, D. Goldstein, R. Little, J. England, and C. M. Peterson, “What is hemoglobin a1c? an analysis of glycated hemoglobins by electrospray ionization mass spectrometry,” *Clinical Chemistry*, vol. 44, no. 9, pp. 1951–1958, 1998.

A | P-values

A.1. P-values

In order to calculate the p-values between the two groups of patients, namely disease control and progressive disease, different statistical tests have been used. For categorical variables the chi-squared and fisher's exact test is used and for continuous variables the Wilcoxon rank test is used. Standard we test the following two hypotheses, all with a significance level $\alpha = 0.05$ taken:

- H_0 = there is no difference between the disease control and progressive disease groups based on the considered categorical or numerical variable
- H_1 = there is a difference between the disease control and progressive disease groups based on the considered categorical or numerical variable

A.1.1 Chi-squared Test

Chi-squared tests are used to test whether there is a difference between groups of unpaired data based on categorical data. There are two types of tests: Chi-squared test of independence and Chi-squared goodness of fit test. We will be focusing on the chi-squared test of independence as we want to compare the two final response groups: disease control and progressive disease with each other.

Assumptions of the chi-square test:

- Independence: the observations used in the test must be independent of each other.
- Sample size: the chi-squared test assumes that the sample size is sufficiently large. There is no fixed rule for determining the sample size, but a general guideline is that the expected frequency count for each cell in the contingency table should at least be ≥ 5. Otherwise alternative tests (like the Fisher's exact test) should be used.

Limitations of the chi-square test:

- Categorical variables: the chi-squared test can only be applied to categorical variables and it is not suitable for continuous or ordinal variables.
- Cell frequencies: the chi-squared test assumes that the expected frequency count for each cell in the contingency table is at least 5. If this condition is violated the test may become unreliable and alternative tests, such as the Fisher's exact test, should be considered.
- Does not indicate the strength of an association, e.g. $p < 0.001$ does not indicate a stronger association than $p < 0.05$.

In order to explain how the Chi-squared test is performed a corresponding p-value is calculated, consider a contingency table with r rows and c columns:

	Column 1	Column 2	...	Column c
Row 1	O_{11}	O_{12}	...	O_{1c}
Row 2	O_{21}	O_{22}	...	O_{2c}
⋮	⋮	⋮	⋱	⋮
Row r	O_{r1}	O_{r2}	...	O_{rc}

where O_{ij} represents the observed count in the cell at the intersection of Row i and Column j. The null hypothesis H_0 of the Chi-Squared test states that there is no association between the row and column variables, and the alternative hypothesis suggests that there is a significant association. To calculate the test statistic, we first calculate the expected frequencies (E_{ij}) assuming the null hypothesis is true. The expected frequency for each cell is given by:

$$E_{ij} = \frac{(\text{row total for Row } i) \times (\text{column total for Column } j)}{\text{total sample size}}$$

The test statistic is then calculated as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Under the null hypothesis, the test statistic approximately follows a chi-squared distribution with degrees of freedom equal to $(r-1)\times(c-1)$. Finally, we compare the obtained test statistic to the critical value from the chi-squared distribution with the desired significance level α. If the obtained test statistic is greater than the critical value, we reject the null hypothesis and conclude that there is evidence of an association between the variables. The chi-squared distribution table or statistical software can be used to find the p-value associated with the calculated chi-squared test statistic and degrees of freedom. Then compare the p-value to the predetermined significance level α to either reject the null hypothesis if $p < \alpha$ or not reject if $p > \alpha$ [97].

A.1.2 Fisher’s Exact Test

Similar to the Chi-squared test, the Fisher’s exact test is used to compare two groups of unpaired categorical data and does not assume any underlying distribution. Similar assumptions as for the chi-squared test are valid. However, the Fisher’s exact test is better when dealing with small sample sizes, even though it is valid for all sample sizes. To explain how a p-value is calculated using the Fisher’s exact test, consider the following 2×2 contingency table.

$$\begin{array}{cc|c} a & b & a+b \\ c & d & c+d \\ \hline a+c & b+d & n \end{array}$$

Here, a, b, c, and d represent the observed frequencies and n the total number of observations. To calculate the p-value for Fisher’s exact test, we use the hypergeometric distribution. The p-value represents the probability of obtaining a table as extreme as or more extreme than the observed table, assuming the null hypothesis is true. In the context of statistical hypothesis testing, “as extreme as" refers to outcomes that are at least as unlikely or as different from the null hypothesis as the observed data. It is used to assess the significance of the observed results. When conducting a statistical test, we compare the observed data to what would be expected under the null hypothesis, which assumes no association or effect. If the observed data is significantly different from what would be expected under the null hypothesis, we consider it to be “extreme." The p-value represents the probability of observing data as extreme as, or more extreme than, the observed data, assuming the null hypothesis is true. For example, in the context of Fisher’s exact test, if the observed contingency table shows a strong association between categorical variables, the p-value quantifies the probability of obtaining a contingency table with a similar or stronger association (as extreme or more extreme) purely by chance, assuming the null hypothesis of no association is true.

$$P(\text{table}) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

To determine the p-value, we calculate the probabilities of all tables as or more extreme than the observed table by considering all possible rearrangements of the counts while keeping the row and column totals fixed. The p-value is then obtained as the sum of these probabilities. Finally, we compare the p-value to a chosen significance level (e.g., 0.05) to make a decision. If the p-value is less than the significance level, we reject the null hypothesis and conclude that there is evidence of an association between the categorical variables [98].

A.1.2.1 Calculation example

Suppose we have two groups, Group 1 and Group 2, and we want to investigate if there is an association between their smoking habits (smoker/non-smoker) and the occurrence of a respiratory disease (present/absent). We collect data from a sample of 50 individuals and obtain the following contingency table:

	Group 1	Group 2
Smoker	10	5
Non-Smoker	15	20

Here, $a = 10$, $b = 5$, $c = 15$, and $d = 20$ represent the observed frequencies.

To calculate the p-value using Fisher's exact test, we use the formula:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

where $\binom{n}{k}$ represents the binomial coefficient. Here, $n = a + b + c + d = 50$.

Substituting the observed frequencies into the formula:

$$p = \frac{\binom{10+5}{10}\binom{15+20}{15}}{\binom{50}{10+15}}$$

Simplifying further:

$$p = \frac{\binom{15}{10}\binom{35}{15}}{\binom{50}{25}}$$

Evaluating the binomial coefficients:

$$p \approx 0.078$$

Therefore, the calculated p-value using Fisher's exact test is approximately 0.078 and we would not reject the null hypothesis if a significance level of $\alpha = 0.05$ is taken and conclude that there is no significance association between smoking habits and the occurrence of the respiratory disease.

A.1.3 Wilcoxon Rank Sum Test / Mann-Whitney U test

The Wilcoxon Rank Sum test, also known as Mann-Whitney U test, is a nonparametric statistical test used to compare two groups and analyzes if they are statistically significant from each other. It is a nonparametric equivalent of the unpaired t-test in which the data does not follow a normal distribution. It tests whether the two samples come from the same population (i.e. have the same median) or alternatively differ from each other [99]. Assumptions of the Wilcoxon Rank test are that the sample drawn from the population is random and that these samples are independent.

Consider two independent samples, Sample 1 and Sample 2, with sizes n_1 and n_2, respectively. The null hypothesis H_0 of the Wilcoxon rank sum test states that there is no difference between the distributions of the two samples, while the alternative hypothesis H_1 suggests that there is a significant difference. Then to determine the critical value or p-value using the Wilcoxon ranksum test, do the following:

1. Combine the data from both samples and assign ranks to the combined data. If there are ties, assign the average rank to the tied values.
2. Calculate the sum of ranks for each sample:
   - S_1: Sum of ranks for Sample 1
   - S_2: Sum of ranks for Sample 2
3. Calculate the test statistic, U:
   - If $n_1 \leq n_2$, $U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - S_1$
   - If $n_1 > n_2$, $U = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - S_2$

4. Determine the critical value or p-value:
    - The critical value or p-value can be obtained from the standard normal distribution or appropriate tables.
    - If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that there is a significant difference between the distributions of the two samples.

A.1.3.1 Example Calculation

Suppose we have two independent groups, Group 1 and Group 2, with the following observed values:

Group 1: 6, 9, 7, 5, 8
Group 2: 4, 3, 2, 1

Let n_1 and n_2 represent the sample sizes of Group 1 and Group 2, respectively. In this case, $n_1 = 5$ and $n_2 = 4$. To calculate the p-value using the Wilcoxon rank sum test, we do the following steps:

1. Combine the data from both groups and rank the values. In case of ties, assign the average rank.
   Combined data: 1, 2, 3, 4, 5, 6, 7, 8, 9
   Ranking: 1, 2, 3, 4, 5.5, 5.5, 7, 8, 9
2. Sum up the ranks for each group.
   Sum of ranks for Group 1: $R_1 = 1 + 2 + 3 + 4 + 5.5 = 15.5$
   Sum of ranks for Group 2: $R_2 = 5.5 + 7 + 8 + 9 = 29.5$
3. Calculate the U statistic for each group.
   $U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - R_1$
   $U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - R_2$
   Substituting the values:
   $U_1 = 5 \cdot 4 + \frac{5 \cdot (5+1)}{2} - 15.5 = 10$
   $U_2 = 5 \cdot 4 + \frac{4 \cdot (4+1)}{2} - 29.5 = 5$
4. Determine the minimum U statistic as $U = \min(U_1, U_2)$.
   In this case, $U = \min(10, 5) = 5$.
5. Calculate the p-value using the appropriate distribution. For small sample sizes, we refer to critical values or tables. For larger sample sizes, we can approximate the p-value using the normal distribution.
6. Interpret the result:
    - If the p-value is less than the predetermined significance level (e.g., 0.05), we reject the null hypothesis and conclude that there is a statistically significant difference between the groups.
    - If the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis, indicating insufficient evidence to establish a significant difference between the groups.

In this example, the calculated p-value would be determined based on the appropriate distribution or approximation method. By calculating the p-value using the Wilcoxon rank sum test, we can assess the significance of the difference between the two independent groups [97].

B | Exploratory Data Analysis (full)

B.1. Data Analysis

This section presents a comprehensive analysis of the PREOPANC2-iKnowIT data provided by the Erasmus MC, consisting of 249 patients spanning the years 2018-2020, with one person from 2015. Two patients, identified with patient ID 059PP20013 and 059PP20014, have been excluded from the analysis due to their lack of data, with the exception of a Date IC, hence the total dataset comprises of 247 observations. In addition, there were eight patients with missing final response, these will be classified in the NA category in the subsequent data analysis graphs and charts provided.

B.1.1 Analysis of (baseline) Patient characteristics

B.1.1.1 Gender

With respect to gender, the dataset considers male and female, which are denoted by 0 and 1, respectively. Of the 239 subjects, 137 (55%) were male, while 110 (45%) were female. The stacked bar chart indicates that gender does not appear to significantly influence the final response to chemotherapy as no significant differences across the two groups can be seen.

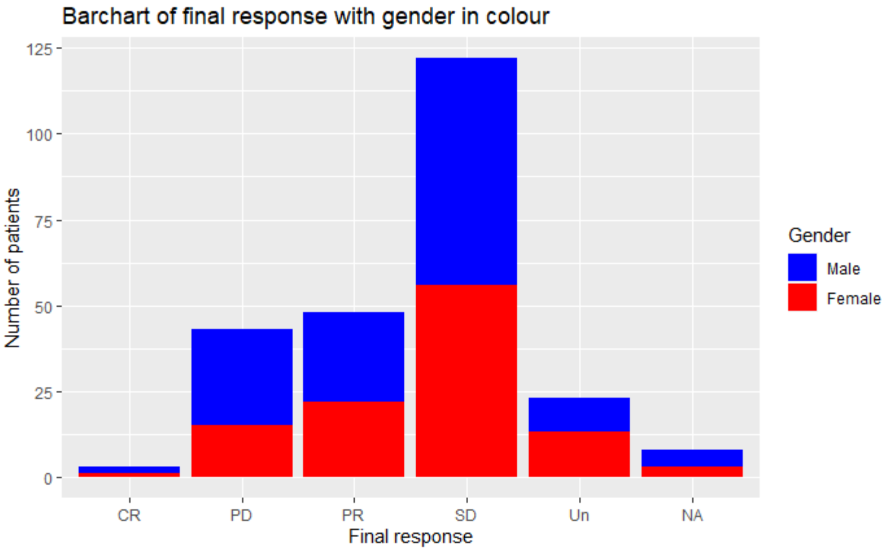


Figure B.1: Stacked bar chart displaying the final outcome categorized by gender, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), total n=247 (137 male, 110 female).

B.1.1.2 Age

The provided dataset lacks explicit information on the age of patients. Therefore, age is calculated by computing the difference between the date of informed consent (DATE IC E1 C1) and the date of birth (DATE BIRTH E1 C2), as given in (B.1):

$$Age = \text{DATE IC E1 C1} - \text{DATE BIRTH E1 C2}. \tag{B.1}$$

This method was deemed most suitable for estimating the patient's age at the start of the research. The resulting age values are in years and not rounded (i.e. patients can have an age of 64.28 years). Thus, a new column named 'Agerounded' has been added to the dataset, where age values are rounded to the nearest integer. The analysis of the Age variable shows a minimum age of 47 years, a mean age of 64 years, and the oldest patient in the dataset being 82 years old. Due to missing values in the DATE BIRTH E1 C2, 27 patients have missing age information. Furthermore, during the analysis, a clear outlier in the dataset was detected, corresponding to patient number 239 with ID 165PP20017. The recorded date of birth, 2963-07-08, was evidently erroneous and has been corrected to 1963-07-08.

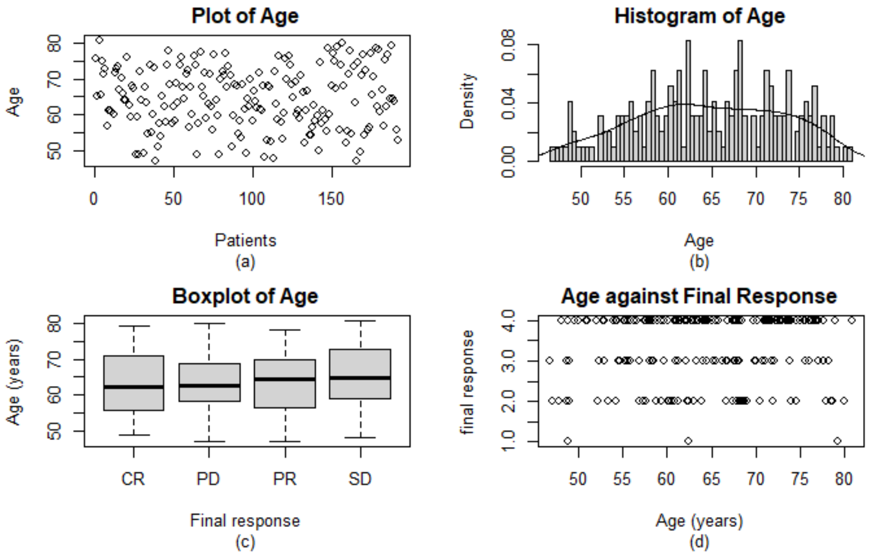


Figure B.2: Age distribution: (a) scatterplot of age data, (b) histogram with a fitted density curve (c) boxplot showing the distribution of ages across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

B.1.1.3 Weight

The weight is provided in kg and range from a minimum of 42kg to 124kg with a mean value of 78kg.

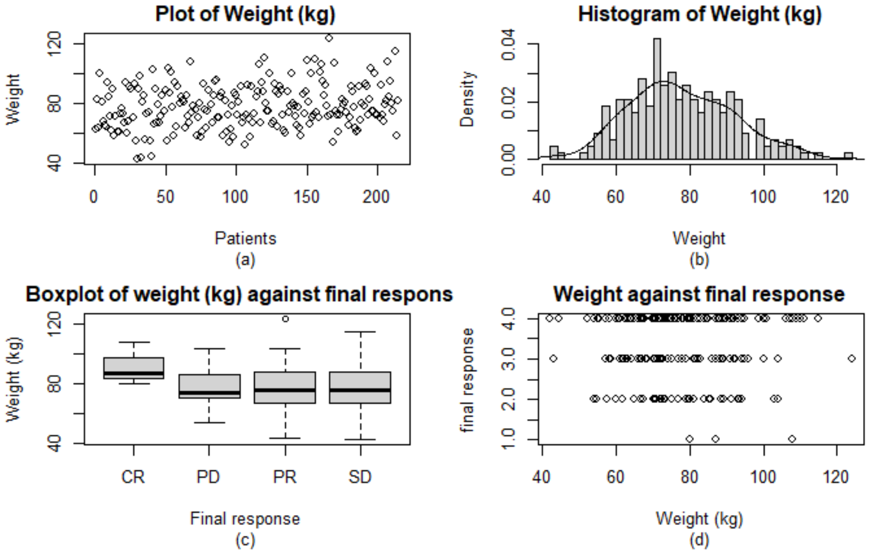


Figure B.3: Weight distribution: (a) scatterplot of weight data, (b) histogram with a fitted density curve (c) boxplot showing the distribution of weights across different final response categories, includingComplete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

B.1.1.4 Length

The data defined a patient's height with the variable 'LENGTH E1 C2' and measured in centimeters. Upon initial inspection, an abnormal value of 1886cm was observed, likely resulting from a typing error, which has been adjusted to the more realistic value of 186cm for patient ID 078PP20038. The analysis revealed a minimum height of 152cm, a maximum of 200cm, and a mean height of 174cm among patients. Additionally, there were 5 instances where height information was missing.

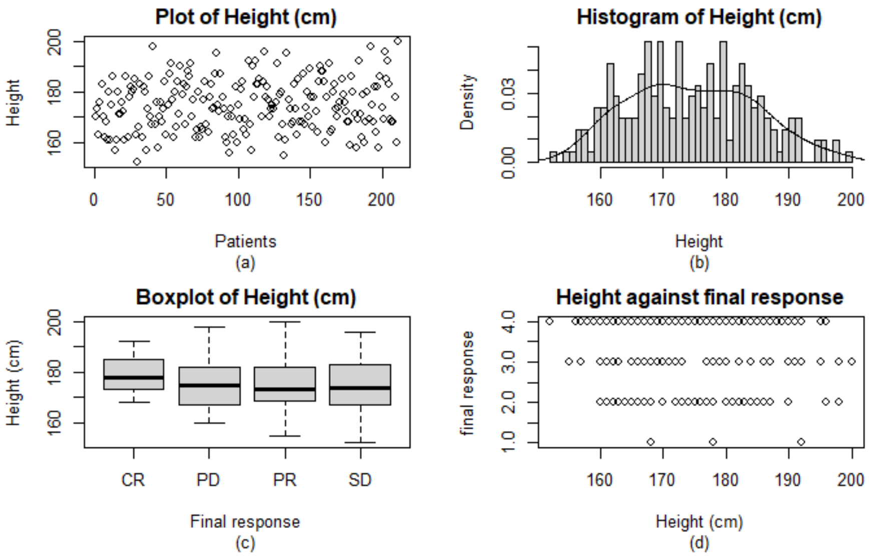


Figure B.4: Height distribution: (a) scatterplot of height data, (b) histogram with a fitted density curve (c) boxplot showing the distribution of heights across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

B.1.1.5 BMI

Solely examining the height and weight of patients does not provide enough information to determine their final response to chemotherapy. Therefore, it is more informative to combine these measurements into a single metric. One of the most commonly employed metrics is the Body Mass Index (BMI), computed using the formula presented in Equation (B.2):

$$BMI = \frac{weight\ (kg)}{height\ (m)^2} \tag{B.2}$$

Based on the available data, BMI values range from a minimum of 15.9 (indicating underweight) to 37.1 (signifying obesity), with a mean BMI of 25.5. This mean BMI value falls just outside the healthy BMI range and is classified as overweight. Thus, the majority of patients are on the heavier side. As a side note, there were 5 instances where BMI information was missing.

B.1.1.6 Stage of Disease

The stage of disease is defined as 0 = (Borderline) resectable, 1 = Locally Advanced, 2 = Metastatic disease. In the dataset provided, we have a total of 152 (62%) patients with (borderline) resectable PDAC, 54 (22%) patients with LAPC and 41 (17%) with metastatic PDAC.

B.1.1.7 Family history of pancreatic cancer

The presence of a positive family history of pancreatic cancer is indicated by whether or not the patient has one or more family members who have been diagnosed with the disease in the past. In the present dataset, 147 patients (60%)

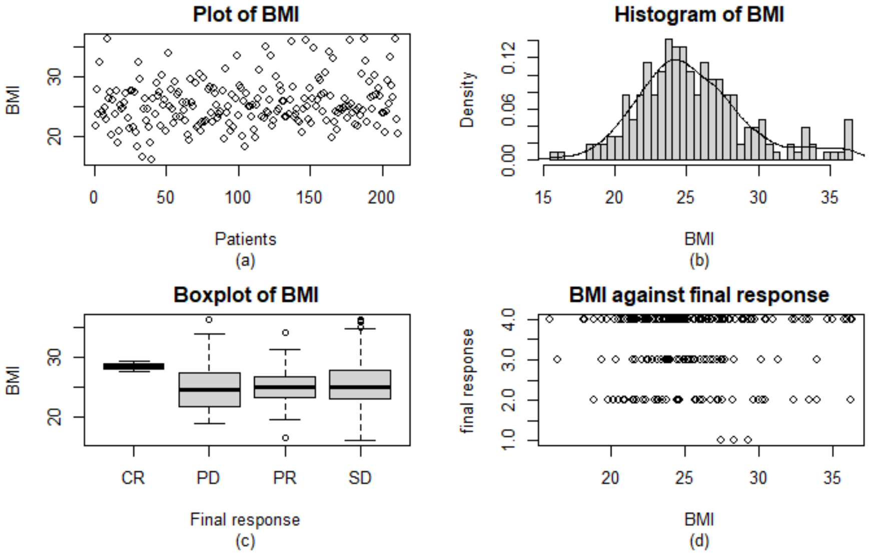


Figure B.5: BMI distribution: (a) scatterplot of BMI data, (b) histogram with a fitted density curve (c) boxplot showing the distribution of BMI across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

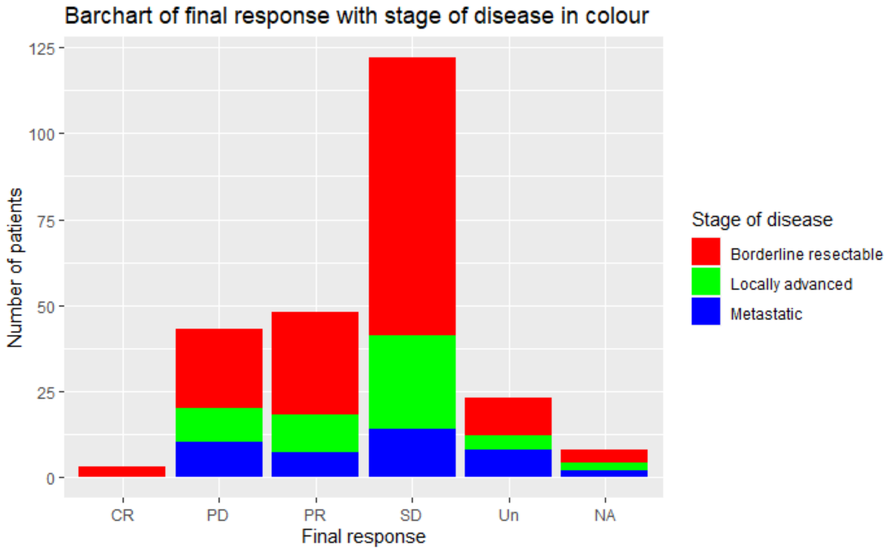


Figure B.6: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by stage of disease: (borderline) resectable (n=152), locally advanced (n=54), metastatic disease (n=41)), total n=247.

had no family history of PDAC, 23 (9%) had a positive family history, and in 77 cases (31%), such information was unknown. Consequently, it remains unclear whether a positive family history of pancreatic cancer exerts any influence on the outcome. In particular, based on the provided graph, it appears that a majority of patients with a positive family history of pancreatic cancer exhibited stable disease. However, the large number of unknown cases makes drawing definitive conclusions challenging.

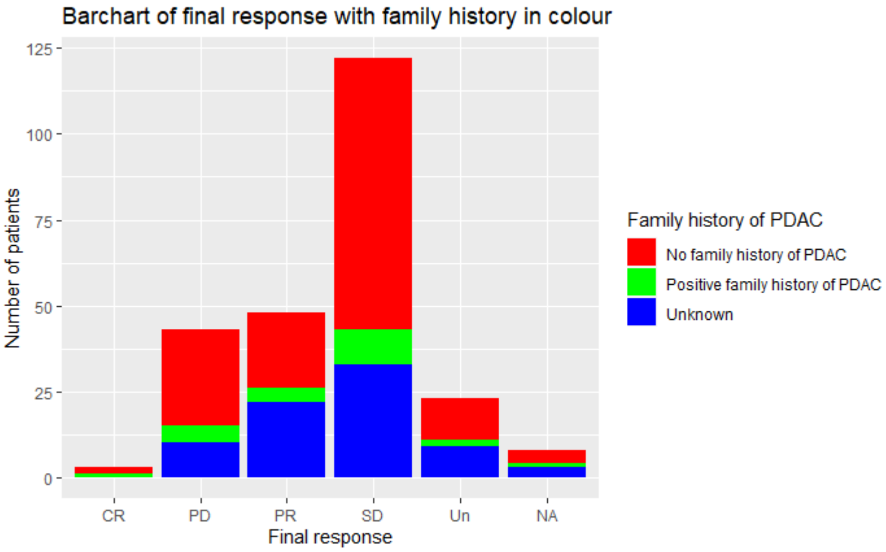


Figure B.7: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by family history of pancreatic cancer of disease: no family history of pancreatic cancer (n=147), positive family history of pancreatic cancer (n=23), unknown (n=77) total n=247.

B.1.1.8 Smoking

Smoking has been identified as one of the most significant risk factors for the development of PDAC. Cigarette smoke contains various carcinogens that can damage the DNA of pancreatic cells and lead to the formation of cancerous tumors. In the provided dataset, it is observed that a considerable proportion of the patients had a history of smoking. Specifically, 124 (50%) of the PDAC patients were current or former smokers, while 108 (44%) had never smoked. The remaining patients had unknown smoking status. After considering the stacked bar chart, it can be hypothesized that a significant proportion of the former and current smokers exhibit stable disease as well as progressive disease as their final response. However, the impact of smoking on the final response of PDAC patients cannot be conclusively determined due to the limited sample size and the high number of patients with unknown smoking status.

B.1.1.9 Alcohol usage

Furthermore, alcohol consumption is often considered as another risk factor for pancreatic cancer. In the provided dataset, the patients' alcohol use is categorized as either no usage (0), usage (1), stopped (2), or unknown (3). The range of alcohol consumption is quite wide, as occasional consumption of a glass of wine is also classified as "usage" compared to habitual heavy drinking. Out of the 247 patients, 78 (32%) reported no alcohol consumption, 113 (46%) reported alcohol consumption, 38 (15%) reported quitting alcohol consumption, and 18 (7%) had unknown alcohol use. After examining the stacked bar chart, it appears that all of the patients who showed a complete response to chemotherapy reported alcohol consumption. However, how much alcohol these patients drank is unknown. Therefore, clear conclusions cannot be drawn.

B.1.1.10 Diabetes Mellitus

According to research by Li [100], there is a suggested relationship between diabetes and pancreatic cancer. The underlying mechanisms contributing to the development of diabetes-associated pancreatic cancer include insulin resistance, hyperglycemia, hyper-insulinemia, and inflammation. In the provided dataset, 189 (77%) patients have no history of diabetes while 58 (23%) have been diagnosed with the condition. This indicates that approximately one quarter of the patients may have developed pancreatic cancer due to diabetes. A visual examination of Figure B.10 shows that the majority of patients

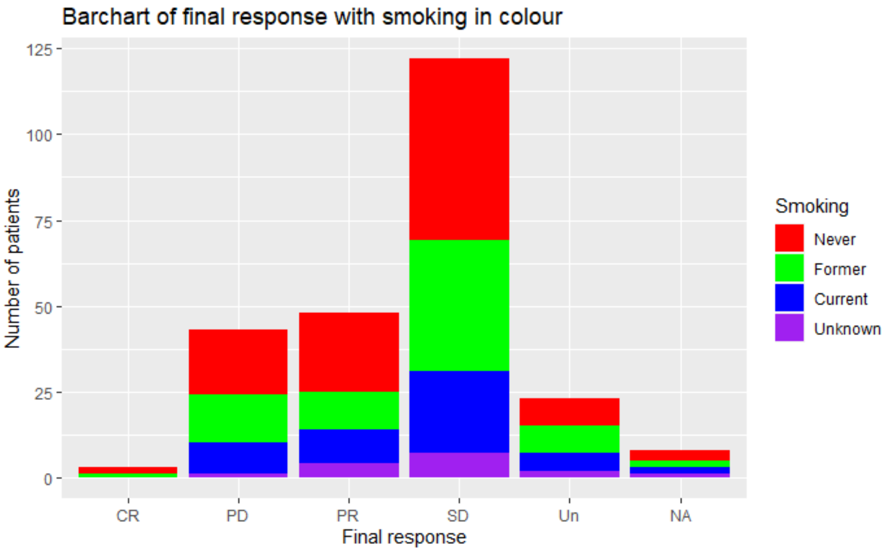


Figure B.8: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by smoking: never (n=108), former (n=74), current (n=50), unknown (n=15), total n=247.

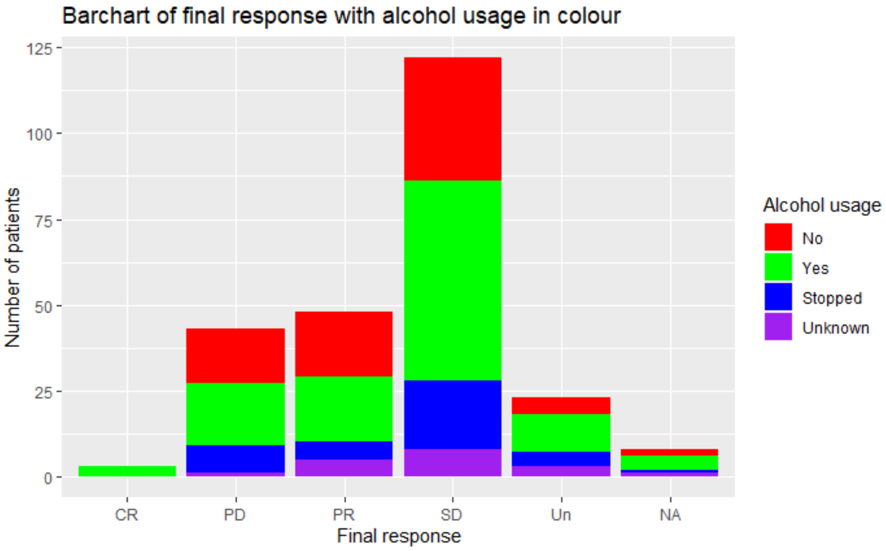


Figure B.9: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by alcohol usage: no (n=78), yes (n=113), stopped (n=38), unknown (n=18), total n=247.

with diabetes reported stable disease. However, a similar trend is observed in the non-diabetic patients, making it difficult to draw firm conclusions about the relationship between diabetes and pancreatic cancer in this dataset.

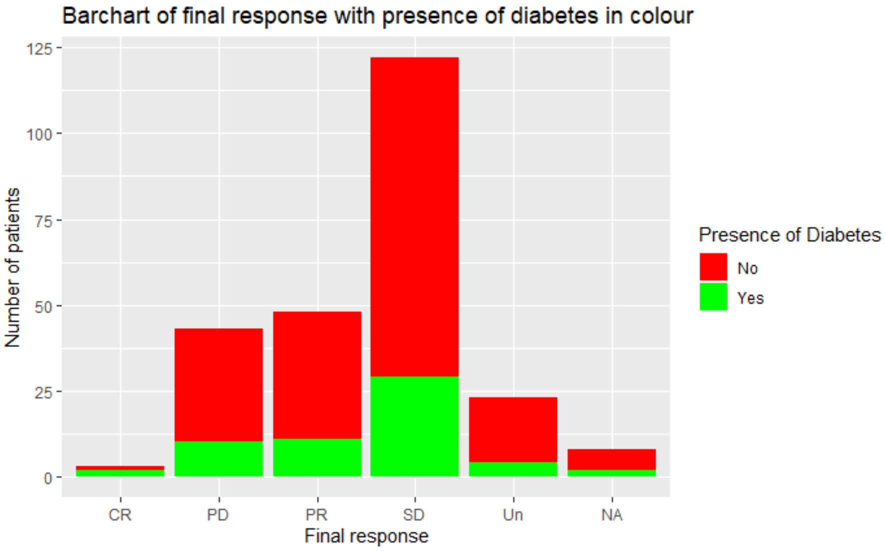


Figure B.10: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by presence of diabetes mellitus: no (n=189), yes (n=58), total n=247.

B.1.1.11 Pancreatitis

Pancreatitis is a condition characterized by inflammation of the pancreas, which may present as acute or chronic pancreatitis. The latter may occur over an extended period and often recur, whereas acute pancreatitis is sudden and lasts for a few days [101]. PDAC is a highly aggressive form of pancreatic cancer that originates from the exocrine cells of the pancreas. These exocrine pancreatic cells produce enzymes that are secreted into the small intestine. In the provided dataset of 239 patients with PDAC, (97%) had no history of pancreatitis, while eight (3%) had a history of pancreatitis. Interestingly, the majority of patients with pancreatitis showed stable disease. Although this suggests that a past history of pancreatitis may not be a significant risk factor for the development of PDAC, it is important to note that the study was limited in scope and may not represent the general population. Therefore, more research is needed to fully understand the relationship between pancreatitis and PDAC.

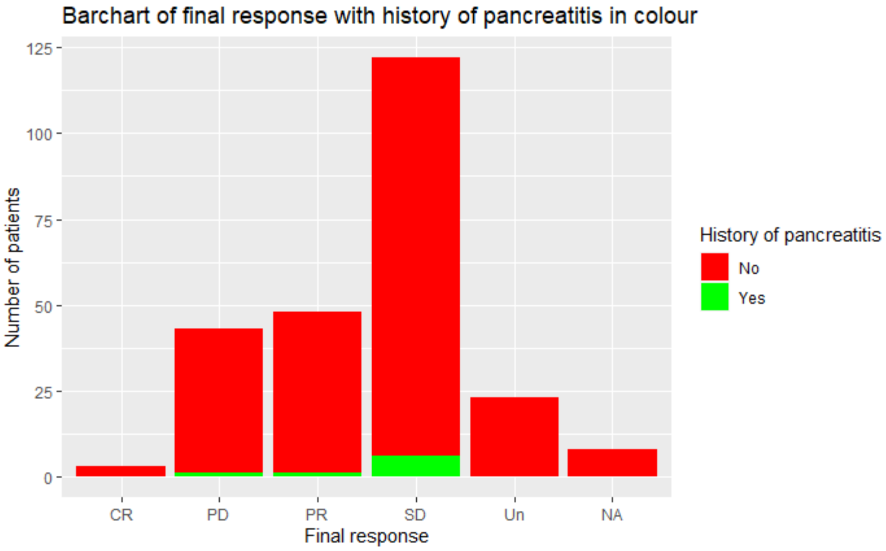


Figure B.11: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by presence of pancreatitis: no (n=239), yes (n=8), total n=247.

B.1.1.12 History of malignancy

History of malignancy refers to the presence of earlier cancer in the patient. It can be any type of cancer. The present study involves a cohort of patients, of which 210 patients had no prior history of malignancy, while 37 had. The findings were analyzed using a stacked bar chart, which shows that patients with a history of malignancy displayed mainly stable disease as their final response. This observation could potentially be attributed to the possibility that patients with a history of malignancy might have developed a heightened sensitivity to treatment as a result of their prior exposure to cancer therapy. However, further research is necessary to establish a causal link between a history of malignancy and improved response to treatment in PDAC patients.

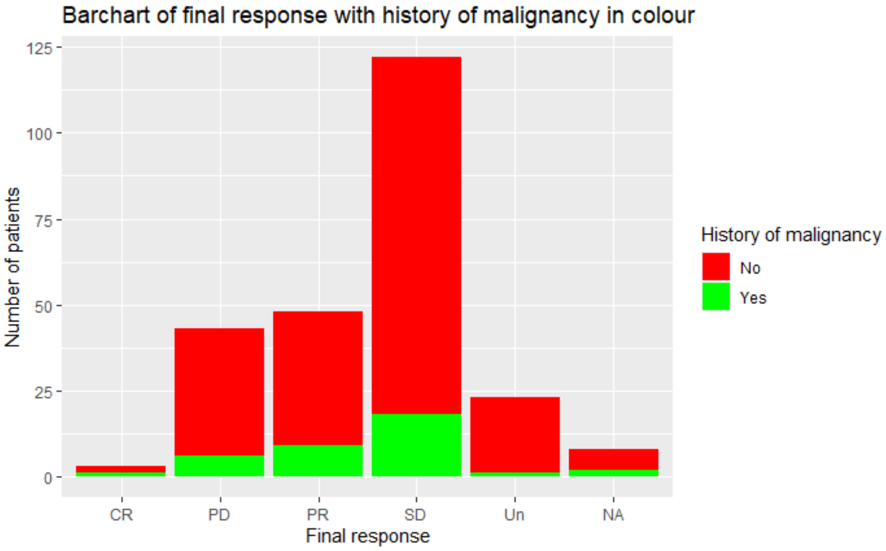


Figure B.12: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by their history of previous malignancy: no (n=210), yes (n=37), total n=247.

B.1.1.13 History of chemotherapy

According to the analysis of the studied subjects, the majority of patients 234 (95%) did not have a prior history of chemotherapy, whereas a minority 13(5%) did. The stacked bar chart in Figure B.13 depicts the distribution of final response to chemotherapy among the patients based on their history of chemotherapy. Due to the sample size of patients with a history of chemotherapy, solid conclusions cannot be drawn.

B.1.1.14 Cycles of FOLFIRINOX

The number of FOLFIRINOX cycles given to each patient in the data set varies per patient. It is worth mentioning that three patients within the dataset did not undergo any chemotherapy treatment. Nevertheless, these patients are excluded from the dataset after the removal of missing values. Among the included patients, the majority (n=128) received eight cycles of FOLFIRINOX, followed by four cycles (n=32). A subset of fifteen patients received the highest number of cycles recorded in the dataset, totaling twelve cycles of FOLFIRINOX, prior to the documentation of the final chemotherapy response. The remaining patients received any number between 0 and 12 cycles of FOLFIRINOX. It is important to note that the blood data measurements captured in this dataset were obtained both before and after the initial cycle, despite subsequent cycles being administered to the majority of patients. Interestingly, all the patients who received 4 cycles of FOLFIRINOX have been classified as PD or SD and patients who received in general more than 4 cycles are classified as PR. The three patients who were classified as CR have received 7 or 8 cycles of FOLFIRINOX.

B.1.1.15 G-CSF received

Granulocyte-Colony Stimulating Factor (G-CSF) is a pharmaceutical agent that facilitates the production of additional neutrophils within the body. Neutrophils are a specific subtype of granulocytes, which are a type of white blood cells. Further information on white blood cells can be found in Section 2.1 and Appendix E. Colony-stimulating factors are

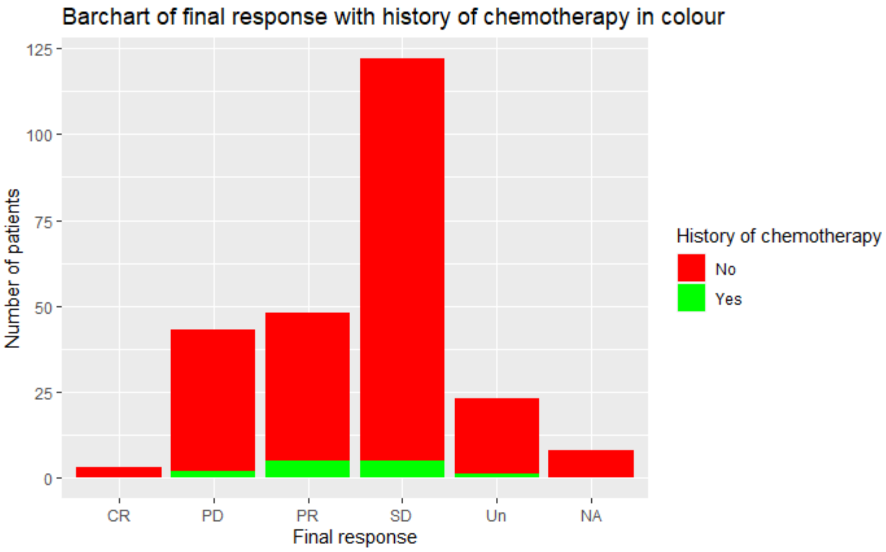


Figure B.13: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by their history of chemotherapy: no (n=234), yes (n=13), total n=247.

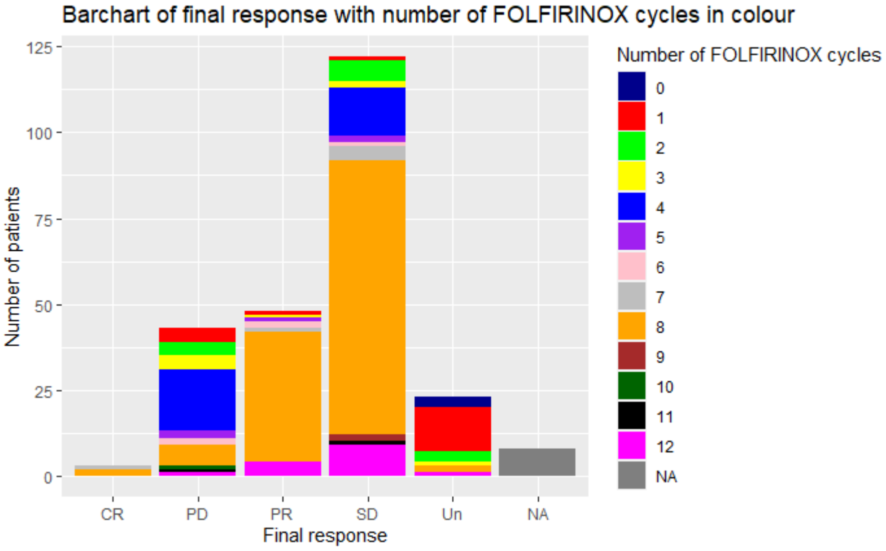


Figure B.14: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by the number of FOLFIRINOX cycles they have received: 0 (n=3), 1 (n=19), 2 (n=13), 3 (n=8), 4 (n=32), 5 (n=5), 6 (n=5), 7 (n=6), 8 (n=128), 9 (n=2), 10 (n=1), 11 (n=2), 12 (n=15), Unknown (n=8), total n=247.

proteins that give instructions to cells what to do. G-CSF has the ability to instruct stem cells to differentiate into neutrophils, leading to an increase in the number of neutrophils present in the body. Specifically, G-CSF is used during the treatment of pancreatic ductal adenocarcinoma to manage and prevent neutropenia. Neutropenia refers to a condition characterized by a lower-than-normal level of neutrophils in the blood, which can arise as a side effect of cancer treatment and hinder the body's ability to combat infections [43]. By stimulating the bone marrow, G-CSF encourages the production of additional neutrophils, thereby reducing the risk of serious infections or fevers following chemotherapy. Additionally, chemotherapy not only targets cancer cells but also damages healthy cells, including neutrophils. If left untreated, this situation can pose a life-threatening risk [102]. From all the n=247 patients 52 have not received G-CSF, while 186 have received it. This could potentially explain the large increasing trend seen in the number of neutrophils in many patients.

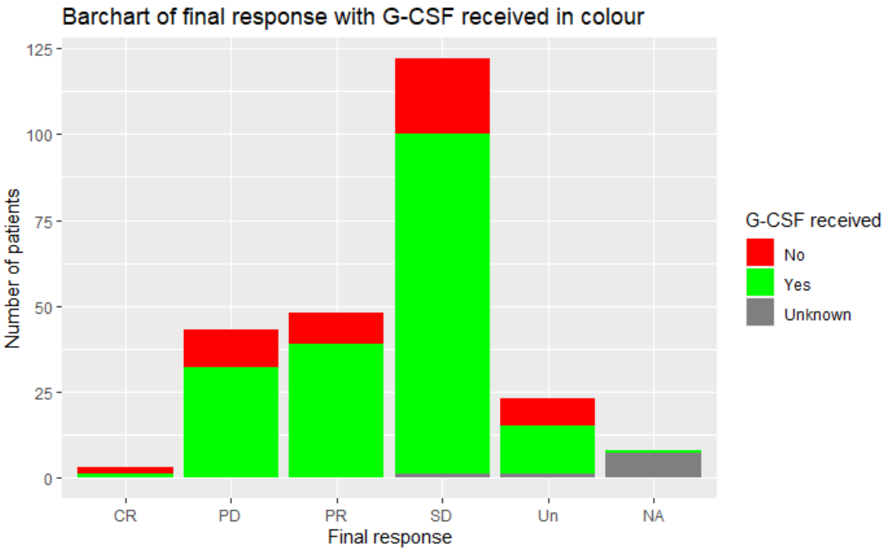


Figure B.15: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by whethery they have received G-CSF: no (n=52), yes (n=186), unknown (n=9), total n=247.

B.1.1.16 G-CSF prophylaxis

G-CSF prophylaxis refers to whether G-CSF has been given to the patient before the start of chemotherapy or when it was observed to be necessary. Patients are categorized into three groups: “no," “yes," and “unknown." The classification of “no" signifies that the patient did not receive G-CSF prior to chemotherapy. On the other hand, “yes" indicates that the patient was given G-CSF prior to the start of chemotherapy. The “unknown" category is assigned to patients for whom there is insufficient information available regarding G-CSF administration. This classification system allows for a distinction to be made based on whether G-CSF was administered as a preventive measure or if it was initiated in response to observed requirements due to declining heatlth conditions. Expanding on this topic, it is important to note that the decision to administer G-CSF prophylaxis is typically influenced by factors such as the patient's individual risk factors, treatment protocol, and the anticipated impact of chemotherapy on neutrophil levels. The majority of the patients who have received G-CSF (n=162) have been given the drug before the start of chemotherapy, while 23 patients received it during treatment and of 62 patients it is unknown.

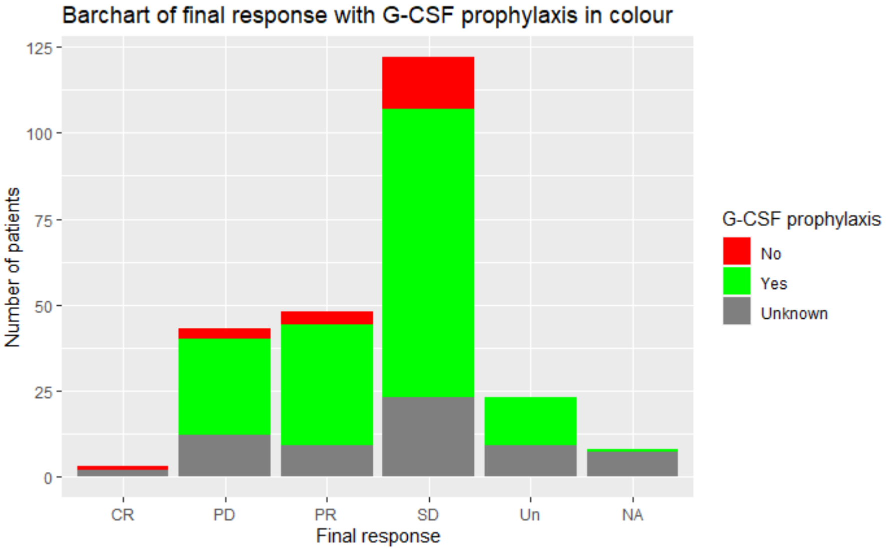


Figure B.16: Stacked bar chart displaying the distribution of treatment responses, Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), among patients grouped by G-CSF prophylaxis: no (n=23), yes (n=162), unknown (n=62), total n=247.

B.1.2 Tumor markers

B.1.2.1 Cancer Antigen 19-9 (CA19-9)

CA19-9, also known as cancer antigen 19-9, is a well-established tumor marker that is frequently used in the diagnosis and treatment of pancreatic ductal adenocarcinoma. Tumor markers are substances produced by cancer cells or normal cells in the body in response to cancer cells. Although healthy individuals may have low levels of CA19-9, it is a glycoprotein that is primarily produced by pancreatic cancer cells, as well as other types of cancer cells and normal cells in the pancreas, bile duct, and gastrointestinal tract. Elevated levels of CA19-9 in the bloodstream are most frequently associated with PDAC. However, it is essential to note that approximately 10% of the population is unable to produce this marker [103], and therefore, other diagnostic tests must be combined to accurately detect the presence of PDAC. According to medical experts from the Erasmus Medical Centre Rotterdam, the normal range for CA19-9 in healthy males is below 37 U/mL, and for females, it is below 27 U/mL. It is worth noting that the interpretation of CA19-9 levels may vary depending on the specific laboratory and the clinical context in which the test is conducted. In general, a significant increase in CA19-9 levels over time may indicate the progression of PDAC or the emergence of a recurrence following treatment. Conversely, a decrease in CA19-9 levels may suggest a positive response to treatment or the absence of disease [24].

CA19-9 Value	*Diagnosis*	*Before 1st Chemo*	*After 1st Chemo*
Min	0	0	0
1st Quartile	50.3	50.0	41.5
Median	167.0	162.5	167.0
Mean	1964.2	2533.1	2509.5
3rd Quartile	782.5	838.2	943.0
Max	149230	121632	95696
Missing Values	9	43	95

Table B.1: Summary statistics of CA19-9 values (kU/L) measured at diagnosis, before the first chemotherapy and after the first chemotherapy cycle for all 239 patients in the dataset.

Figures B.17, B.18, and B.19 show plots of the CA19-9 values measured at diagnosis, before the first chemotherapy cycle, and after the first chemotherapy cycle, respectively. Note that subplots (c) and (d) are made without the eight missing values in the data set. The first plot in each figure displays a scatterplot of the CA19-9 values plotted against the patient. The second plot shows a histogram of the CA19-9 values with a fitted density estimate. The third plot (lower left) shows a boxplot of the CA19-9 values categorized per final response value. The lower right plot shows the same boxplot but plotted in a different way. Analysis of the data shows that the first quartile and median values of CA19-9 did not change significantly across the three time periods. The mean value did increase after diagnosis before the first chemotherapy, which is expected

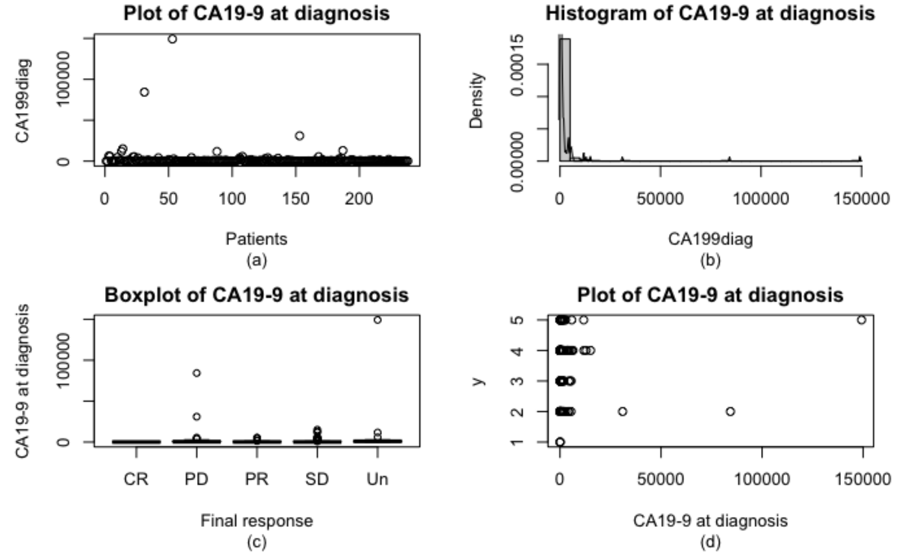


Figure B.17: CA19-9 distribution at diagnosis (kU/L) for the entire dataset: (a) scatterplot of CA19-9 at diagnosis, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CA19-9 at diagnosis across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

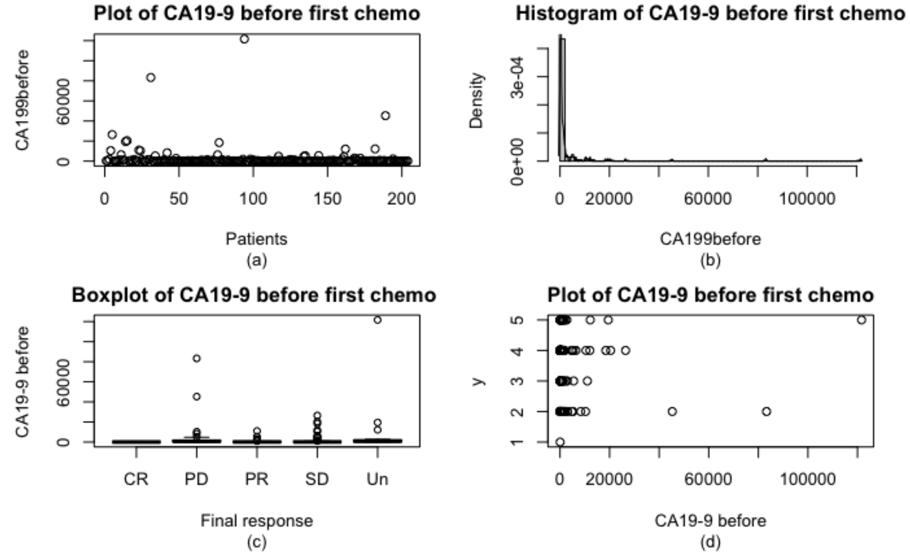


Figure B.18: CA19-9 distribution before the first chemotherapy cycle (kU/L) for the entire dataset: (a) scatterplot of CA19-9 before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CA19-9 before the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

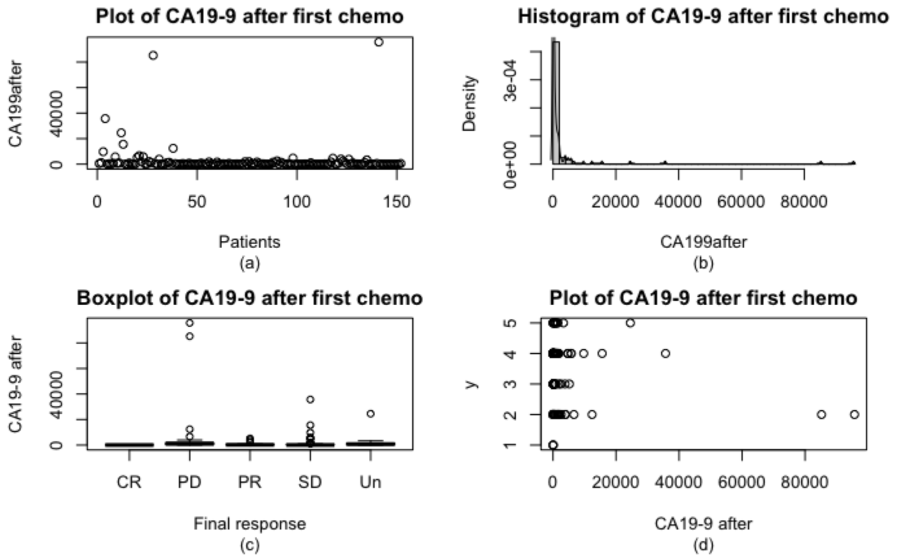


Figure B.19: CA19-9 distribution after the first chemotherapy cycle (kU/L) for the entire dataset: (a) scatterplot of CA19-9 after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CA19-9 after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

as elevated levels of CA19-9 indicate pancreatic cancer. The third quartile increased before the first chemotherapy session and after that as well, while the maximum value decreased after diagnosis. However, considering the increase in the number of missing values, a comparison of only the patients with complete data might be more appropriate. This will be done next.

After removing all missing values from the CA19-9 data, a total of 131 patients remain, with 67 male and 64 female patients. The summary statistics in Table B.2 show that the minimum values and the first quartile values remained approximately the same across the three time periods. However, the median and mean values increased. The third quartile shows a slight decrease in the CA19-9 value after the first chemotherapy, and the maximum value is slightly lower after diagnosis just before the first chemotherapy. Out of the 131 patients with complete data, only one patient had a complete response as a final response, while 25 patients showed partial response, 27 showed progressive disease, 69 showed stable disease, and 9 had an unknown final outcome.

CA19-9 value	*Diagnosis*	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	0.0	0.0	0.0	-50391.0
1st Quartile	39.5	40.0	41.0	-91.0
Median	147.0	159.0	167.0	0.0
Mean	1470.7	2422.7	2803.0	-380.6
3rd Quartile	662.6	992.5	946.0	35.0
Max	84341	83313	95696	8881

Table B.2: Summary statistics of CA19-9 levels (kU/L) in the dataset of all patients with missing values removed at diagnosis, before the first cycle, after the first chemotherapy cycle and the difference between the before and after chemotherapy values, n=131.

It may be more informative to rearrange the patient data in accordance with their final response (0, 1, 2, 3, 4) to detect any trends in the CA19-9 value plot. Figure B.21 presents the plot of the CA19-9 values for the five categories based on their final response. Notably, only one subject in this dataset exhibited a complete response, thus preventing any conclusive observations from a single observation. However, it is noticeable that the CA19-9 values of partial responders are typically lower, in contrast to the higher peak values among progressive disease patients. The stable disease group had a few elevated values, but overall presented a comparable pattern to that of the partial response group. The unknown group's values cannot be interpreted, given the absence of knowledge about their final response.

An additional interesting analysis involves examining the difference between CA19-9 values measured prior to and following the first chemotherapy session. To calculate this difference, the following formula is applied:

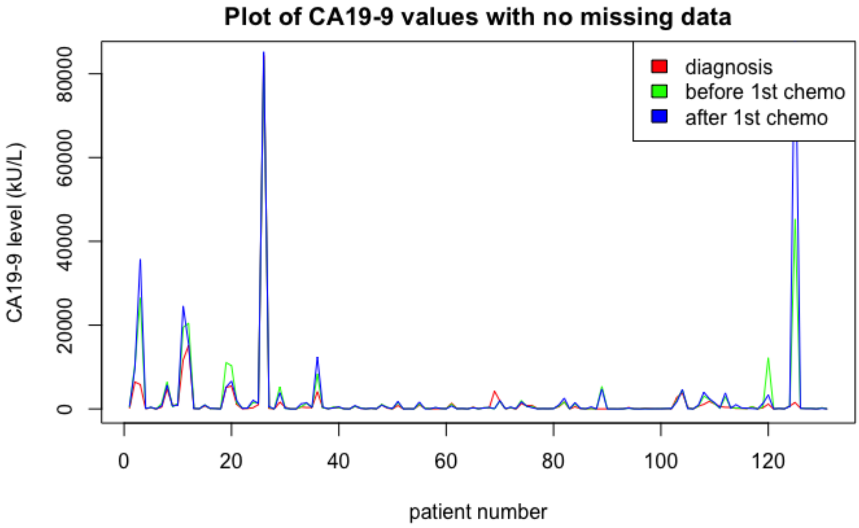


Figure B.20: CA19-9 value at diagnosis, before the first chemotherapy and after the first chemotherapy session for the CA19-9 dataset with all missing values removed, n = 131.

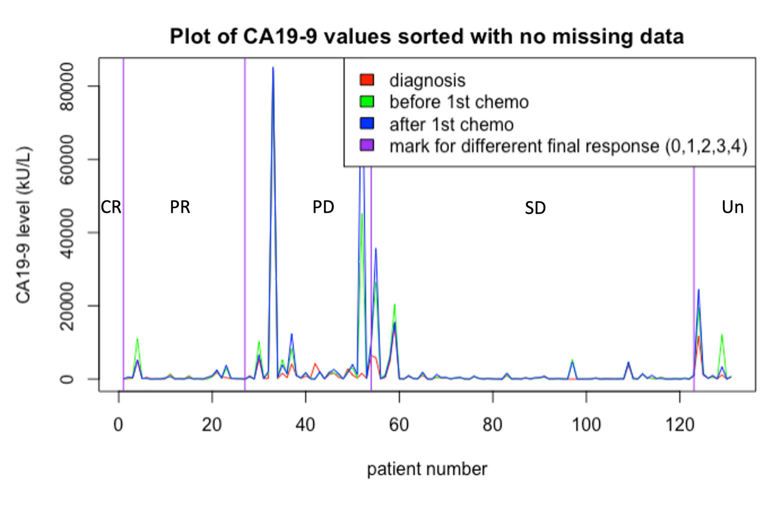


Figure B.21: CA19-9 values for each patient at diagnosis with all missing values removed, before the first chemotherapy and after the first chemotherapy cycle, sorted by their final response to chemotherapy. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=25) Progressive Disease (PD, n=27), Stable Disease (SD, n=69) and Unknown (Un, n=9), total n=131. Red = at diagnosis, Green = before first cycle, Blue = after first cycle, Purple = marks boundary between the final response.

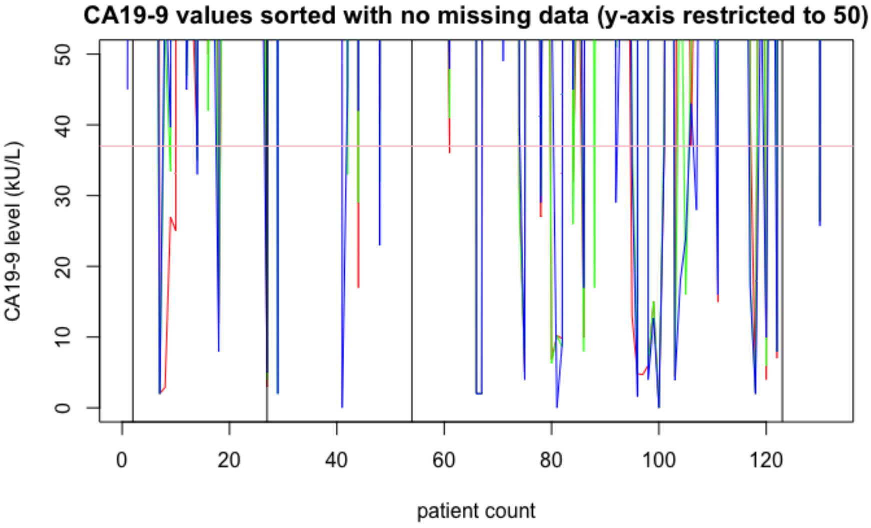


Figure B.22: Zoomed in plot of Figure B.21 with CA19-9 values restricted to levels below 50 kU/L sorted by final response.

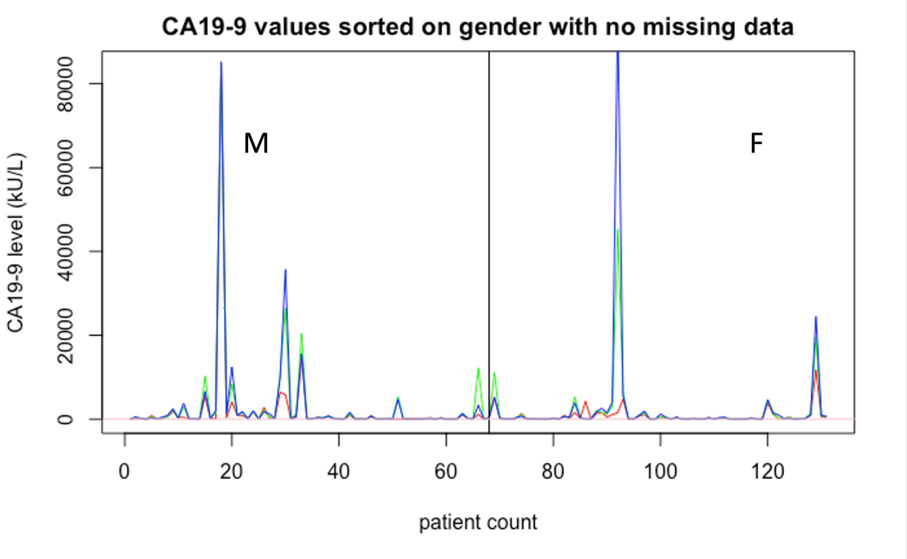


Figure B.23: CA19-9 values for each patient at diagnosis with all missing values removed, before the first chemotherapy and after the first chemotherapy cycle, sorted by gender, n = 131 (67 male, 64 female). The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=25) Progressive Disease (PD, n=27), Stable Disease (SD, n=69) and Unknown (Un, n=9), total n=131. Red = at diagnosis, Green = before first cycle, Blue = after first cycle.

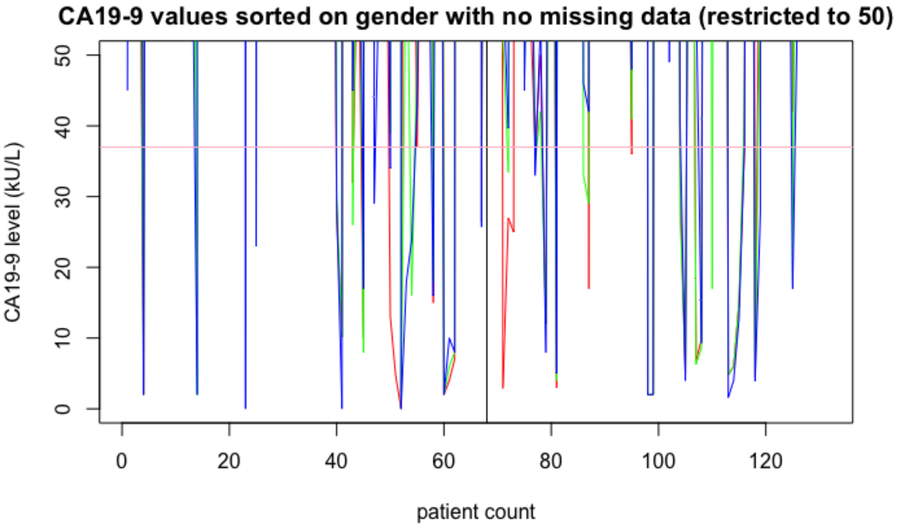


Figure B.24: Same figure as Figure B.24 with CA19-9 values restricted to 50 kU/L, with a pink line indicating the healthy range (below 37 kU/L).

$$\Delta = v_b - v_a \tag{B.3}$$

Here, Δ represents the difference between the values prior to (v_b) and after (v_a) the first chemotherapy cycle. While the term 'v' is an acronym for 'value' and represents the CA19-9 marker in this section. However, in the following subsections, the meaning of 'v' will differ depending on the investigated bloodmarker. Figure B.25 displays a graph depicting the difference between CA19-9 values before and after chemotherapy. The most significant difference is apparent within the cohort of patients whose final response was progressive disease. Nevertheless, this is valid for only one patient, who might be an outlier. Therefore, clear conclusions cannot be drawn from this plot.

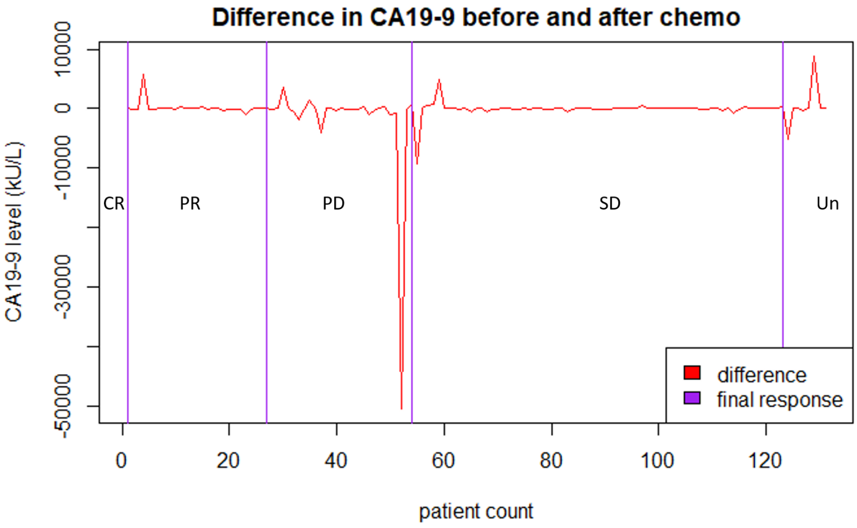


Figure B.25: Difference in CA19-9 values (kU/L) before and after first cycle for all patients with no missing data, sorted by their final response. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=25), Progressive Disease (PD, n=27), Stable Disease (SD, n=69) and Unknown (Un, n=9), total n=131.

As previously noted, a proportion of the global population, approximately 5-10%, are unable to produce the CA19-9 tumor marker, and are classified as “non-producers". To ensure a more accurate analysis, it is preferable to exclude these individuals from the dataset. In this context, non-producers are identified as patients whose CA19-9 concentration falls below 35 kU/L. These patients are excluded from the dataset as well next to the patients with missing values. A similar analysis as previously done is conducted and the summary values can be found in Table B.3. The plots provided below show that the progressive disease group contains the largest values for CA19-9 after the first chemotherapy cycle as well as

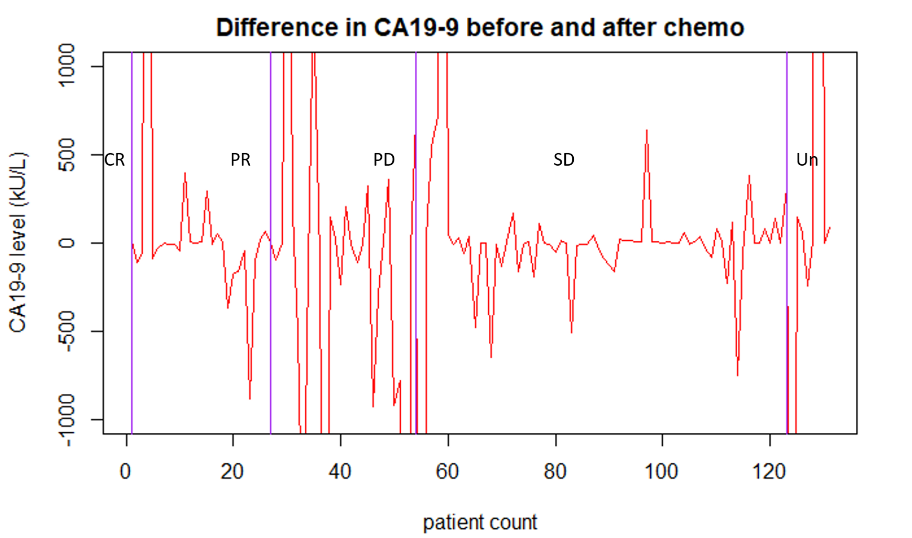


Figure B.26: Same plot as Figure B.25 but with CA19-9 values restricted between -1000 to +1000 kU/L.

the biggest differences before and after the first cycle.

CA19-9 value	*Diagnosis*	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	36.0	41.0	36.8	-50391.0
1st Quartile	127.0	138.9	130.5	-164.0
Median	354.0	388.0	467.0	-9.0
Mean	2056.0	3415.0	3971.4	-556.4
3rd Quartile	1006.0	1435.5	1668.0	77.9
Max	84341.0	83313.0	95696.0	8881.0

Table B.3: Summary statistics of CA19-9 levels (kU/L) of all patients with no missing values and non-producers excluded at diagnosis, before the first chemotherapy, after the first chemotherapy cycle, and the difference between the before and after first cycle values, n = 91.

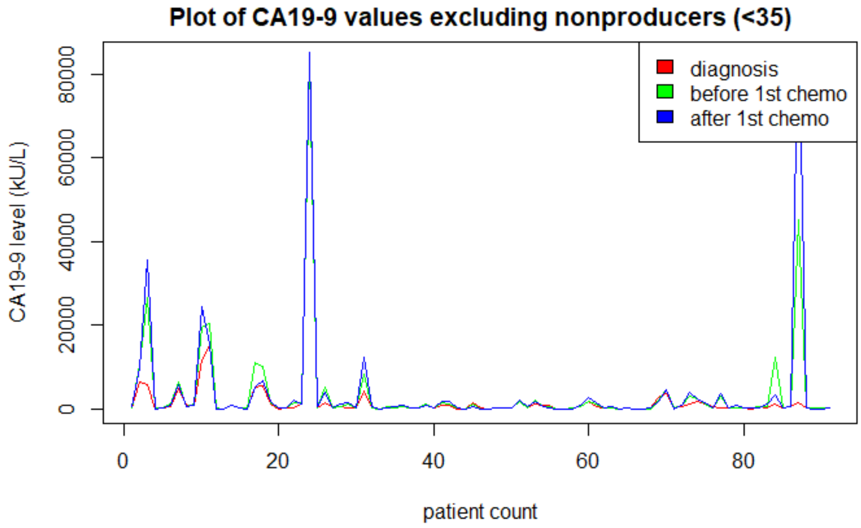


Figure B.27: CA19-9 values, at diagnosis, before the first chemotherapy and after the first chemotherapy cycle in a subset of patients with no missing data or non-producers (CA19-9 < 35 kU/L), n = 91.

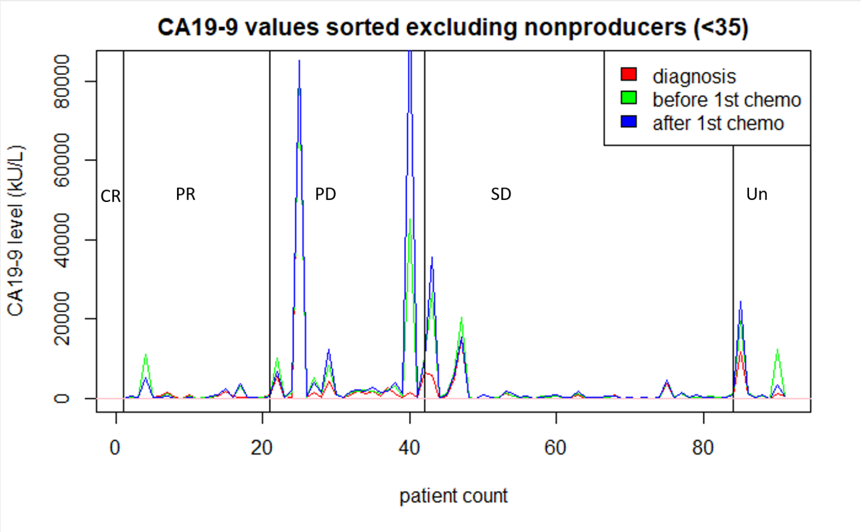


Figure B.28: CA19-9 values at diagnosis, before the first chemotherapy and after the first chemotherapy cycle, in a subset of patients with no missing data or non-producers (CA19-9 $<$ 35 kU/L) sorted by their final response to chemotherapy, n = 91. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=19), Progressive Disease (PD, n=21), Stable Disease (SD, n=42) and Unknown (Un, n=8), total n = 91.

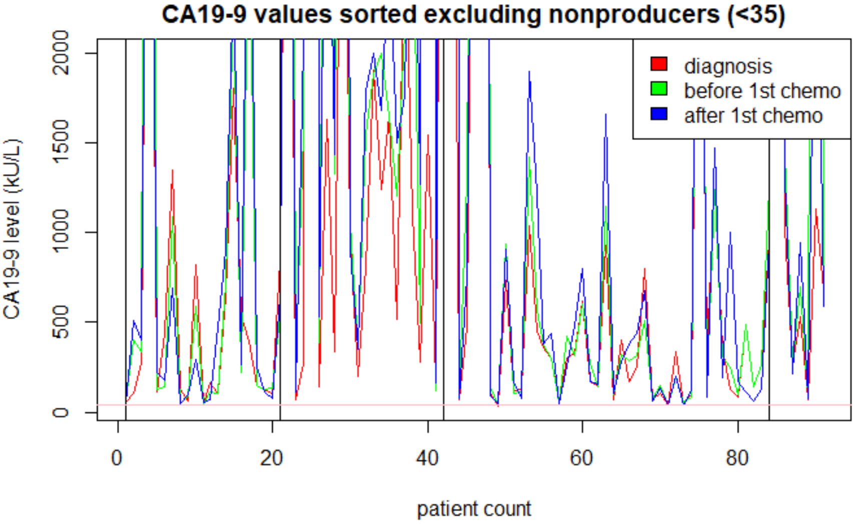


Figure B.29: Same plot as Figure B.28 but with CA19-9 values ranging between 35-2000 kU/L. This zoomed-in view allows for a closer examination of the CA19-9 levels within the clinically relevant range.

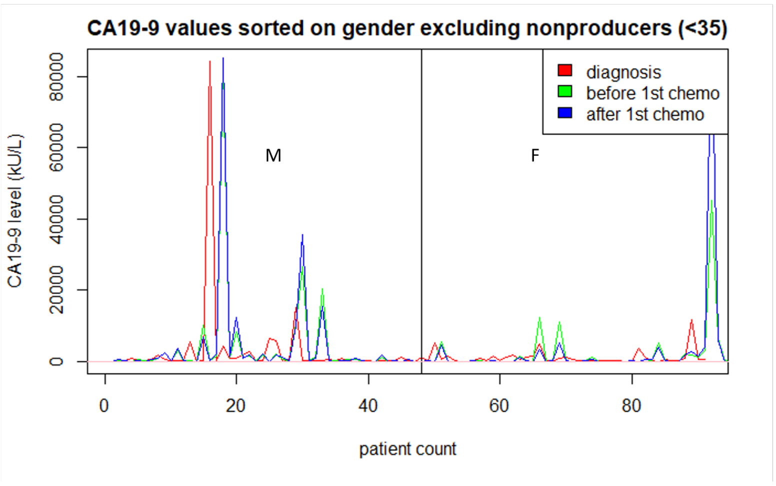


Figure B.30: CA19-9 values, at diagnosis, before the first chemotherapy and after the first chemotherapy cycle in a subset of patients with no missing data or non-producers (CA19-9 $<$ 35 kU/L) sorted by gender, n = 91 (48 male, 43 female).

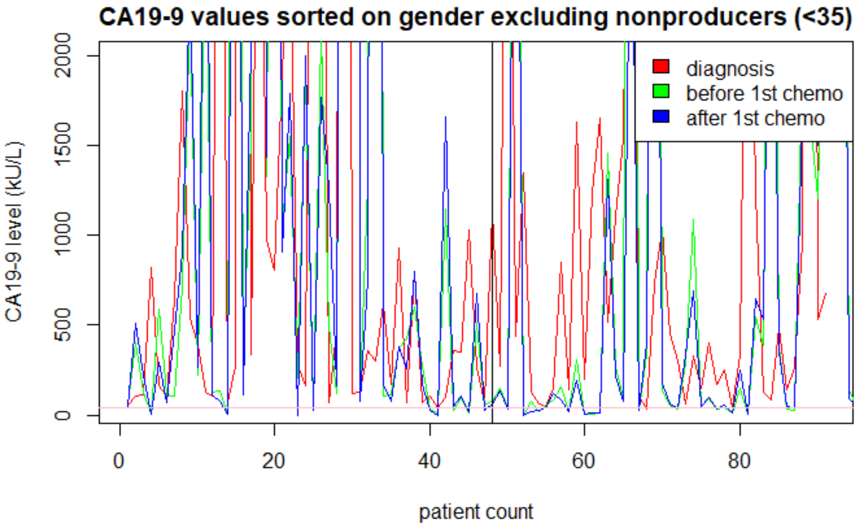


Figure B.31: Same plot as Figure B.30 with CA19-9 values ranging between 35-2000 kU/L, n = 91 (48 male, 43 female).

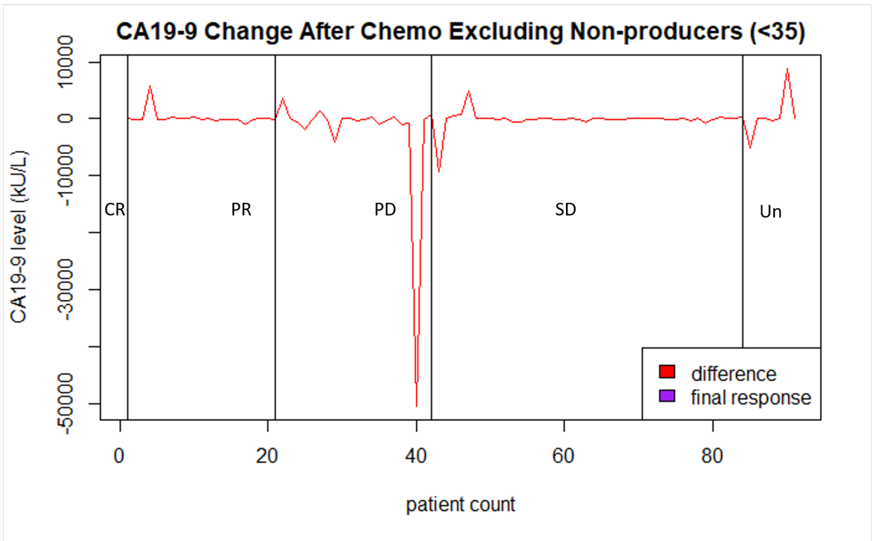


Figure B.32: Difference in CA19-9 values (kU/L) before and after first cycle for subset of patients with no missing data or non-producers (CA19-9 < 35 kU/L) , sorted by their final response. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=19), Progressive Disease (PD, n=21), Stable Disease (SD, n=42) and Unknown (Un, n=8), total n = 91.

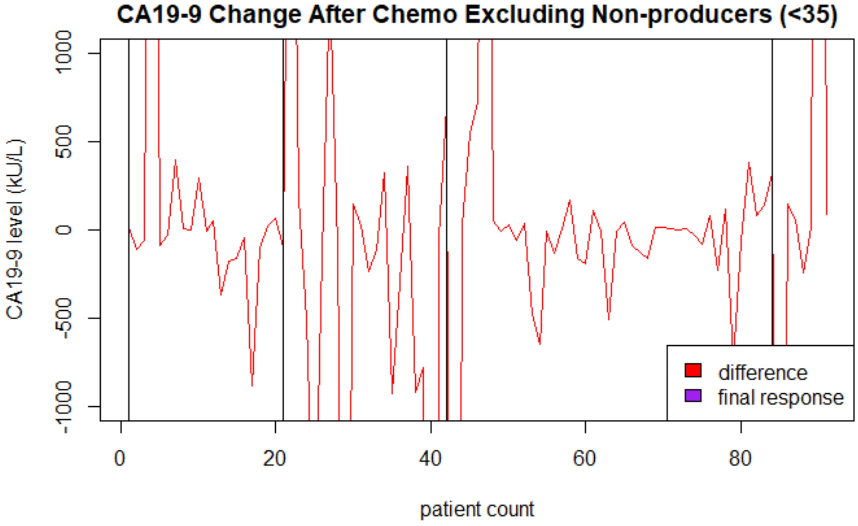


Figure B.33: Same plot as Figure B.32 with CA19-9 difference values ranging between -1000 to + 1000 kU/L.

B.1.2.2 Carcinoembryonic antigen (CEA)

Carcinoembryonic antigen (CEA) is another tumor marker, next to CA19-9, frequently used in the diagnosis and treatment of different types of cancer. While CEA is normally produced during fetal development, its presence in healthy adults is either negligible or completely absent. In cancer, the cancerous cells produce CEA, which is then released into the bloodstream. It is essential to note that elevated CEA levels in the blood can also be caused by non-cancerous conditions, such as inflammation, infection, and smoking. However, not all individuals with cancer have elevated levels of CEA, and some individuals with non-cancerous conditions may have elevated CEA levels. Therefore, it is important to use CEA tests in combination with other diagnostic tests to draw accurate conclusions. Some patients with PDAC have elevated levels of CA19-9 but normal CEA levels, while others have elevated levels of both CA19-9 and CEA. For healthy adults, normal CEA levels are below $2.5ng/mL$ for females and below $3.8ng/mL$ for males. However, in smokers, slightly higher levels up to $5ng/mL$ may be considered normal. CEA levels greater than $10ng/mL$ are suggestive of extensive disease [104].

CEA value	*Diagnosis*	*Before 1st Chemo*	*After 1st Chemo*
Min	0	0	0
1st Quartile	2.2	2.1	2.2
Median	3.9	3.7	3.9
Mean	54.3	16.4	33.6
3rd Quartile	7.7	8.1	8.2
Max	5374	543	2822
Number of NA's	64	63	107

Table B.4: Summary statistics of CEA values measured at diagnosis, before and after the first chemotherapy cycle for all 247 patients in the dataset.

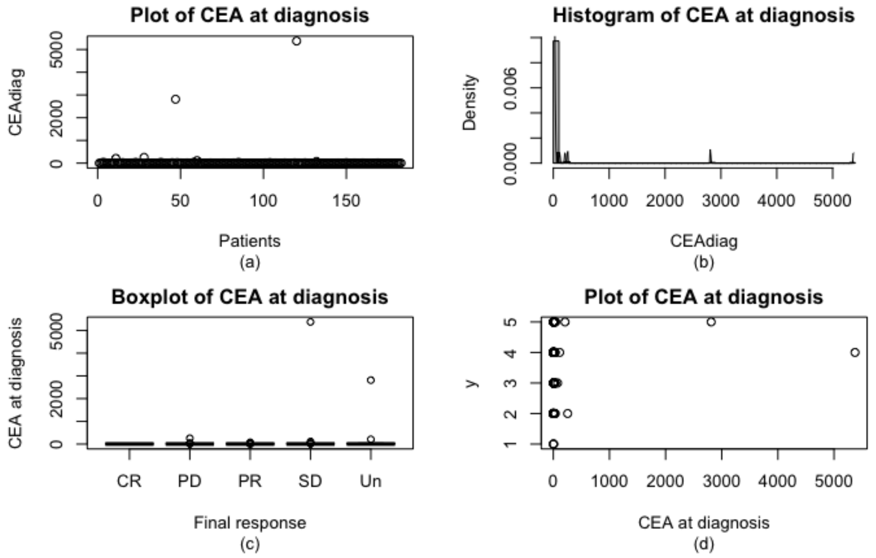


Figure B.34: CEA distribution at diagnosis ($\mu g/L$) for the entire dataset: (a) scatterplot of CEA at diagnosis, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CEA at diagnosis across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

The summary statistics presented in Table B.4 show that the minimum, first quartile, median and third quartile values of CEA levels are similar across the three measurements, whereas the mean value is highest at diagnosis and lowest before the first chemotherapy. Moreover, the maximum value is highest at diagnosis, but after the first chemotherapy, it is only about half of the maximum value at diagnosis, while the maximum value before the first chemotherapy is only about 10% of the largest value found at diagnosis. Further inspection of the data shows that the high CEA values at diagnosis are primarily found in two patients, with the five highest values being 5374, 2812, 255, 209, and 112.7 $\mu g/L$. The five highest values before the first chemotherapy are 543, 299, 231, 226, and 119 $\mu g/L$. Additionally, only one patient exhibited a very high CEA value of 2822 $\mu g/L$ after the first chemotherapy, with the five highest values being 2822, 384, 241, 139, and 98.8 $\mu g/L$. Thus, there are three outlier values of 5374, 2812, and 2822 $\mu g/L$ that deviate from the rest of the data, with these values belonging to patients 078PANC0004, 002PANC0010, and 078PANC0006, respectively.

After removing all missing values, the dataset contains n=103 observations, with only one patient exhibiting a complete

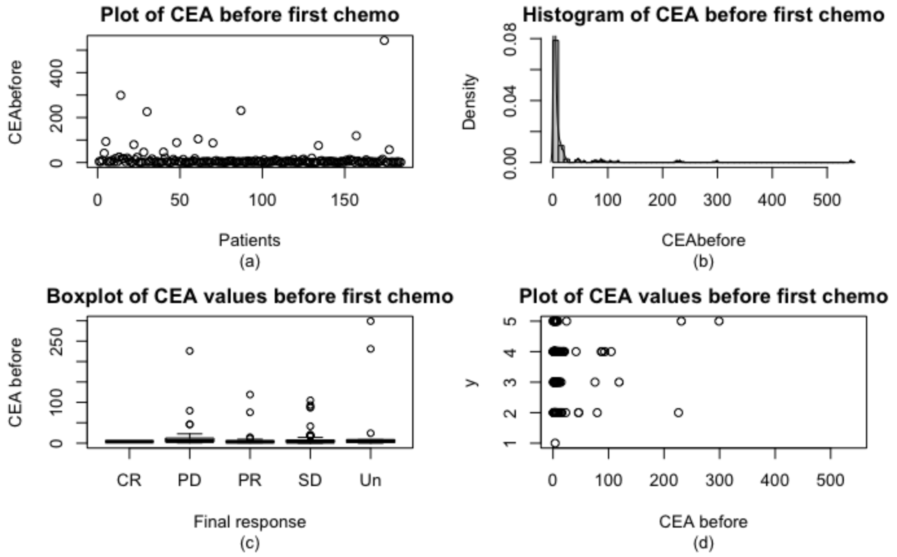


Figure B.35: CEA distribution before the first chemotherapy cycle ($\mu g/L$) for the entire dataset:(a) scatterplot of CEA before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CEA before the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

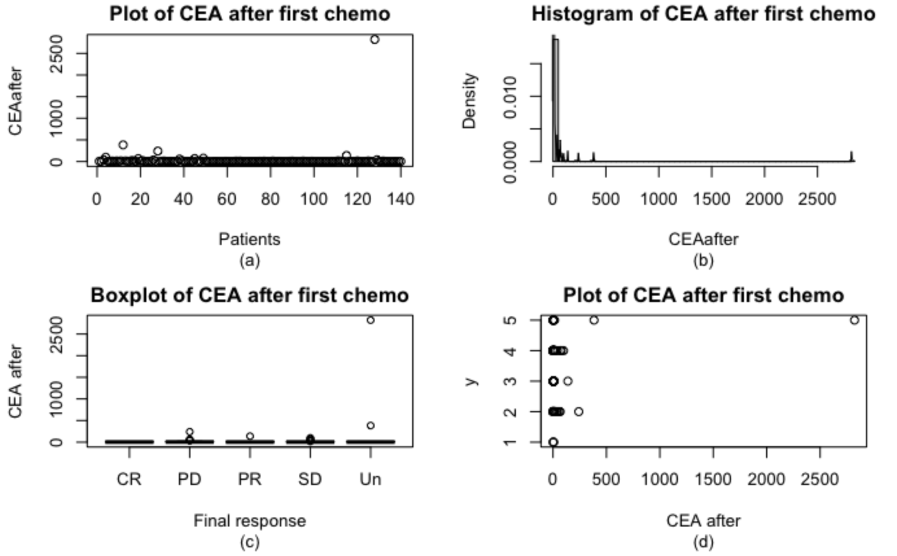


Figure B.36: CEA distribution after the first chemotherapy cycle ($\mu g/L$) for the entire dataset: (a) scatterplot of CEA after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CEA after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

response, 19 (18%) showing partial response, 20 (19%) showing progressive disease, 56 (54%) showing stable disease, and 7 (7%) having an unknown response. Grouping the patients based on their final response and plotting the CEA measurements at the three time points in Figure B.38 reveals that the highest spikes occur in the progressive disease group, but patients in the progressive response and stable disease group also contain patients with enormous peaks. However, the majority of the patients in these two groups have a CEA value lower than 50 $\mu g/L$. In Figure B.42 a plot of the difference between the CEA values before and after chemotherapy can be found and it can be seen that aside from a few outliers in the unknown final response category, all the other values seem to be relatively stable around the origin. It is better to look at it zoomed in, this figure can be found in Figure B.43.

CEA value	*Diagnosis*	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	0.0	0.0	0.0	-2822.0
1st Quartile	2.0	2.0	2.2	-1.1
Median	3.8	3.6	3.9	-0.1
Mean	11.3	15.1	42.4	-27.3
3rd Quartile	8.0	9.3	9.0	0.8
Max	225	299	2822	98

Table B.5: Summary statistics of CEA levels ($\mu g/L$) in the dataset of all patients with missing values removed at diagnosis, before the first cycle, after the first chemotherapy cycle and the difference between the before and after chemotherapy values, n=103.

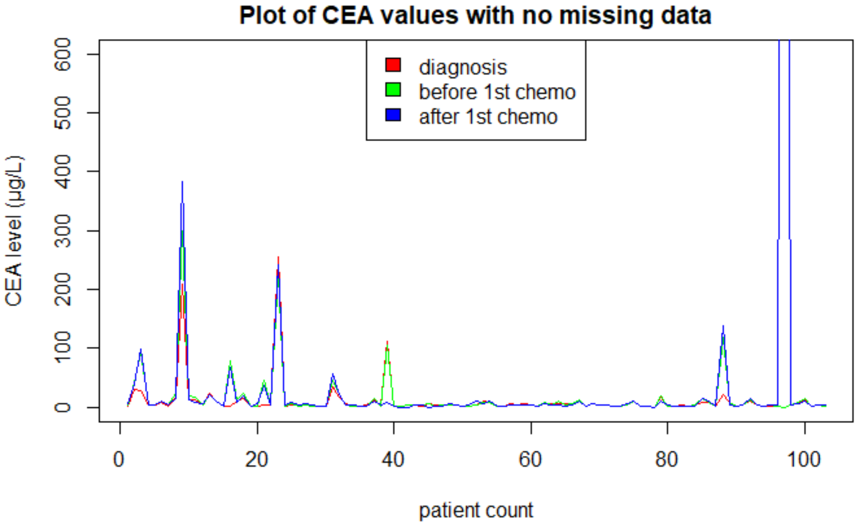


Figure B.37: CEA values at diagnosis, before the first chemotherapy and after the first chemotherapy cycle with all missing values removed, n = 103.

B.1.2.3 Comparison of CA19-9 and CEA levels

Investigating the relationship between CA19-9 and CEA levels in pancreatic ductal adenocarcinoma (PDAC) patients can provide useful clinical insights. The potential relationship between these two biomarkers is examined in Figure B.44. Although the plot suggests that high CA19-9 levels may be associated with high CEA levels, it is not a consistent trend across all patients. To further explore this relationship, a more detailed visualization is presented in Figure B.45, Figure B.46, Figure B.47, and Figure B.48. However, there appears to be no clear pattern or trend, except that most low CA19-9 values correspond to low CEA levels. The data set has a total of n=103 observations with n=1 CR, n=19 PR, n=20 PD, n=56 SD and n=7 Unknown values. The relationship between CA19-9 and CEA in PDAC patients is complex and not fully understood. While there may be some association between CA19-9 and CEA levels in PDAC patients, the relationship is not straightforward, and further research is needed to fully understand their clinical significance.

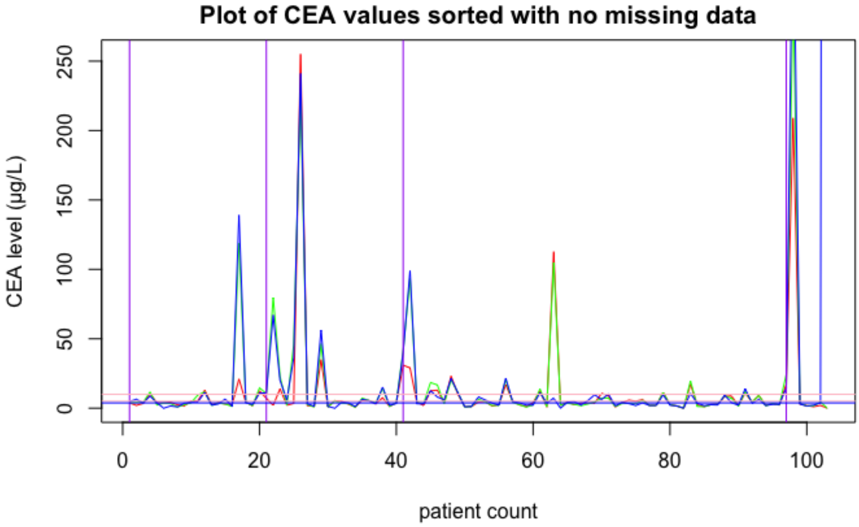


Figure B.38: CEA values at diagnosis, before the first chemotherapy and after the first chemotherapy cycle with all missing values removed, sorted by final response to chemotherapy. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=19) Progressive Disease (PD, n=20), Stable Disease (SD, n=56) and Unknown (Un, n=7), total n=103. Red = at diagnosis, Green = before first cycle, Blue = after first cycle, Purple = marks boundary between the final response.

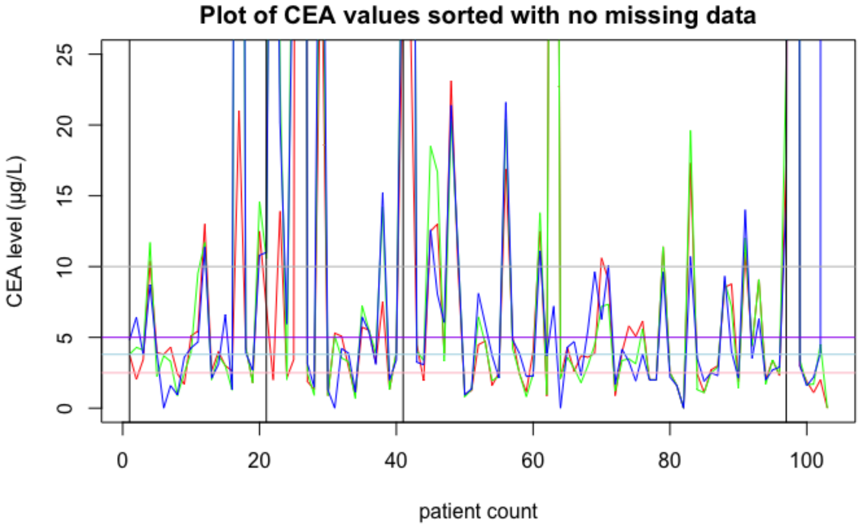


Figure B.39: Same plot as Figure B.38, while constraining CEA values to the range of 0-25 $\mu g/L$. The chart displays four distinct lines: a grey line at 10 $\mu g/L$, a purple line at 5 $\mu g/L$, a blue line at the male healthy threshold of 3.8 $\mu g/L$, and a pink line at the female healthy threshold of 2.5$\mu g/L$. Additionally, black lines separate the different response groups, namely: Complete response (CR, n=1), Partial Response (PR, n=19) Progressive Disease (PD, n=20), Stable Disease (SD, n=56) and Unknown (Un, n=7), total n=103.

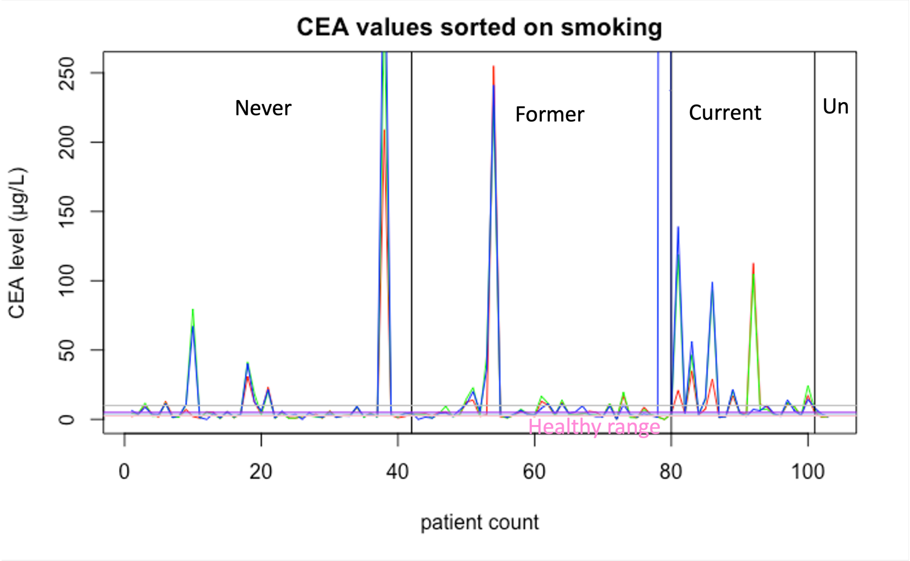


Figure B.40: CEA values for a cohort of patients with missing data removed, sorted based on their smoking status. Grey = $\mu g/L$, Purple = $5\mu g/L$, Blue = healthy male boundary at 3.8 $\mu g/L$, Pink line = healthy female boundary at $2.5\mu g/L$. The distribution among the smokers is never (n=41), former (n=38), current (n=21) and unknown (n=3), total n=103.

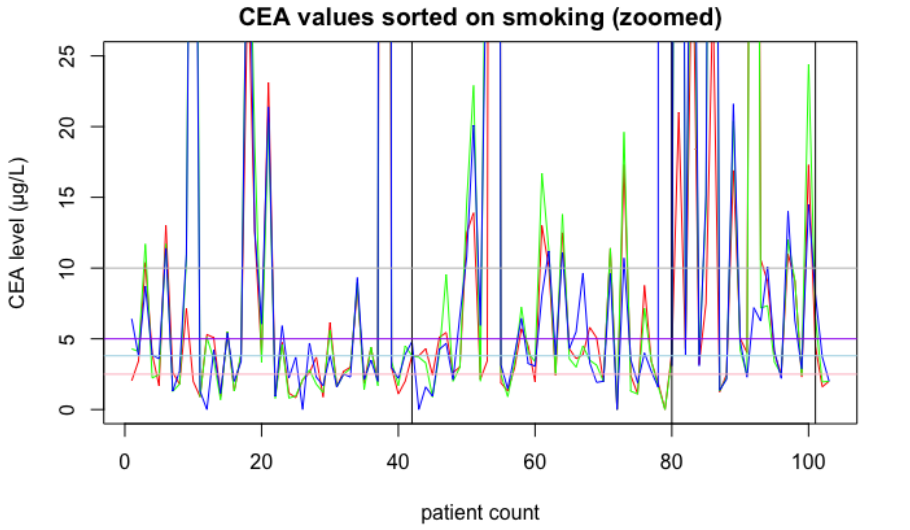


Figure B.41: Same plot as Figure B.41 with CEA values restricted from 0-15 $\mu g/L$.

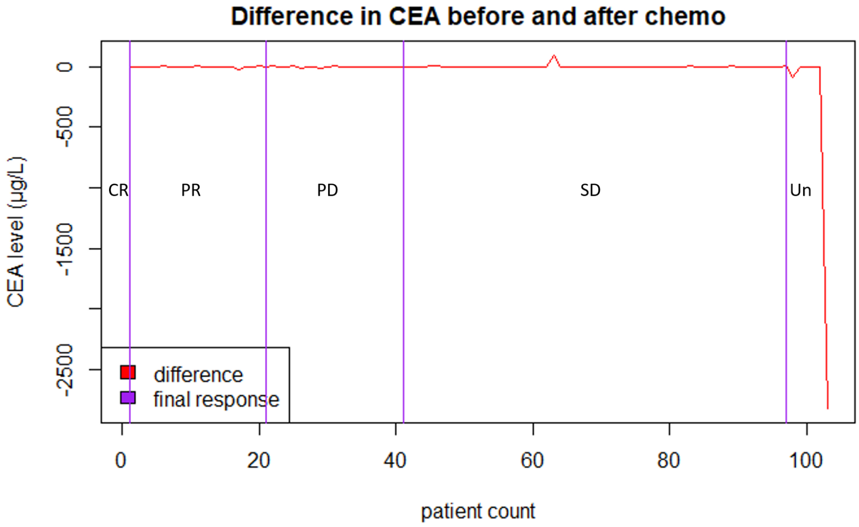


Figure B.42: Difference in CEA values (kU/L) before and after first cycle for all patients with no missing data, sorted by their final response. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=19) Progressive Disease (PD, n=20), Stable Disease (SD, n=56) and Unknown (Un, n=7), total n=103.

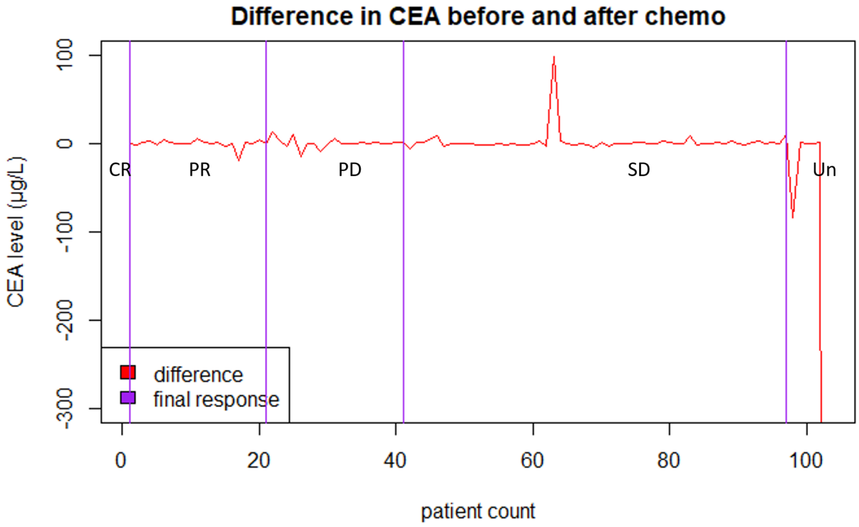


Figure B.43: Same plot as Figure B.42 with CEA values ranging between -300 to + 100 $\mu g/L$.

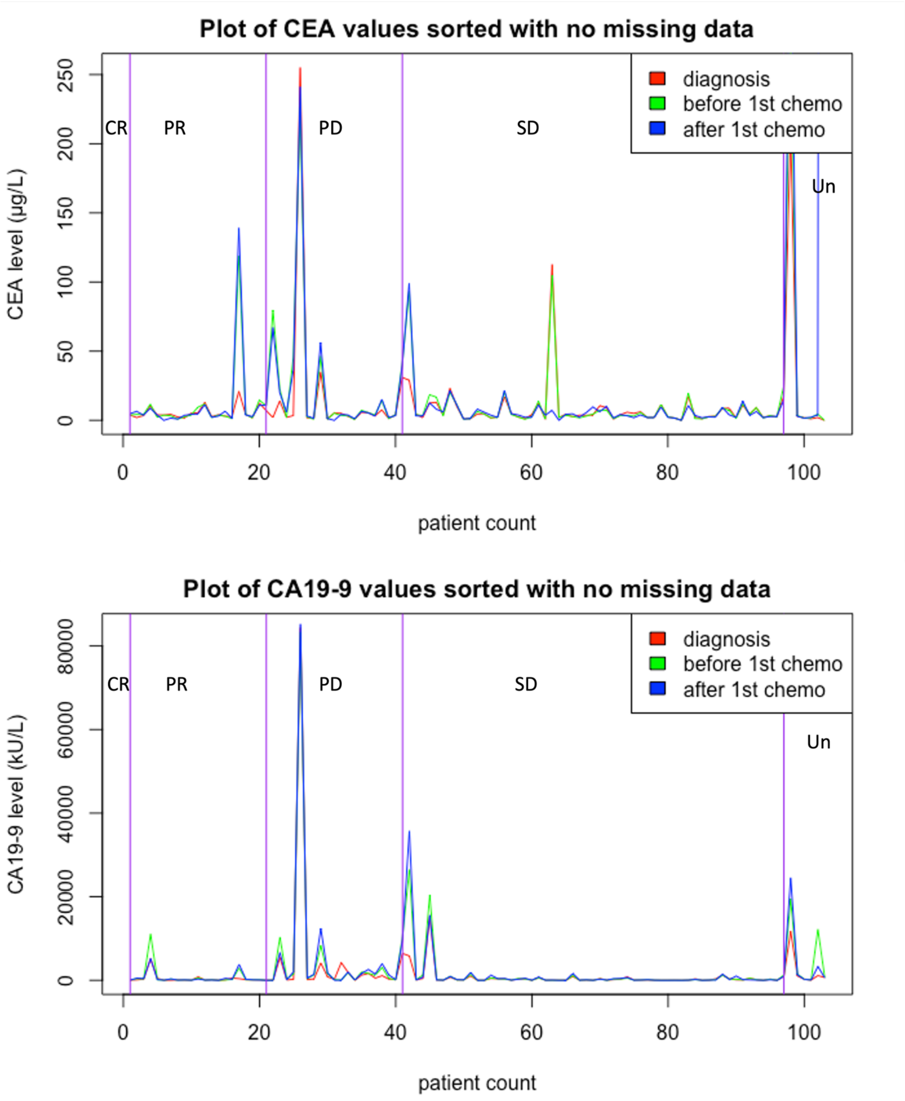


Figure B.44: CA19-9 and CEA values at diagnosis, before the first chemotherapy and after the first chemotherapy cycle for a subset of patients with missing values removed data, sorted based on their final response. The final response is categorized as: Complete response (CR, n=1), Partial Response (PR, n=19) Progressive Disease (PD, n=20), Stable Disease (SD, n=56) and Unknown (Un, n=7), total n=103.

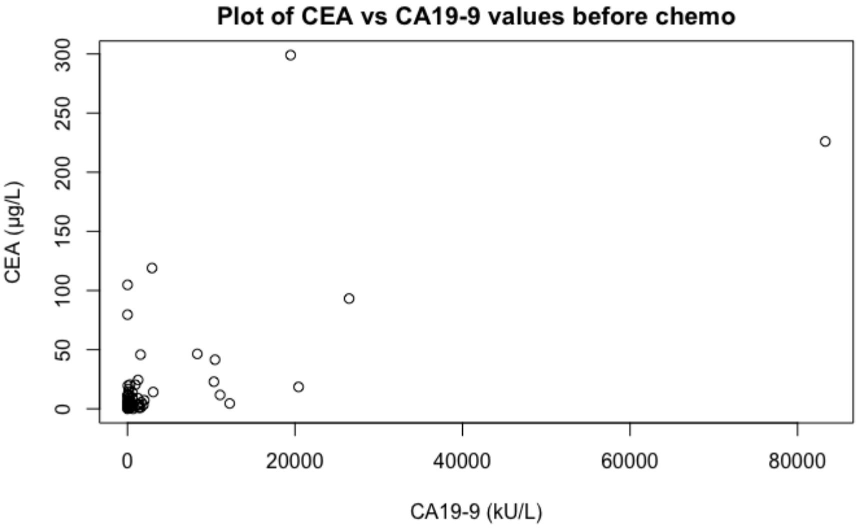


Figure B.45: Scatter plot of the cancer markers CEA (μg) and CA19-9 (kU/L) against each other, for a cohort of patients prior to the first chemotherapy cycle, n=103.

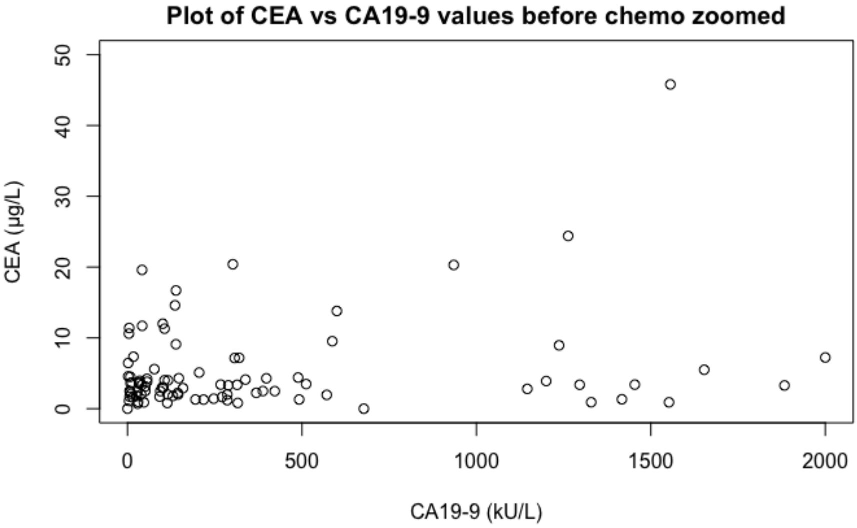


Figure B.46: Same plot as Figure B.45 with CEA levels restricted from 0-50 $\mu g/L$ and CA19-9 from 0-2000 kU/L.

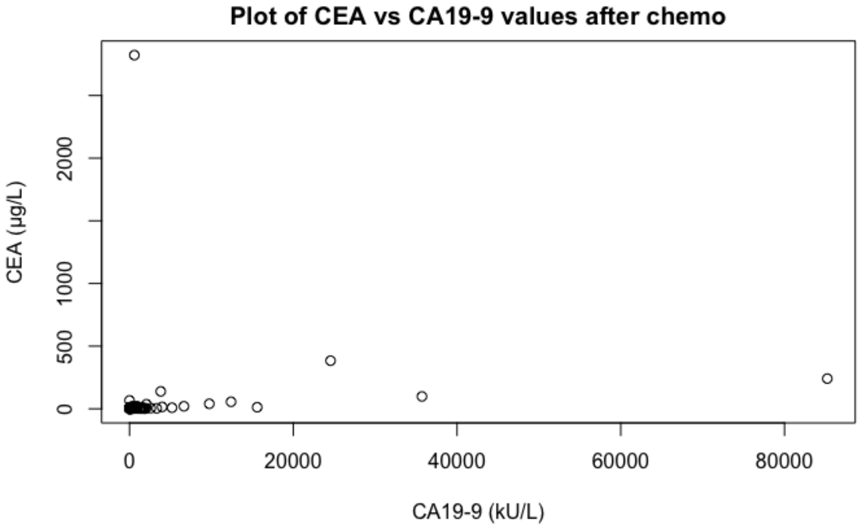


Figure B.47: Scatter plot of the cancer markers CEA (μg) and CA19-9 (kU/L) against each other, for a cohort of patients after the first chemotherapy cycle, n=103.

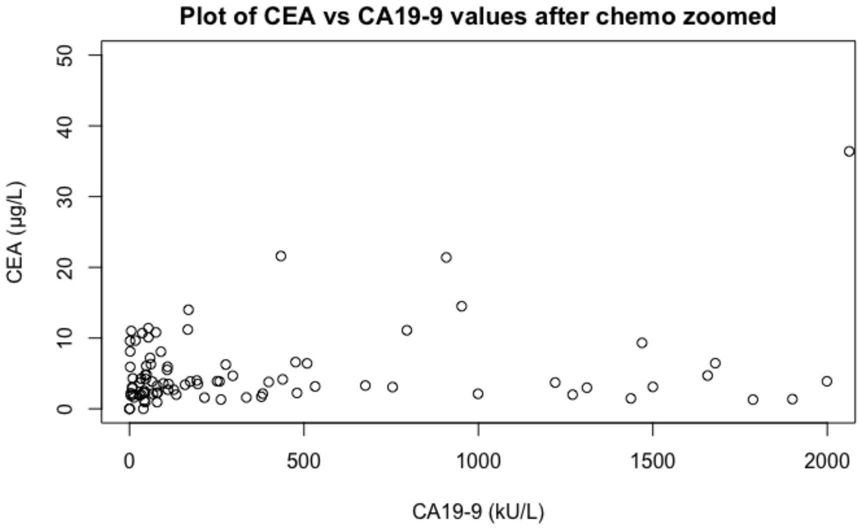


Figure B.48: Same plot as Figure B.47 with CEA levels restricted from 0-50 $\mu g/L$ and CA19-9 from 0-2000 kU/L.

B.1.3 Blood markers

This section contains an analysis of the blood markers measured before and after the first chemotherapy cycle with FOLFIRINOX. The measured blood markers are:

1. Hemoglobin
2. Thrombocytes
3. Leukocytes
4. Neutrophils
5. Lymphocytes
6. Creatinin
7. Glomular Filtration Rate
8. Sodium
9. Potassium
10. Aspartate Aminotransferase
11. Alanine Aminotransferase
12. Alkaline Phosphatase
13. Gamma-Glutamyl Transferase
14. Bilirubin
15. Albumin
16. C-Reatice Protein
17. International Normalized Ratio
18. Systemic Inflammation Index (added)
19. Neutrophil-to-Lymphocyte Ratio (added)
20. Platelet-to-Lymphocyte Ratio (added)

B.1.3.1 Hemoglobin (HB)

Hemoglobin is a crucial protein found in red blood cells that plays a vital role in oxygen transport to the body's organs and tissues and removal of carbon dioxide from these tissues back to the lungs. According to medical sources, the normal range of hemoglobin in healthy individuals is 13.2 to 16.6 g/dL for men and 11.6 to 15 g/dL for women [105]. When converted to SI units, these values are approximately 8.7-11.2 mmol/L for men and 7.4-9.9 mmol/L for women. An examination of the data statistics presented in Table B.6 indicates that the hemoglobin values before and after chemotherapy are relatively similar, with some variations observed in the maximum values. However, these do not differ significantly.

Figure B.49 and Figure B.50 display various plots depicting the distribution of hemoglobin (HB) values. The histogram suggests a normal distribution of HB values. To further investigate this, a QQplot is generated, and the plots for HB values before and after chemotherapy are presented in Figure B.51. The QQplots, similar to the histogram, reveal that the HB values are approximately normally distributed. However, there are some deviations from normality observed in the tails, particularly after the initial chemotherapy session. In order to assess the normality of the data more rigorously, the Shapiro-Wilk test and the Kolmogorov–Smirnov test were conducted on the dataset, and the outcomes are presented in Table B.7. The Shapiro-Wilk test does not reject normality for both the before and after first cycle HB values, while the KS-test indicates non-normality.

To provide some more background information on this observation. The Shapiro-Wilk test and the Kolmogorov-Smirnov (KS) test are both statistical tests used to assess normality of a distribution. However, they are based on different criteria and assumptions. The Shapiro-Wilk test is a sensitive test to detect deviations from normality in the center of the distribution. It tests the null hypothesis that a sample is drawn from a normally distributed population. If the p-value is greater than the significance level (usually 0.05), then the null hypothesis cannot be rejected, and the data is considered to be normally distributed.On the other hand, the KS test is more sensitive to deviations from normality in the tails of the distribution. It tests the null hypothesis that a sample is drawn from a specific continuous distribution, usually a normal distribution (in this case as well). If the p-value is greater than the significance level, then the null hypothesis cannot be rejected, and the data is considered to be consistent with the specified distribution. If the p-value is less than the significance level, then the null hypothesis is rejected, and the data is considered to be inconsistent with the specified distribution. In the case of the hemoglobin values, the histogram and QQplot suggest that the distribution is approximately normal. However, the KS test indicates that the distribution is not exactly normal, especially in the tails. The Shapiro-Wilk test, being less sensitive to deviations in the tails, does not detect non-normality in the data. Therefore, the Shapiro-Wilk test and the KS test are not in contradiction with each other, but rather they provide complementary information about the normality of the distribution.

HB value	*Before 1st Chemo*	*After 1st Chemo*
Min	5.5	5.4
1st Quartile	7.3	7.0
Median	7.9	7.5
Mean	7.8	7.5
3rd Quartile	8.5	8.0
Max	10.7	9.8
NA	7	14

Table B.6: Summary statistics of HB values (mmol/L) of the entire data set, n=247.

HB value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.991	0.175
SW-test pvalue	0.995	0.650
KS test pvalue	Invalid	Invalid

Table B.7: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the hemoglobin values, n=223.

To get a better understanding of the hemoglobin values, the missing data has been removed and the HB values are sorted based on the patients' response to chemotherapy. Table B.8 presents the summary statistics of the hemoglobin (HB) values in mmol/L for a cohort of 223 patients, comprising 124 male and 99 female subjects. The table includes an additional column, HBdiff, which represents the difference in HB values before and after chemotherapy. The final response to chemotherapy was categorized into 3 complete response (CR), 46 partial response (PR), 40 progressive disease (PD), 118 stable disease (SD), and 16 unknown responses (Un). Furthemore, the HB values before and after the first chemotherapy session are displayed in Figure B.52. Figure B.53 illustrates the HB values sorted by the patients' final response, with the healthy female range indicated in pink and the male range in blue. A similar split between male and female can be seen in Figure B.54, where the values differ per group. On average, the HB values were lower than the healthy range. Figure B.55 displays the difference between the HB values before and after the first chemotherapy session. The decrease in hemoglobin can be attributed to chemotherapy-induced anemia, which can result from a decrease in the number of red blood cells produced in

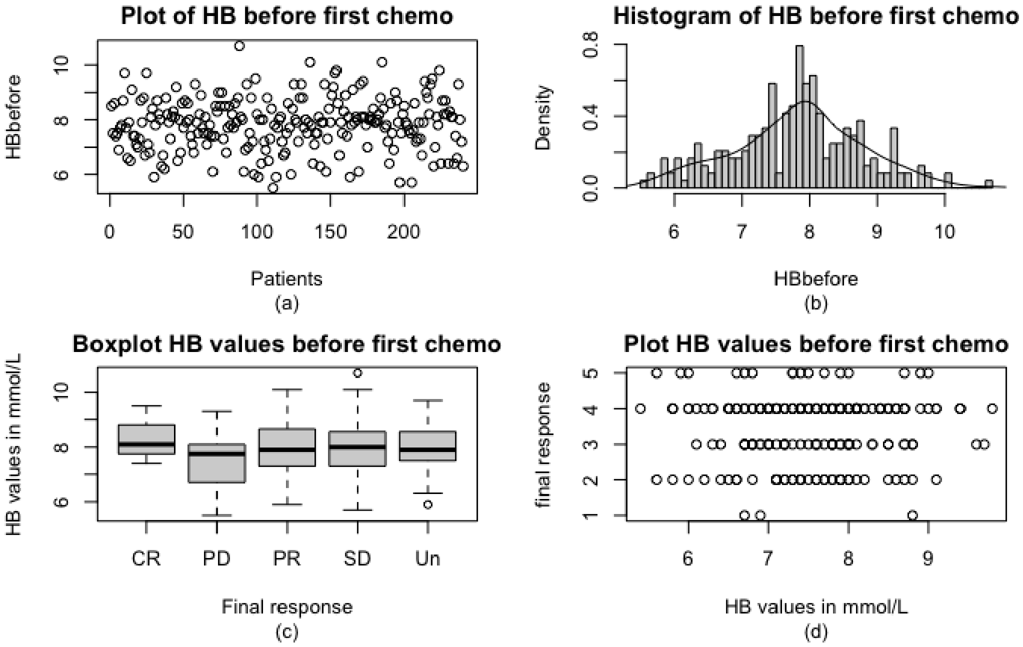


Figure B.49: Hemoglobin distribution before the first chemotherapy cycle (mmol/L) for the entire dataset: (a) scatterplot of hemoglobin values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of hemoglobin values before the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

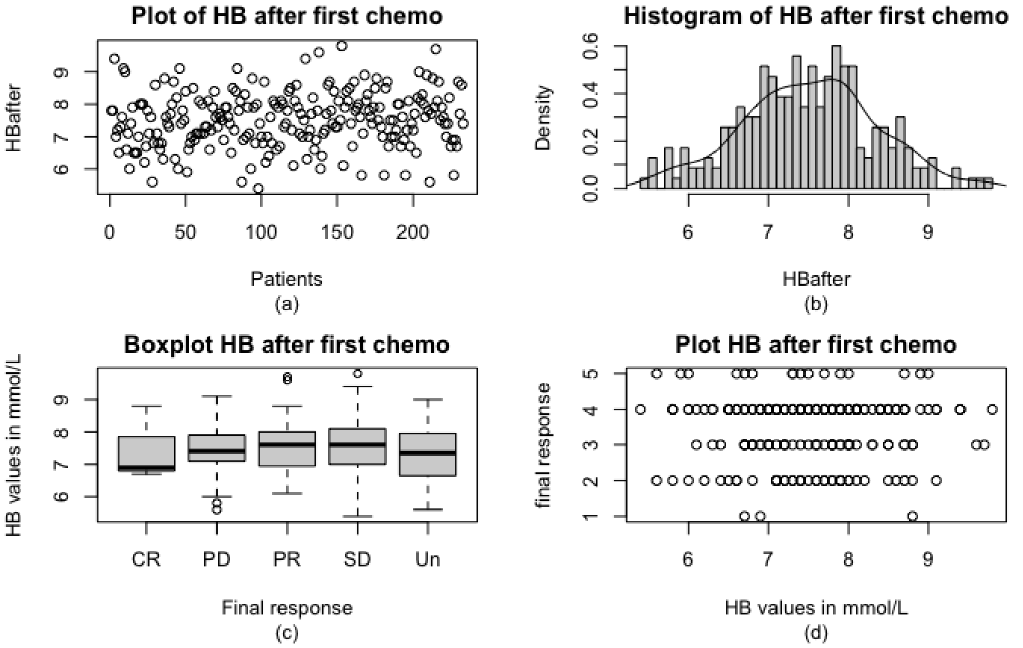


Figure B.50: Hemoglobin distribution after the first chemotherapy cycle (mmol/L) for the entire dataset: (a) scatterplot of hemoglobin values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of hemoglobin values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

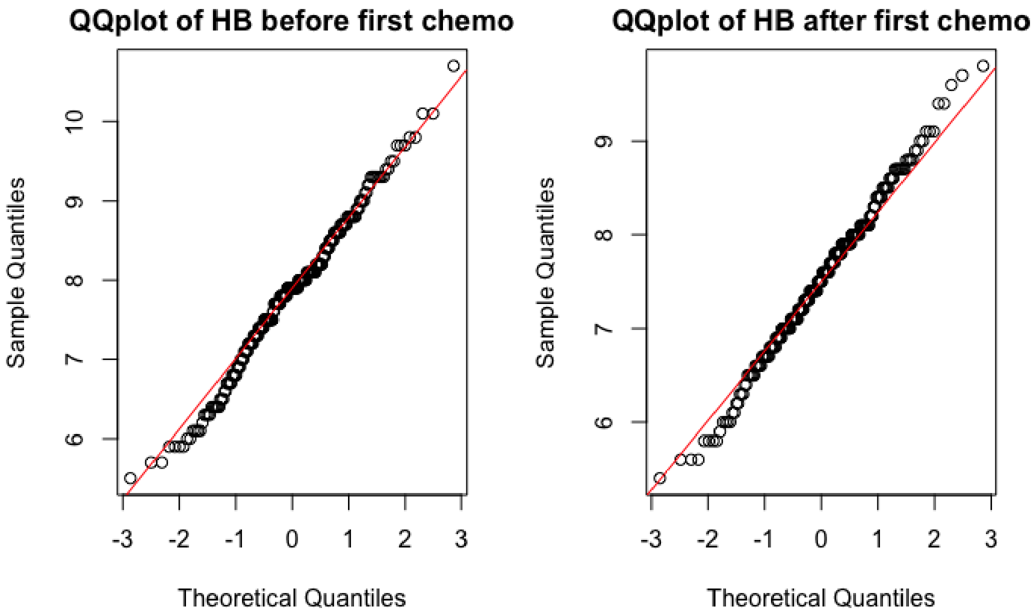


Figure B.51: QQplot of the hemoglobin values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=223.

the bone marrow or damage to the red blood cells themselves. Additionally, chemotherapy drugs can cause gastrointestinal symptoms, such as nausea, vomiting, and diarrhea, which can lead to poor nutrient absorption and subsequent anemia. Additionally, some chemotherapy drugs can cause damage to the red blood cells themselves, leading to hemolysis and a decrease in hemoglobin levels.

HB value	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	5.5	5.4	-1.7
1st Quartile	7.3	7.0	-0.05
Median	7.9	7.5	0.3
Mean	7.8	7.5	0.3
3rd Quartile	8.5	8.1	0.7
Max	10.7	9.8	2.1

Table B.8: Summary statistics of HB values (mmol/L) before and after the first chemotherapy cycle with all missing values removed, n=223. An additional column displaying the difference between hemoglobin values before and after the first cycle, denoted as HBdiff = HBbefore - HBafter, has been included.

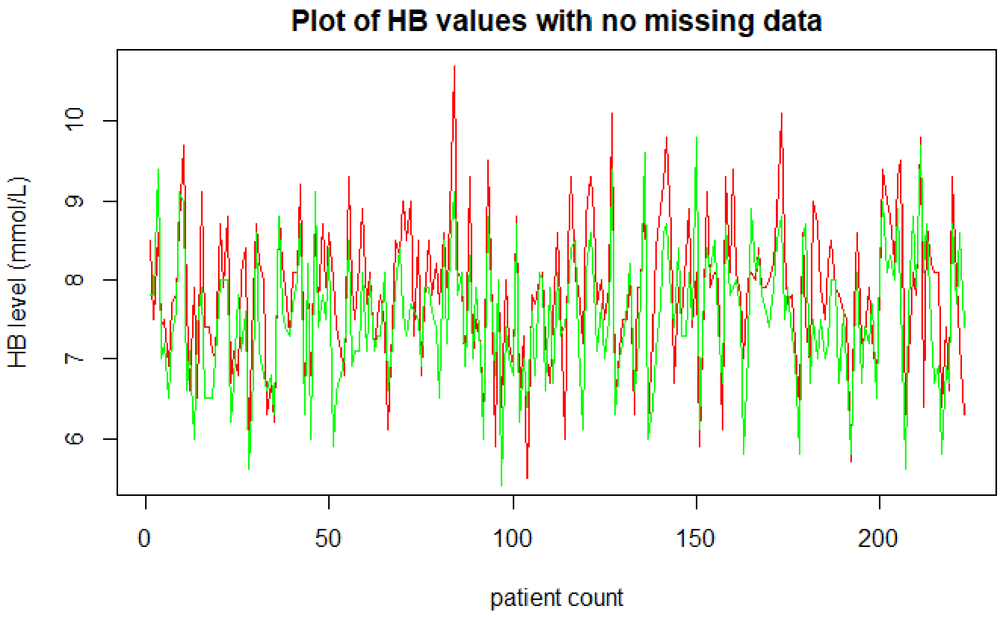


Figure B.52: Hemoglobin values (mmol/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=223.

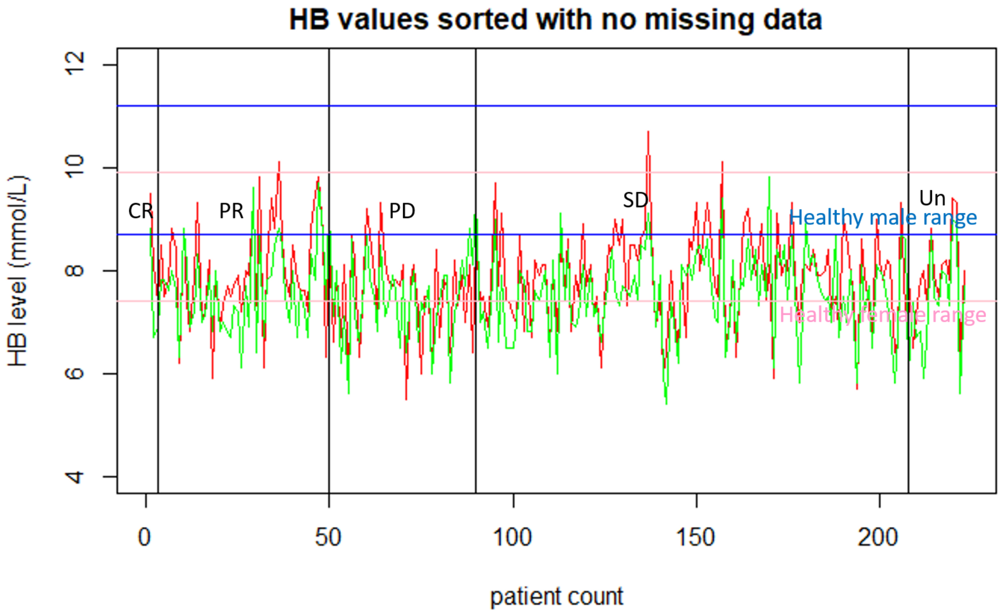


Figure B.53: Hemoglobin values before and after the first chemotherapy cycle sorted by final response with no missing values, n=223. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=46), Progressive Disease (PD, n= 40), Stable Disease (SD, n=118), Unknown (Un, n=16). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Blue = healthy range for males, Pink = healthy range for females.

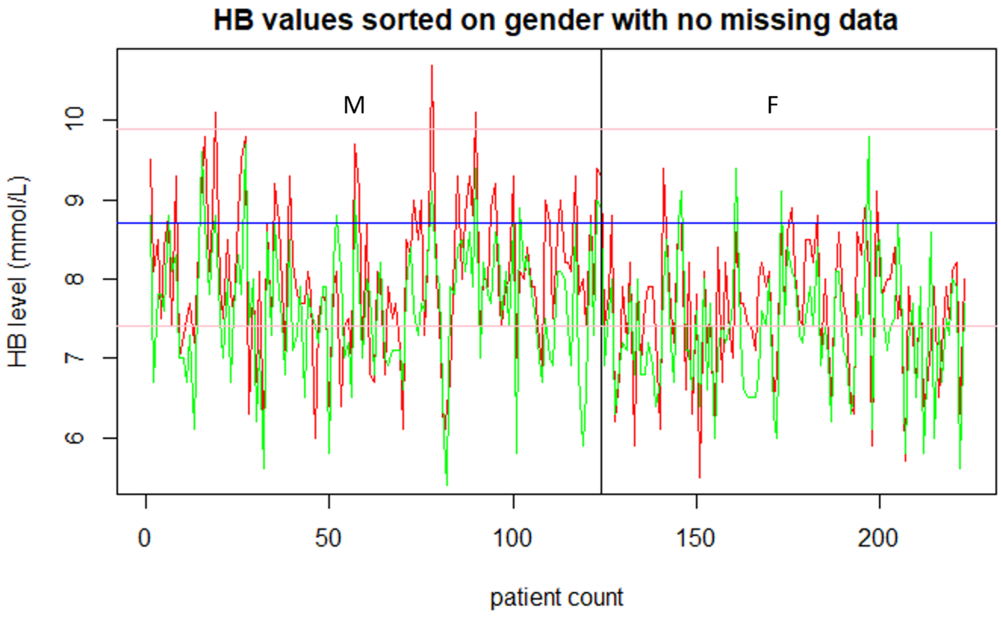


Figure B.54: Hemoglobin values (mmol/L) before and after the first chemotherapy cycle sorted by gender with no missing values, n=223 (124 male, 99 female).

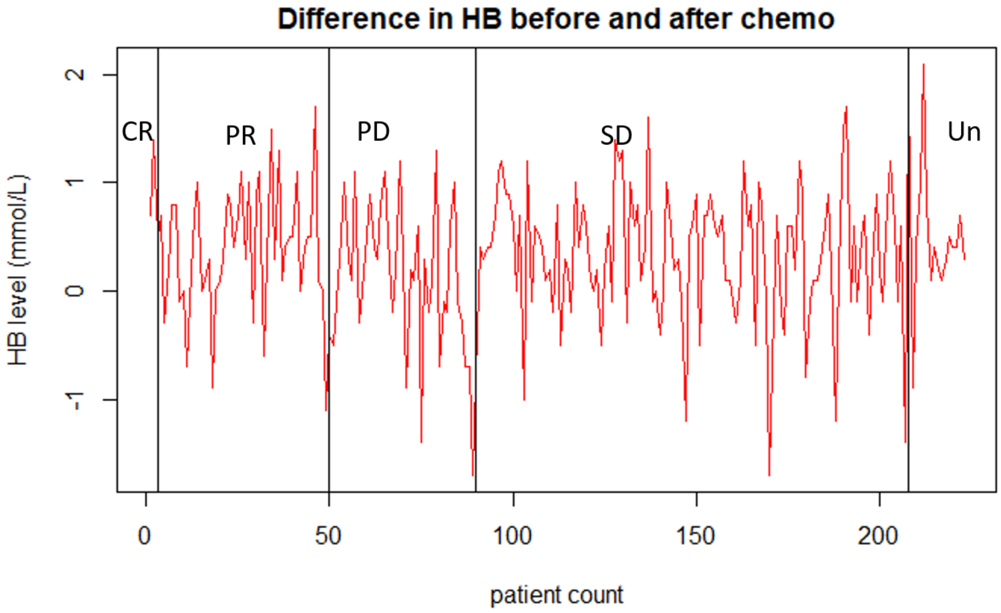


Figure B.55: Difference in hemoglobin values (mmol/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=223. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=46), Progressive Disease (PD, n= 40), Stable Disease (SD, n=118), Unknown (Un, n=16).

B.1.3.2 Thrombocytes TB

Thrombocytes, also known as platelets, are tiny disc-shaped cells that are found in the blood and spleen. They play a crucial role in the formation of blood clots to slow or stop bleeding and promote wound healing. According to medical doctors a normal number of platelets in the blood is about 150-400 $\times 10^9$/L blood. In the present dataset, a patient with ID (9065PP20005) had reported TB value of 4.3 $\times 10^9/L$ blood which is corrected to 43 $\times 10^9/L$ as a more realistic value. In Table B.9 summary statistics of thrombocyte counts in the patient cohort is presented. The results show a tenfold increase in the minimum value after chemotherapy, whereas the first quartile value, median, mean, third quartile, and maximum value have all decreased. The mean thrombocyte count falls within the healthy range of 150-450 as well as the first and third quartile values. (as the first till third quartile values are all in the range). Overall, there is a decrease in the number of platelets after the first chemotherapy session. In Figure B.56 various plots of the TB values before the first chemotherapy are presented. The results indicate that the average values are consistent across all final responses, with the caveat that the number of observations in the complete response group is only three. Therefore, any conclusions that could be drawn are invalid for the complete response group due to the small sample size. This holds true for the further analyses done on the other markers measured in the blood as well. Moreover, the histogram of the platelet values does show deviations from normality, which is supported by the normality tests with p-values given in Table B.10 and confirmed by the QQplots in Figure B.58. The same applies to the thrombocyte count after the first chemotherapy session.

To continue, the dataset is cleaned, whereby all missing values (NA) have been removed. The resulting dataset comprises of 224 observations, with 123 being male and 101 being female. Regarding the final response, 3 individuals had a complete response (CR), 47 had a partial response (PR), 40 had progressive disease (PD), while 118 had stable disease (SD). There were 16 cases where the final response was unknown. A summary of these values is presented in Table B.11. These summary values show close resemblance to those in Table 2 (see Table B.9). Additionally, Figure B.59 illustrates a plot of the values before and after chemotherapy. It is apparent from the plot that the overall trend indicates a lower value of TB count after chemotherapy, with the green graph averaging below the red graph. Figure B.60 presents the sorted values of the patients based on their final response outcome, with the healthy range plotted between the pink horizontal lines. An interesting observation is that the most significant spikes occur for patients with stable and progressive diseases. In contrast, most partial responders remain within the healthy range.

Thrombocyte value	*Before 1st Chemo*	*After 1st Chemo*
Min	43.0	43.0
1st Quartile	225	159
Median	274	202
Mean	280	216
3rd Quartile	325	252
Max	709	630
NA	7	13

Table B.9: Summary statistics values of the thrombocyte count of the entire dataset in $10^9/L$, n=247.

TB value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.954	0.899
SW-test pvalue	8.24e-7	1.87e-11
KS test pvalue	Invalid	Invalid

Table B.10: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the thrombocyte count, n=247. The KS-test is invalid due to the presence of ties in the data.

TB value	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	4.3	43.0	-304.0
1st Quartile	224.2	159.0	26.8
Median	270.5	202.5	64.0
Mean	278.6	216.5	62.1
3rd Quartile	325.0	251.2	106.8
Max	709	630	467

Table B.11: Summary statistics of the thrombocyte count in $10^9/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=224.

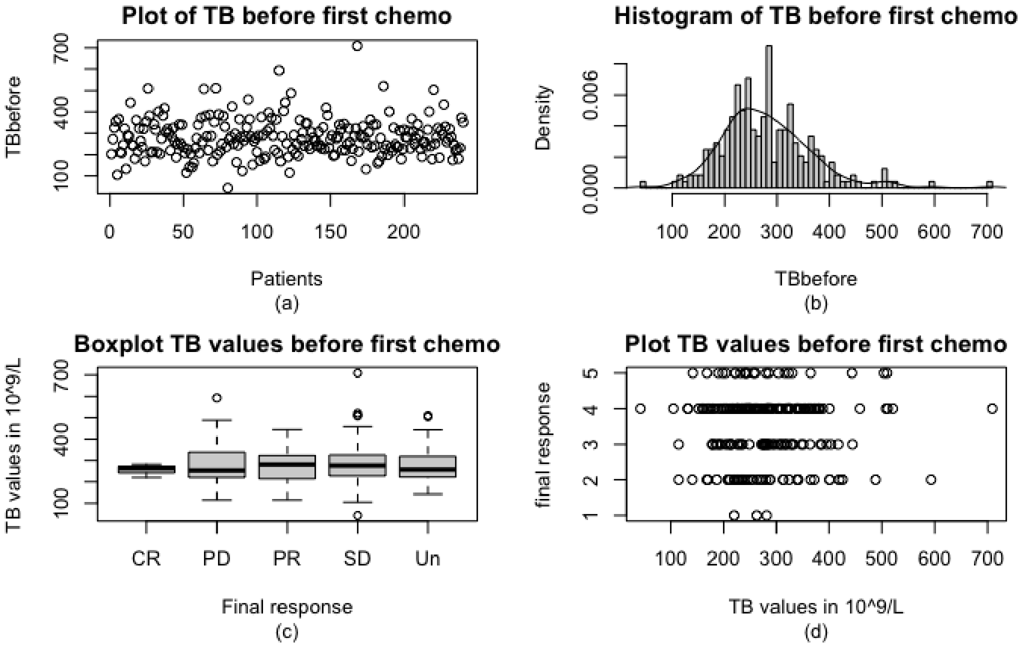


Figure B.56: Thrombocytes distribution before the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of thrombocyte count before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of thrombocyte count before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

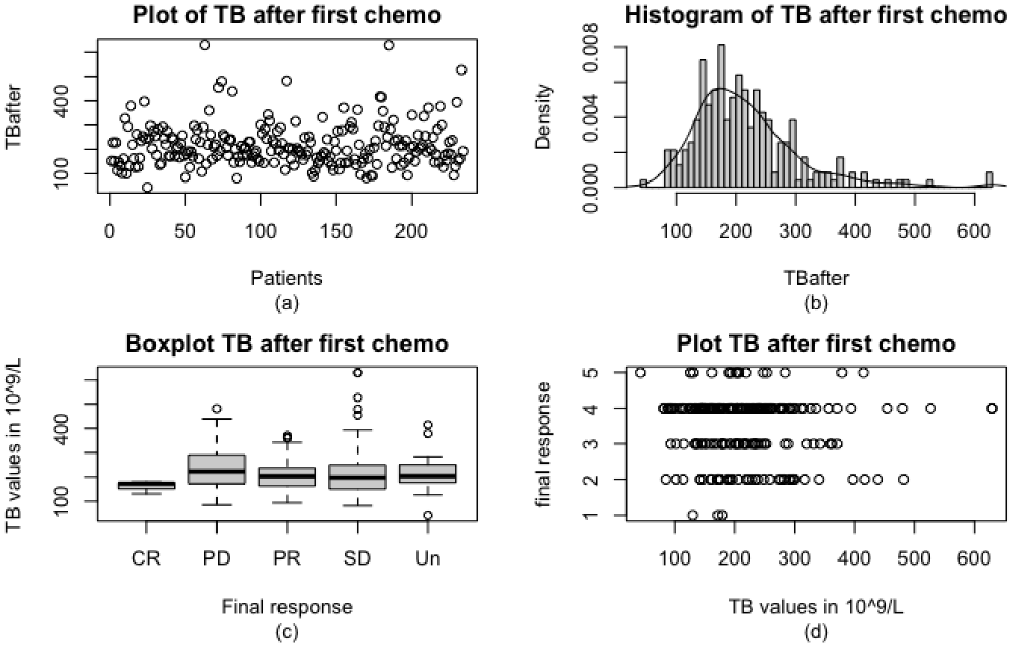


Figure B.57: Thrombocytes distribution after the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of thrombocyte count after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of thrombocyte count after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

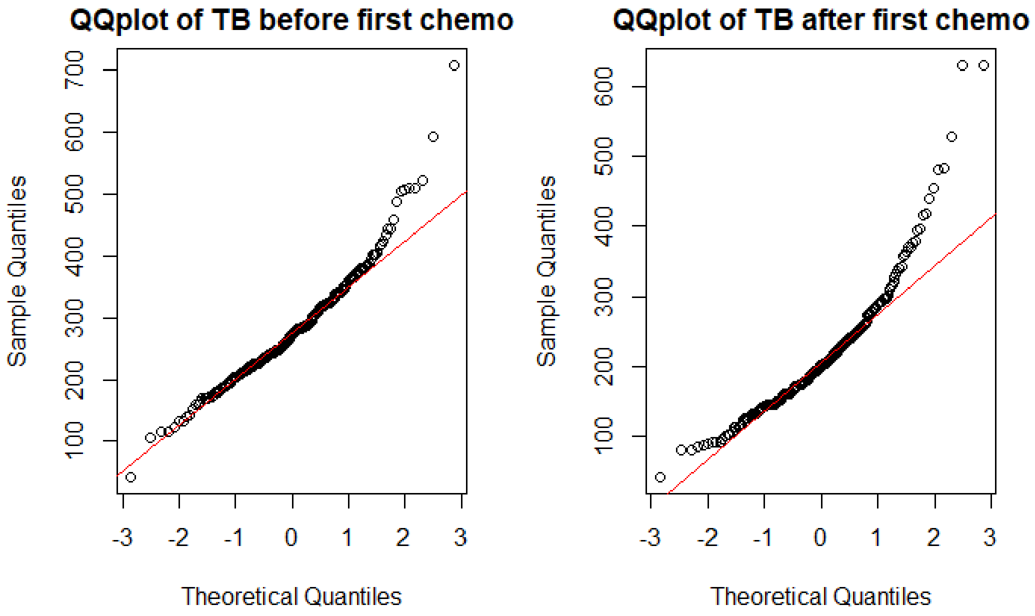


Figure B.58: QQplot of the thrombocyte count before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

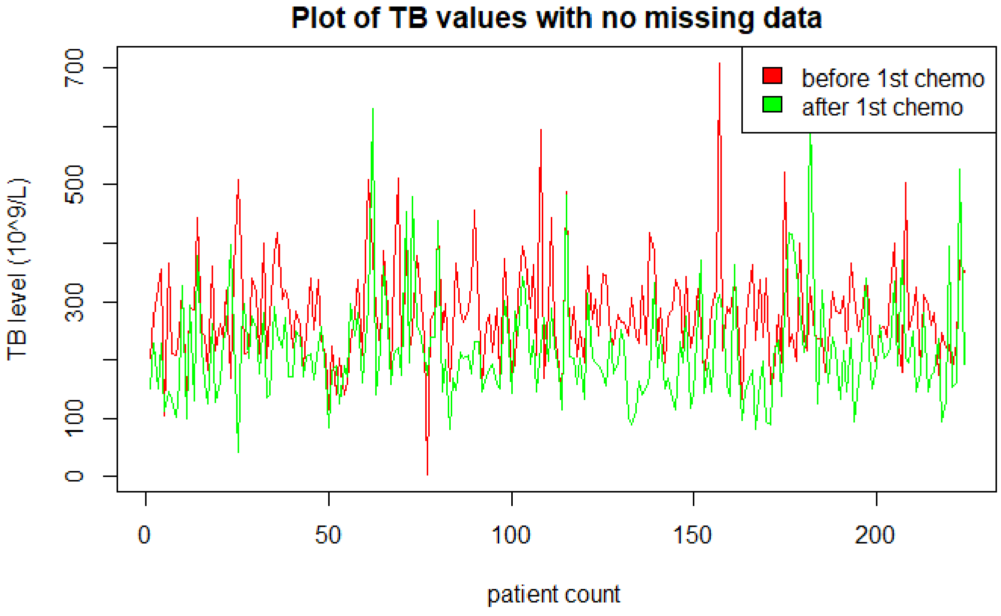


Figure B.59: Thrombocyte count ($10^9/L$) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=224.

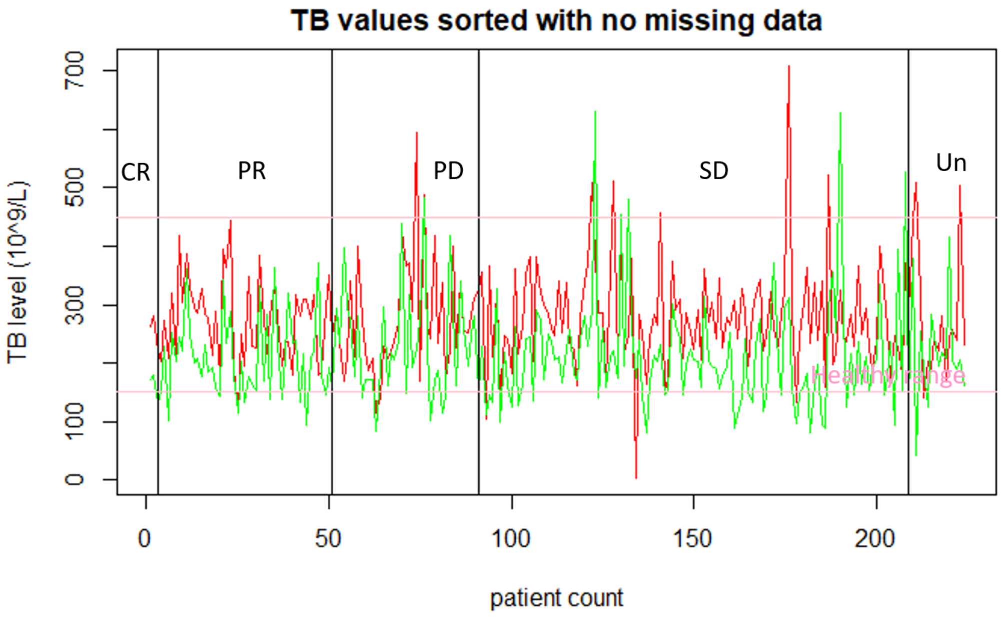


Figure B.60: Thrombocyte count ($10^9/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=224. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 40), Stable Disease (SD, n=118), Unknown (Un, n=16). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

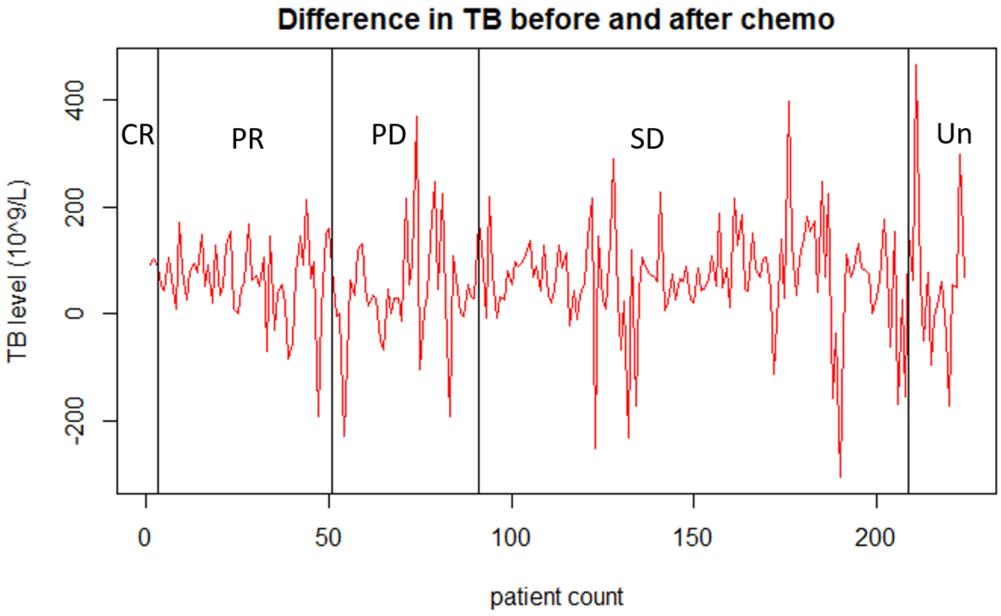


Figure B.61: Difference in thrombocyte count ($10^9/L$ between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=224. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 40), Stable Disease (SD, n=118), Unknown (Un, n=16).

B.1.3.3 Leukocytes LK

Leukocytes, also known as white blood cells, are a vital component of the human immune system, as they defend the body against infections and diseases. Leukocytes are produced in the bone marrow and are distributed throughout the blood and lymphatic tissues. The three main types of leukocytes include granulocytes (neutrophils, eosinophils, and basophils), monocytes, and lymphocytes (B- and T-cells). According to medical experts from the Erasmus Medical Centre Rotterdam an individual's healthy range of leukocytes is typically between 4 and 10 $\times 10^9/L$ blood.

In this data analysis, to find if there is indeed any link between leukocyte count and response to chemotherapy, first all the missing values are removed from the dataset. The resultant dataset contains 226 observations of which 125 are male and 101 are female. The final response has 3 patients with CR, 47 with PR, 40 with PD, 120 with SD and 16 were unknown. Summary statistics for this dataset are presented in Table B.14, indicating only slight changes in the values when compared to those in Table B.12. Remarkably, the leukocyte count demonstrated a significant increase after chemotherapy, as illustrated in Figure B.65. This phenomenon may be attributed to body's ability to stimulate the bone marrow to produce more white blood cells as a natural response to the stress of the chemotherapy treatment. Specifically, chemotherapy can stimulate the bone marrow to produce more white blood cells to help fight infections and support the immune system.

One type of white blood cell that may increase after chemotherapy is the neutrophil (considered in the section below), which plays a critical role in fighting bacterial infections. However, the increase in white blood cells can also be a side effect of the drugs used in chemotherapy. Some chemotherapy drugs can stimulate the production of white blood cells directly, while others can damage the bone marrow, leading to an increase in the release of white blood cells into the bloodstream. It is important to note that not all chemotherapy drugs cause an increase in leukocytes, and the extent and duration of the increase can vary depending on the specific drugs and dosage used, as well as the individual's overall health and response to treatment. After sorting the patients based on their final outcome, as shown in Figure B.66, no noticeable pattern emerges, except for the observation that the highest LK values before chemotherapy are found in the stable disease group.

Leukocyte value	*Before 1st Chemo*	*After 1st Chemo*
Min	2.5	1.2
1st Quartile	6.7	6.7
Median	8.0	12.1
Mean	8.2	13.5
3rd Quartile	9.3	18.2
Max	16.6	42.7
NA	6	12

Table B.12: Summary statistics values of the leukocyte count of the entire dataset in $10^9/L$, n=247.

LK value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.959	2.06e-6
SW-test pvalue	0.936	1.34e-8
KS test pvalue	Invalid	Invalid

Table B.13: Results of the Shapiro-Wilk test and the Kolmogorov Smirnov test for assessing the normality of the leukocyte count, n=247. The KS-test is invalid due to the presence of ties in the data.

LK value	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	2.5	1.2	-32.4
1st Quartile	6.7	6.6	-9.6
Median	8.0	11.9	-3.7
Mean	8.2	13.3	-5.1
3rd Quartile	9.2	18.2	1.1
Max	16.6	42.7	9.5

Table B.14: Summary statistics of the leukocyte count in $10^9/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=226.

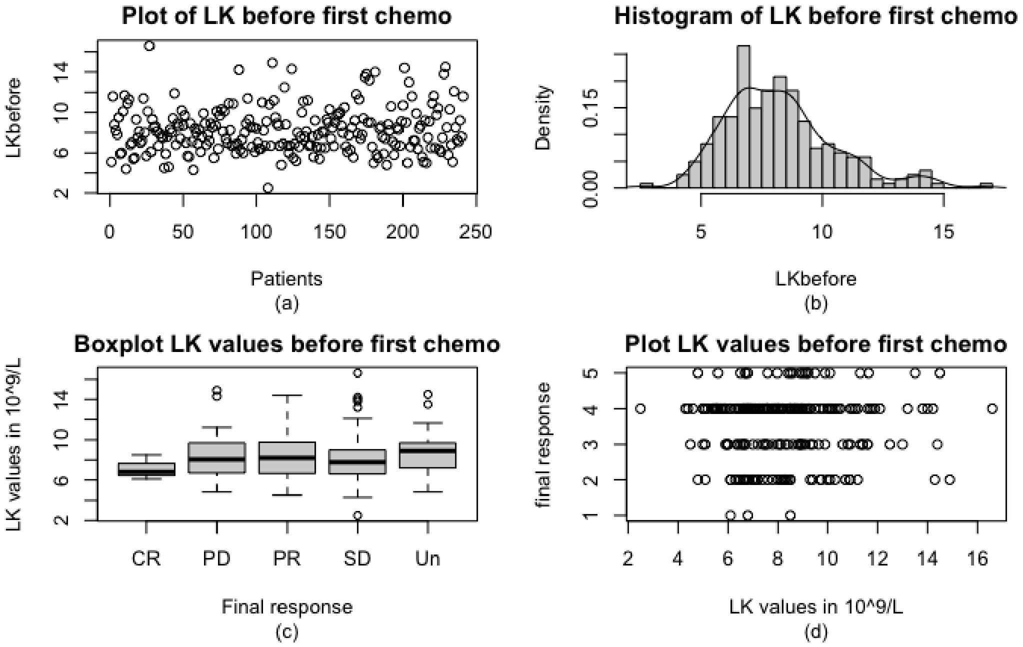


Figure B.62: Leukocyte distribution before the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of leukocyte count before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of leukocyte count before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

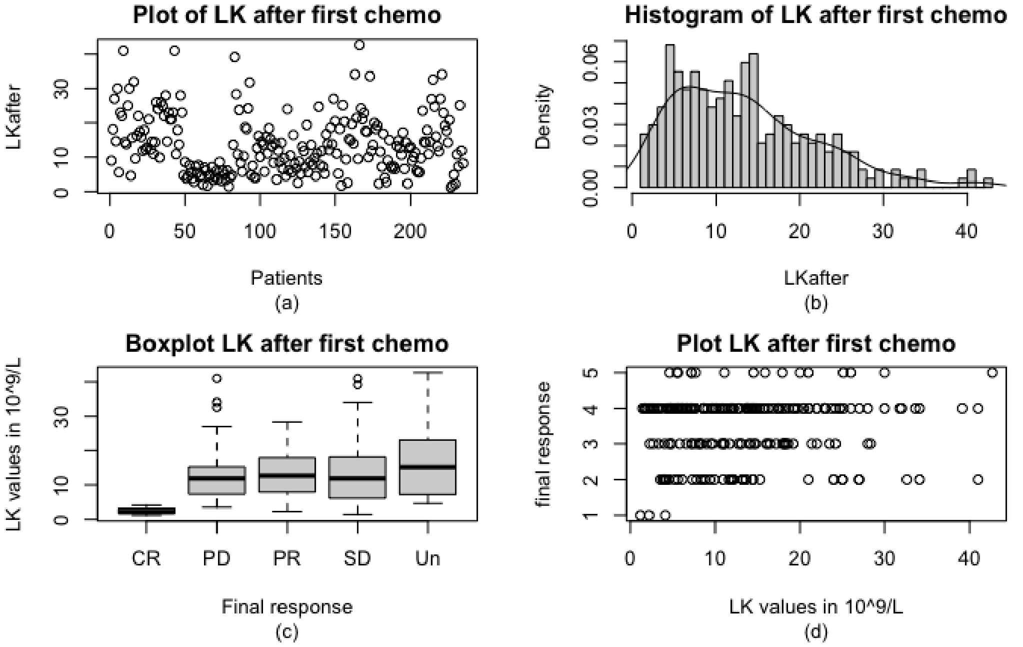


Figure B.63: Leukocyte distribution after the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of leukocyte count after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of leukocyte count after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

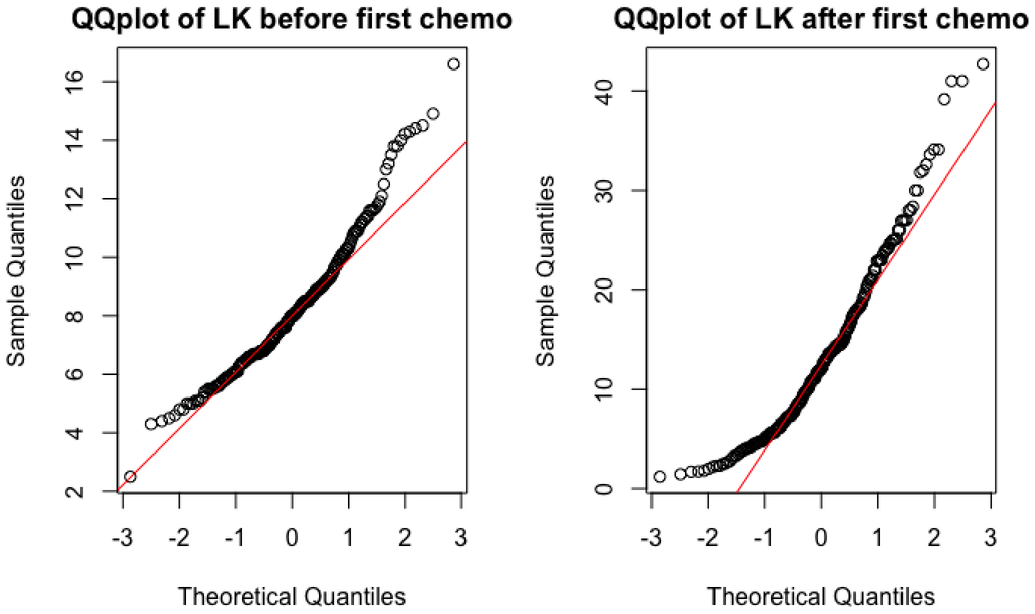


Figure B.64: QQplot of the leukocyte values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

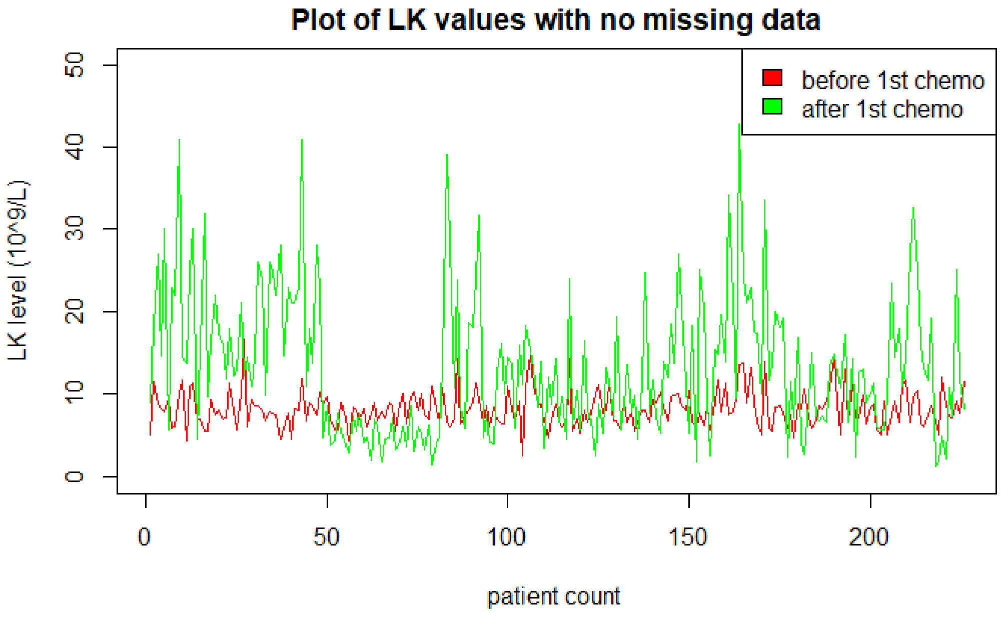


Figure B.65: Leukocyte count $(10^9/L)$ before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=226.

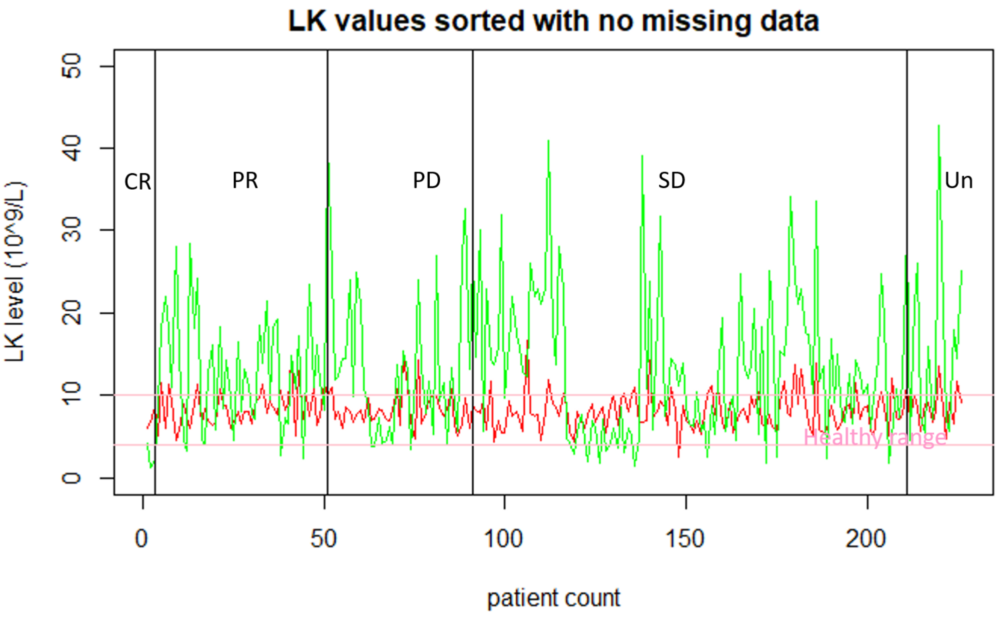


Figure B.66: Leukocyte count ($10^9/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=226. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 40), Stable Disease (SD, n=120), Unknown (Un, n=16). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

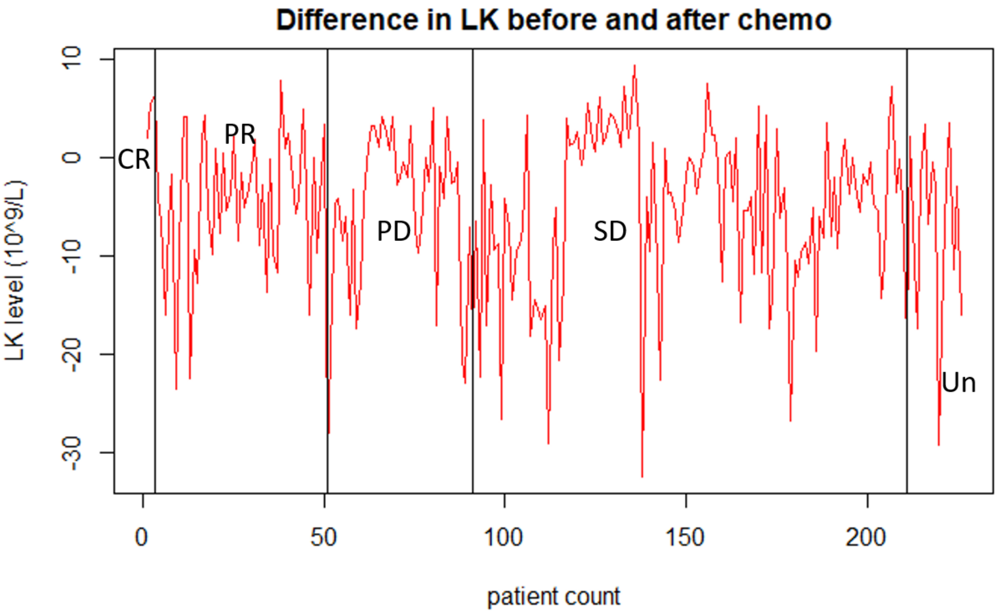


Figure B.67: Difference in leukocyte count ($10^9/L$) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=226. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 40), Stable Disease (SD, n=120), Unknown (Un, n=16).

B.1.3.4 Neutrophils NP

Neutrophils are type of leukocyte, or white blood cell, that assist in the body's immune response against infections and injuries. They are classified as a subset of granulocytes, along with eosinophils and basophil cells. Among all types of leukocytes, neutrophils are the most abundant in the human body. A healthy range of Neutrophils is 1.5-7.5 $\times 10^9$/L blood. Studies have shown that neutrophils play a crucial role in the progression and metastasis of PDAC, as they are recruited to the tumor micro-environment and promote tumor growth and invasion by releasing cytokines, proteases, and other pro-inflammatory factors. One study by Iwai et al. [83] studying the effect of NLR in unresectable pancreatic cancer patients, found that elevated neutrophil-to-lymphocyte ratio (NLR) was independently associated with poor prognosis in patients with PDAC. A high NLR was associated with the metastatic stage and worse performance status. Another study by Yang et al. [84] also reported that high NLR in the peripheral blood is a poor prognostic factor for patients with pancreatic cancer as elevated NLR was associated with poor OS in patients with PDAC receiving either mixed treatment or chemotherapy. Overall, these findings suggest that neutrophil count and function are important factors in the pathogenesis and clinical outcome of PDAC, and highlight the potential of targeting neutrophils as a novel therapeutic strategy for this deadly disease. However, to see if this is seen back in the current data, the following analysis is performed.

The summary statistics in Table B.15 indicate that the first to third quartile values of neutrophil levels were within the healthy range prior to chemotherapy. However, following chemotherapy, all values increased, with the median value exceeding the healthy range at 8.1 $\times 10^9/L$ (which was previously 7.5 $\times 10^9/L$). Additionally, the distribution of neutrophil values before and after chemotherapy, as depicted in Figure B.68 en Figure B.69, respectively, demonstrate non-normality, which is supported by the results of the two tests outlined in Table B.16. Consider the complete dataset on neutrophils without missing values. The data consists of 199 observations, with 110 male and 89 female participants. Among these participants, 3 exhibited complete response (CR), 40 exhibited partial response (PR), 35 exhibited progressive disease (PD), 110 exhibited stable disease (SD), and 11 exhibited an unknown final response. Analysis of the dataset, as depicted in Figure B.71 and Figure B.72, reveals a clear and substantial increase in neutrophil count following chemotherapy, resulting in a shift to unhealthy levels. Prior to chemotherapy, the majority of neutrophil counts fell within the healthy range. Notably, a significant increase in neutrophil count was observed among patients with stable disease as their final response, similar to the pattern observed with leukocyte count. These findings suggest that chemotherapy may have a significant impact on neutrophil count, and that changes in neutrophil count may be associated with treatment response. However, the increase in neutrophil count may also be caused by medication given to the patients during treatment.

Neutrophils Value	*Before 1st Chemo*	*After 1st Chemo*
Min	1.2	0.1
1st Quartile	4.2	3.6
Median	5.1	8.2
Mean	5.4	9.6
3rd Quartile	6.4	13.4
Max	13.2	36.4
NA	28	24

Table B.15: Summary statistics of the Neutrophil count in 10^9/L blood in entire data set, n=247.

NP value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.952	0.918
SW-test pvalue	1.25e-6	9e-10
KS test pvalue	Invalid	Invalid

Table B.16: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the neutrophil values, n=247. The KS-test is invalid due to the presence of ties in the data.

NP value	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	1.2	0.1	-31.1
1st Quartile	4.2	3.5	-7.0
Median	5.1	7.7	-2.3
Mean	5.4	9.1	-3.7
3rd Quartile	6.3	12.3	1.6
Max	13.2	36.1	8.5

Table B.17: Summary statistics of the neutrophil count in $10^9/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=199.

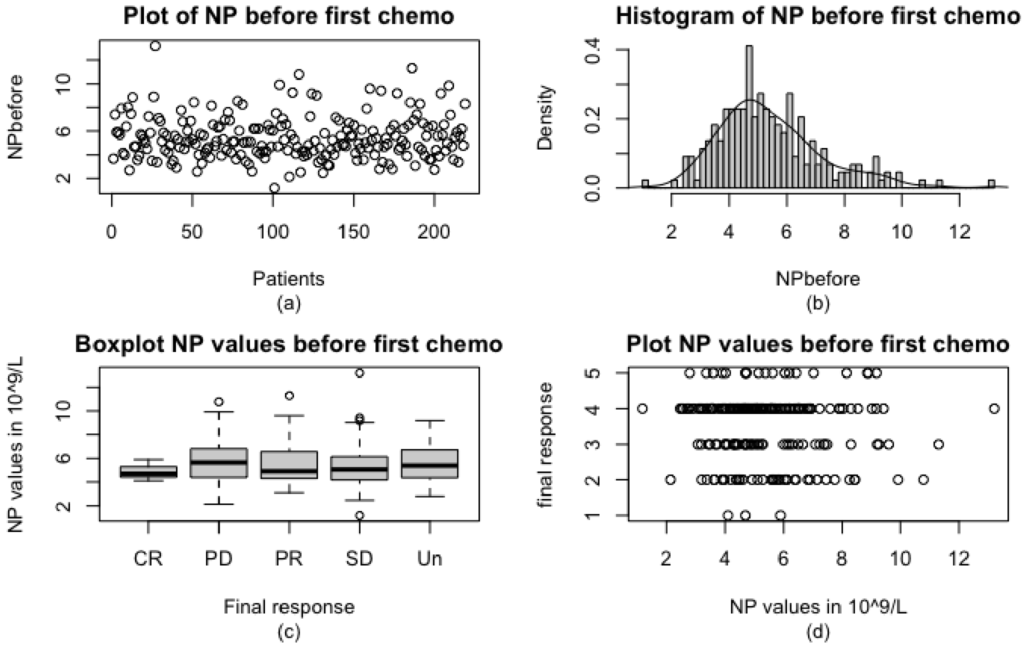


Figure B.68: Neutrophil distribution before the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of neutrophil count before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of neutrophil count before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

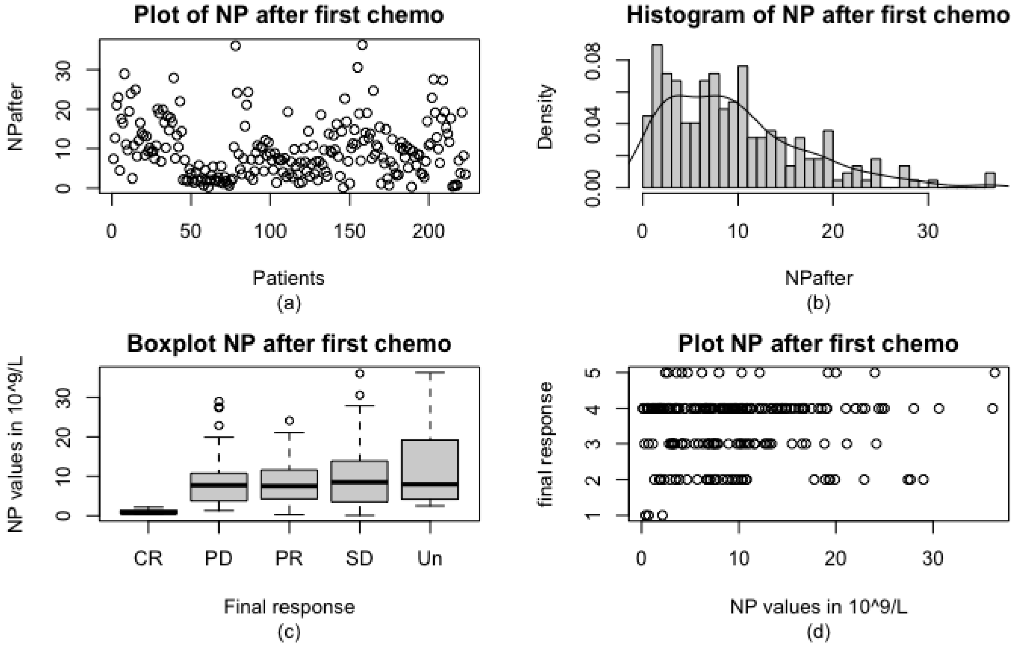


Figure B.69: Neutrophil distribution after the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of neutrophil count after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of neutrophil count after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

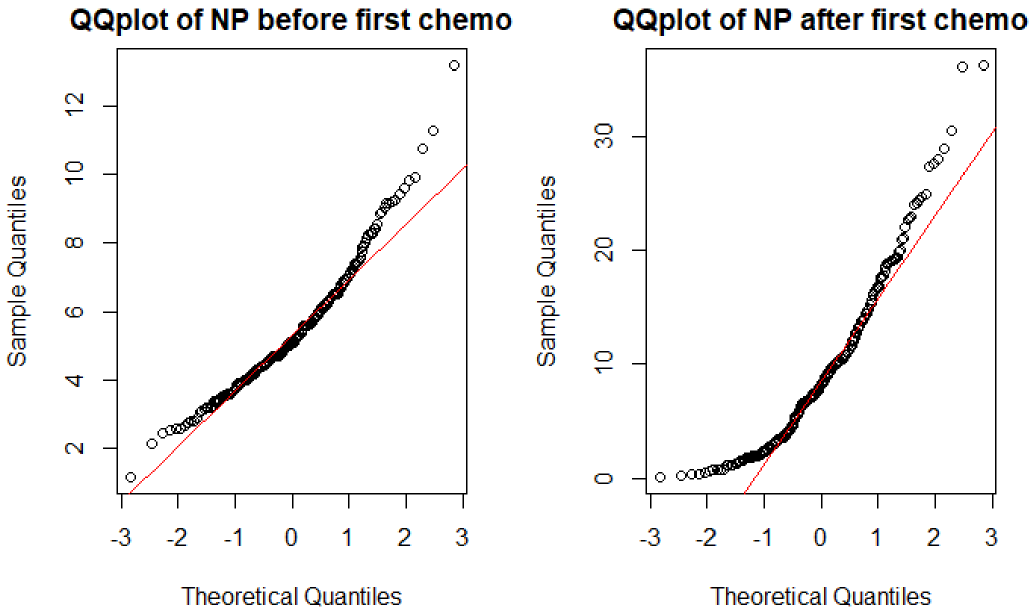


Figure B.70: QQplot of the neutrophil count before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

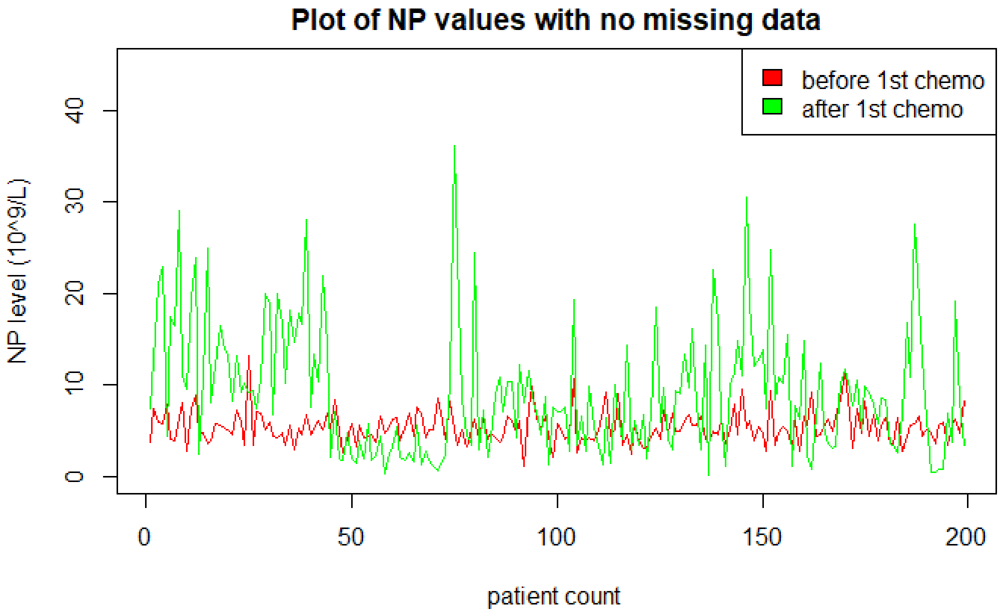


Figure B.71: Neutrophil count ($10^9/L$) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=199.

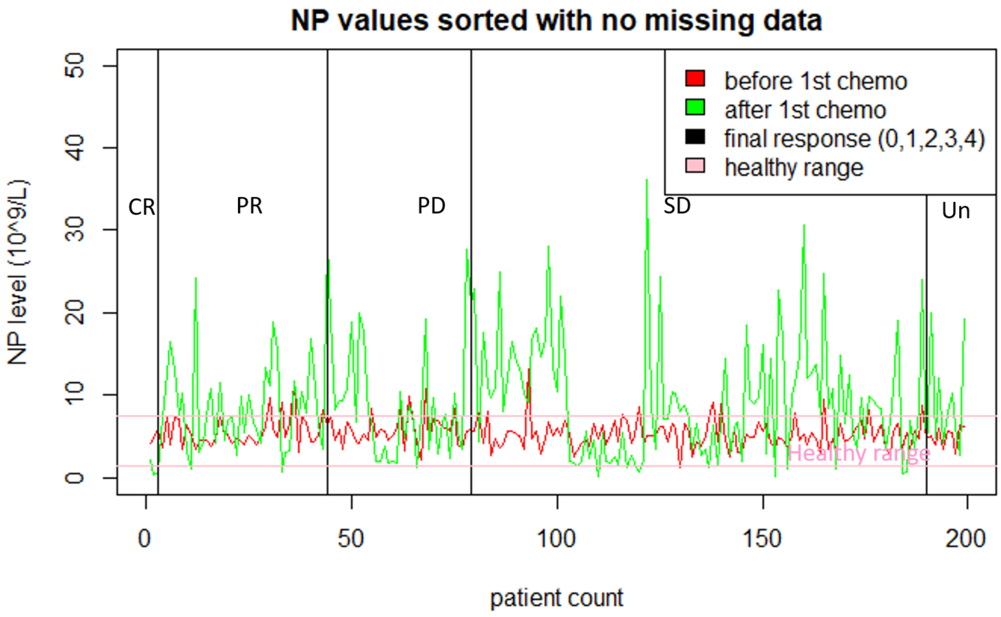


Figure B.72: Neutrophil count ($10^9/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=199. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=40), Progressive Disease (PD, n= 35), Stable Disease (SD, n=110), Unknown (Un, n=11). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

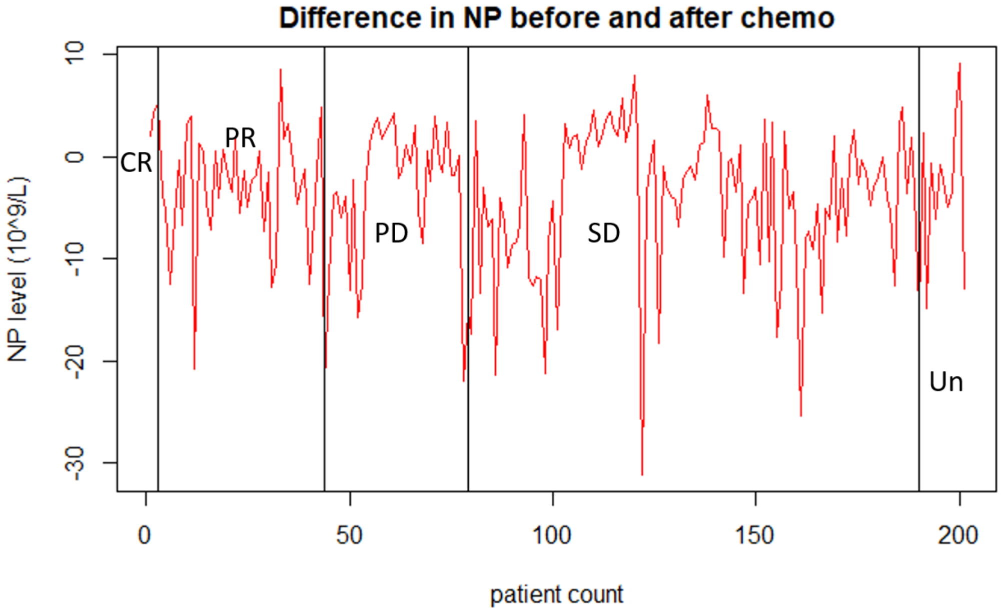


Figure B.73: Difference in Neutrophil count ($10^9/L$) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=199. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=40), Progressive Disease (PD, n= 35), Stable Disease (SD, n=110), Unknown (Un, n=11).

B.1.3.5 Lymphocytes LC

Similar to neutrophils, lymphocytes are a crucial type of white blood cell that play an important role in the body's immune system's defense mechanism against infections and diseases. These cells can be broadly categorized into two subtypes: T-lymphocytes and B-lymphocytes. T-lymphocytes play a vital role in regulating the immune response and directly attacking and eliminating infected cells and cancerous cells. On the contrary, B-lymphocytes produce antibodies, which are specialized proteins that identify and target viruses, bacteria, and other foreign antigens. The summary statistics of the lymphocyte data in $10^9/L$ blood for the complete dataset (n=247) including missing values (NA) are presented in Table B.18. The LC values before and after chemotherapy are illustrated in Figure B.74 and Figure B.75, respectively. The histograms in these figures indicate a relatively normal distribution of the LC count before and after the initial chemotherapy session, with significant outliers on the right side. The boxplot and scatter plots show no distinct patterns or trends with respect to final responses. The QQplots in Figure B.76 and the values in Table B.19 suggest that the LC values are not normally distributed, primarily due to the outliers causing deviations from normality in the tail of the plots.

The lymphocyte data set is analyzed after removing the missing values. The remaining dataset contains n=122 observations (64 male, 58 female). Among these, 2 individuals show a complete response (CR), 22 show a partial response (PR), 24 show progressive disease (PD), 66 show stable disease (SD), and 8 observations were categorized as Unknown in terms of final response. The summary statistics of this dataset are presented in Table B.20, which reveals that the healthy range of values, between 1 and 4 $\times 10^9/L$, is achieved between the first and third quartile. Figure B.77 presents a plot of the lymphocyte values, and Figure B.78 groups them by final outcome. The majority of peaks are observed in the group with stable disease as the final outcome. Additionally, Figure B.79 displays a plot of the difference between the lymphocyte count values before and after chemotherapy sorted by their respective final response outcomes.

Lymphocytes Value	*Before 1st Chemo*	*After 1st Chemo*
Min	0.4	0.2
1st Quartile	1.3	1.3
Median	1.6	1.9
Mean	1.8	2.1
3rd Quartile	2.2	2.6
Max	9.2	11.0
NA	75	100

Table B.18: Summary statistics values of the lymphocyte count of the entire dataset in $10^9/L$, n=247.

NP value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.775	0.796
SW-test pvalue	5.706e-15	4.963-13
KS test pvalue	Invalid	Invalid

Table B.19: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the lymphocyte count, n=247. The KS-test is invalid due to the presence of ties in the data.

Lymphocytes Value	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	0.4	0.2	-3.1
1st Quartile	1.3	1.2	-0.7
Median	1.6	1.9	-0.2
Mean	1.8	2.1	-0.3
3rd Quartile	2.1	2.6	0.2
Max	9.2	11.0	2.1

Table B.20: Summary statistics of the lymphocyte count in $10^9/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=122.

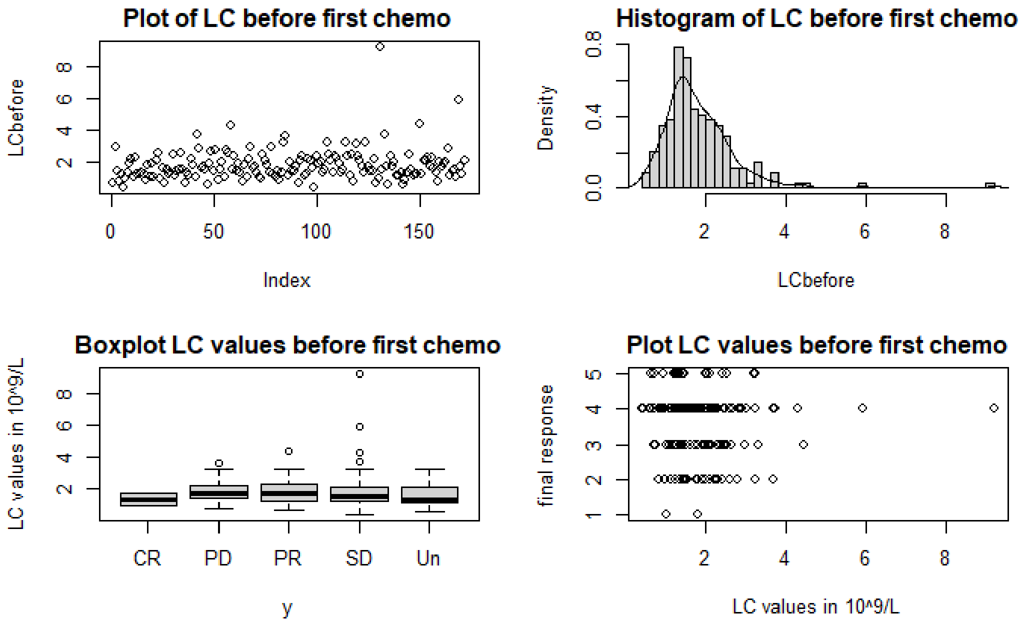


Figure B.74: Lymphocyte distribution before the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of lymphocyte count before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of lymphocyte count before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

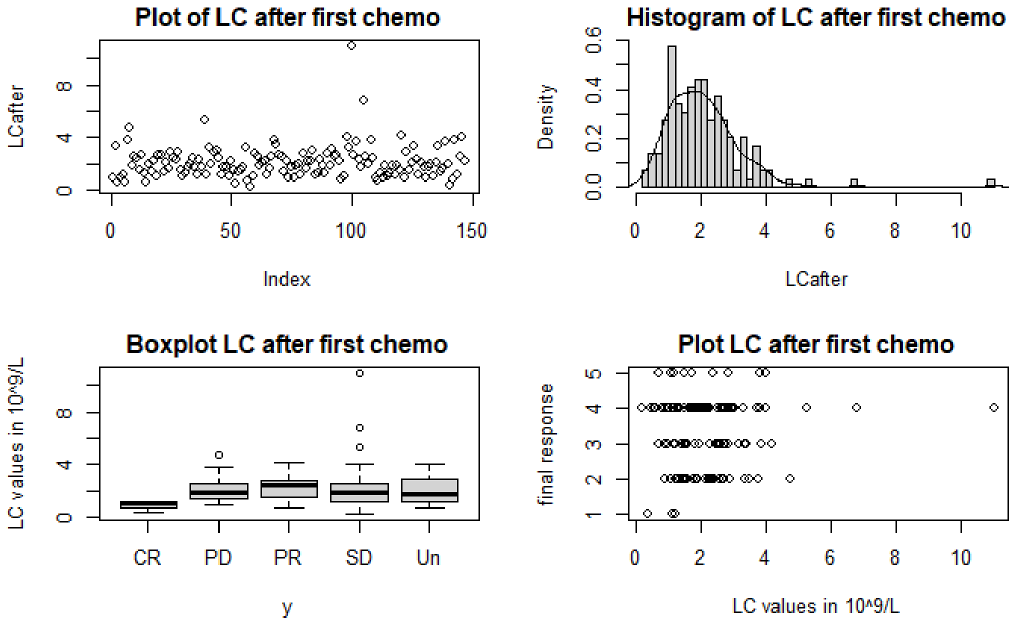


Figure B.75: Lymphocyte distribution after the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of lymphocyte values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of lymphocyte values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

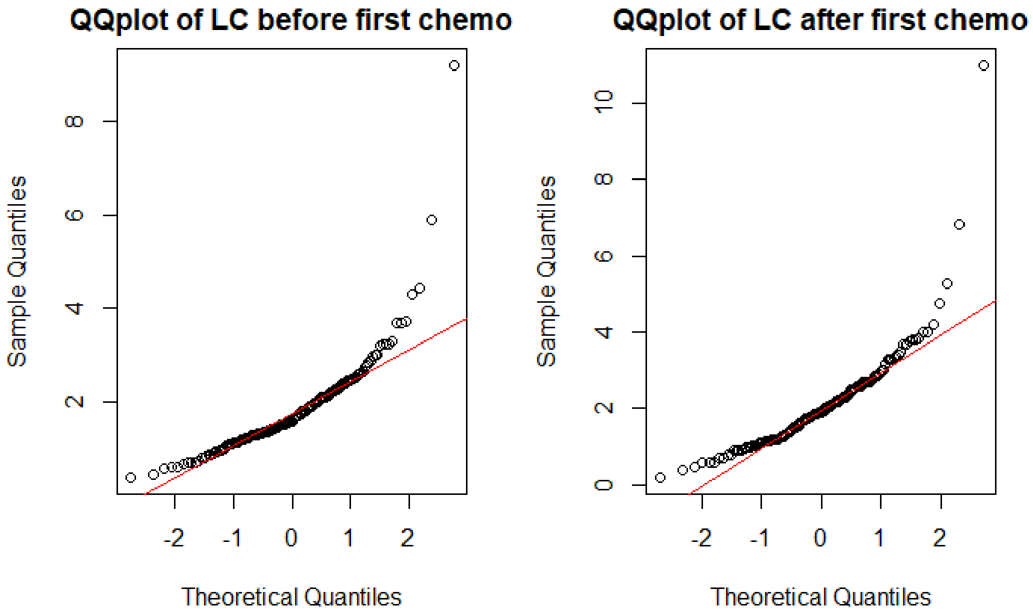


Figure B.76: QQplot of the lymphocyte count before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

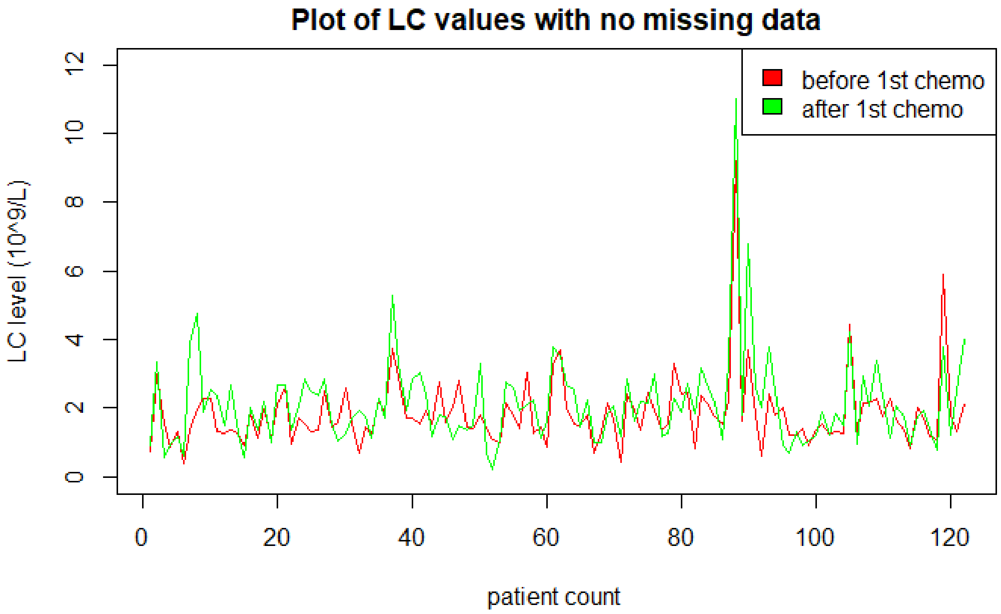


Figure B.77: Lymphocyte count ($10^9/L$) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=122.

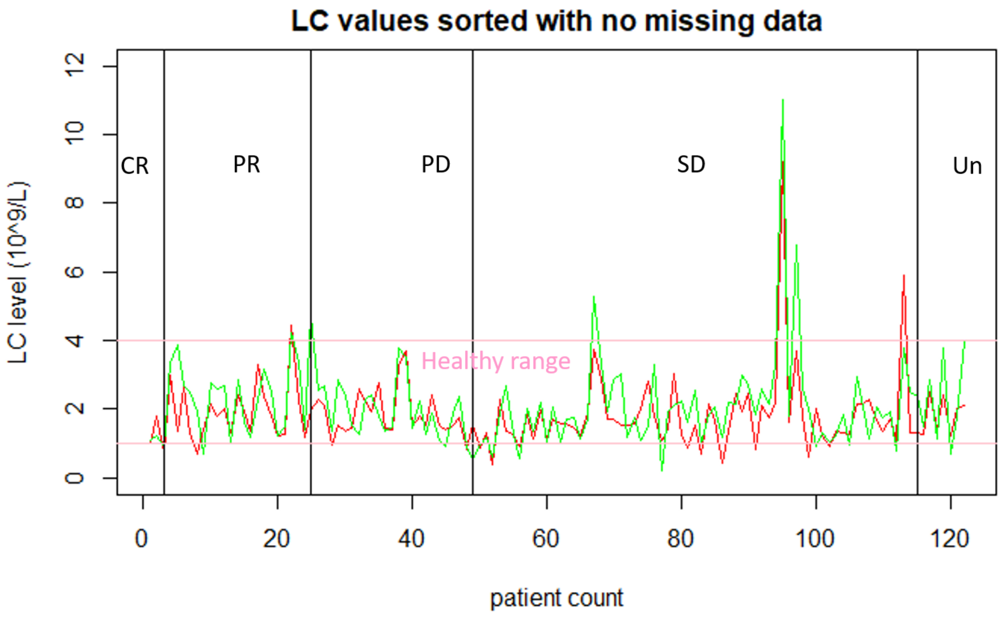


Figure B.78: Lymphocyte count ($10^9/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=122. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=22), Progressive Disease (PD, n= 24), Stable Disease (SD, n=66), Unknown (Un, n=8). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

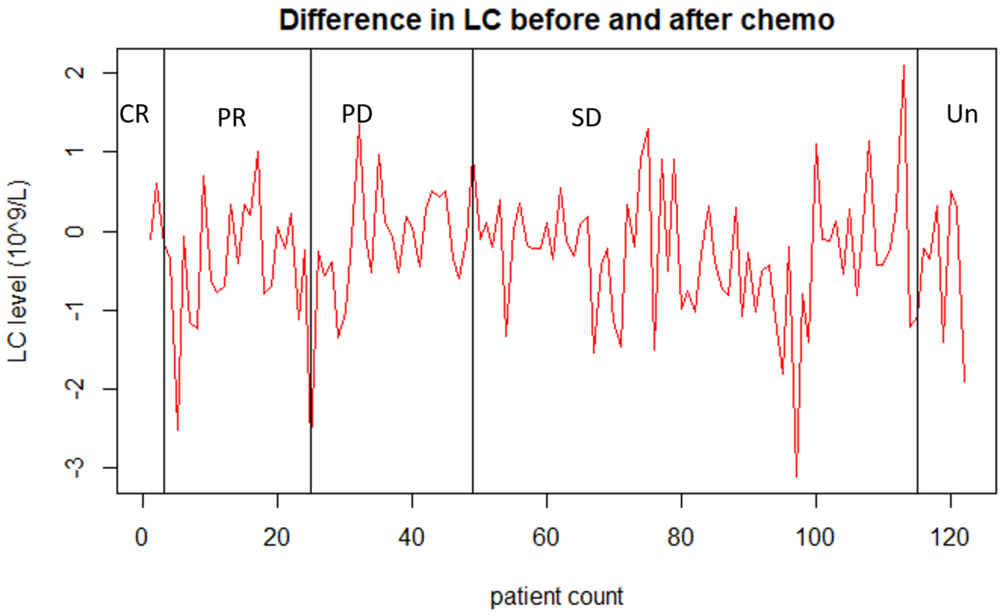


Figure B.79: Difference in lymphocyte count ($10^9/L$) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=122. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=22), Progressive Disease (PD, n= 24), Stable Disease (SD, n=66), Unknown (Un, n=8).

B.1.3.6 Creatinin CR

Creatine is a naturally occurring substance in the body that helps to maintain a continuous supply of energy to working muscles. During the process of energy production in working muscles, creatinin is produced as a byproduct. Normally, healthy kidneys filter creatinin out of the blood, and the level of creatinin in the blood can indicate how well the kidneys are functioning. The healthy range for creatinin levels is between 60-110 $\mu mol/L$ for males and between 50-100 $\mu mol/L$ for females.

The QQplots presented in Figure B.82 indicate that the majority of creatinin values follow a normal distribution, but exhibit some significant deviations in the tail values. This is consistent with the histograms displayed in Figure B.80 and Figure B.81, where it can be observed that most values conform to a normal distribution, albeit with a long right tail. The dataset with missing values comprises of n = 224 subjects (126 male, 98 female). The final responses show that 3 individuals exhibited complete response (CR), 47 partial response (PR), 42 progressive disease (PD), 118 stable disease (SD), and 14 had an unknown final response. Moreover, the summary statistics provided in Table B.23 indicate that the majority of values before and after the first chemotherapy session remain similar, except for the maximum value. The substantial increase in maximum values suggests a considerable decrease in kidney functionality following chemotherapy. However, this increase is not seen as an overall trend, but only in a couple of patients, who most likely have kidney failure or any other kidney related health condition.

Creatinin value	*Before 1st Chemo*	*After 1st Chemo*
Min	33.0	32.0
1st Quartile	59.0	59.0
Median	69.0	70.0
Mean	79.8	76.2
3rd Quartile	80.0	81.0
Max	222.0	952.0
NA	6	14

Table B.21: Summary statistics values of the creatinin values of the entire dataset in μmol/L, n=247.

CR value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.864	0.249
SW-test pvalue	8.307e-14	<2.2e-16
KS test pvalue	Invalid	Invalid

Table B.22: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the creatinin values, n=247. The KS-test is invalid due to the presence of ties in the data.

Creatinin values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	33.0	32.0	-843.0
1st Quartile	59.0	59.0	-6.0
Median	69.0	69.5	-1.0
Mean	70.6	76.0	-5.3
3rd Quartile	80.0	81.0	3.0
Max	222.0	952.0	61.0

Table B.23: Summary statistics of the creatinin levels in μmol/L before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=224.

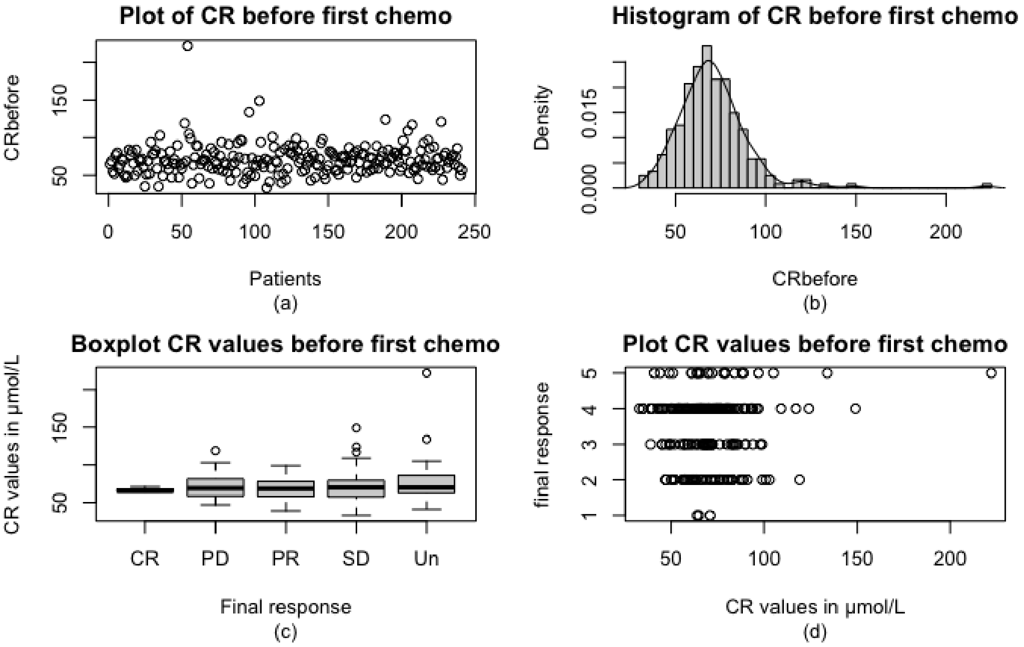


Figure B.80: Creatinin distribution before the first chemotherapy cycle (μmol/L) for the entire dataset: (a) scatterplot of creatinin values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of creatinin values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

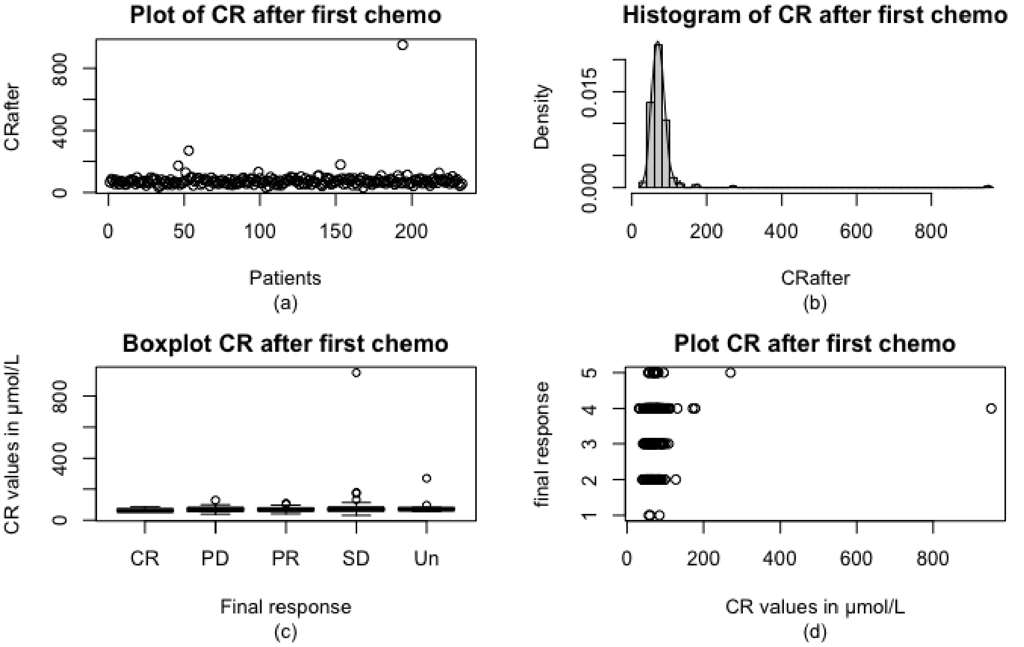


Figure B.81: Creatinin distribution after the first chemotherapy cycle (μmol/L) for the entire dataset: (a) scatterplot of creatinin values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of creatinin values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

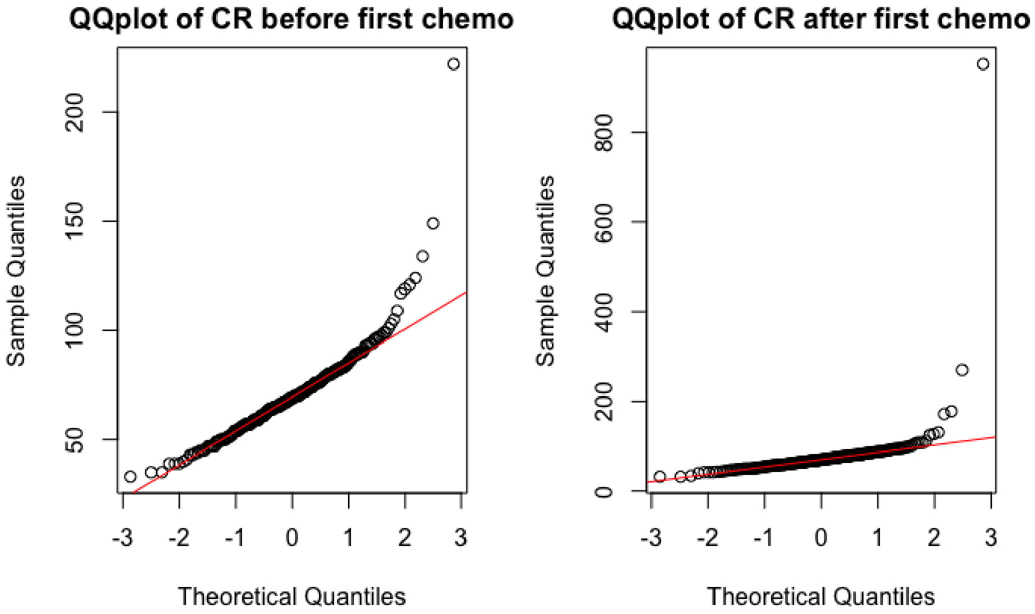


Figure B.82: QQplot of the creatinin values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

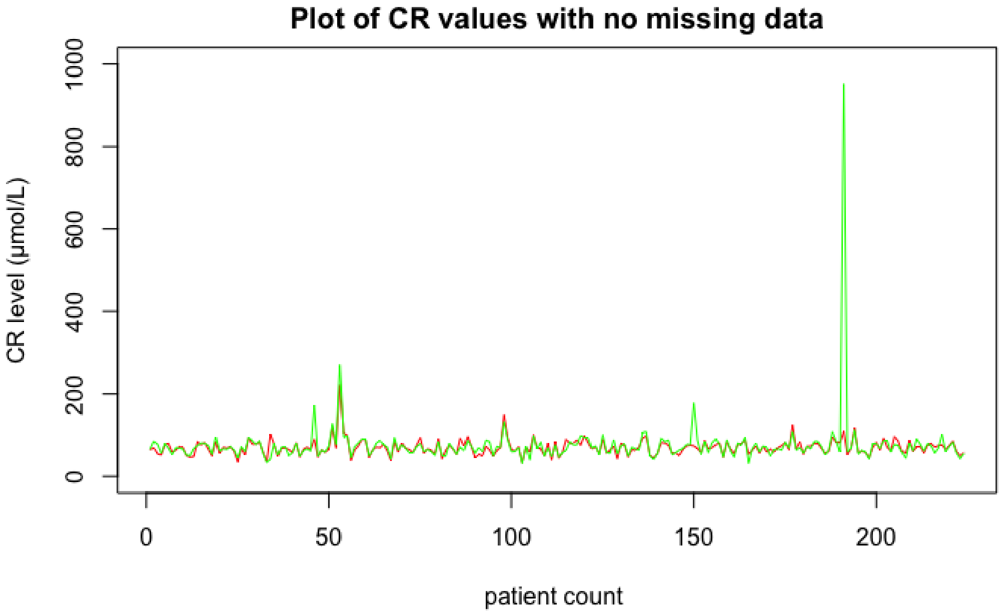


Figure B.83: Creatinin values ($\mu mol/L$) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=224.

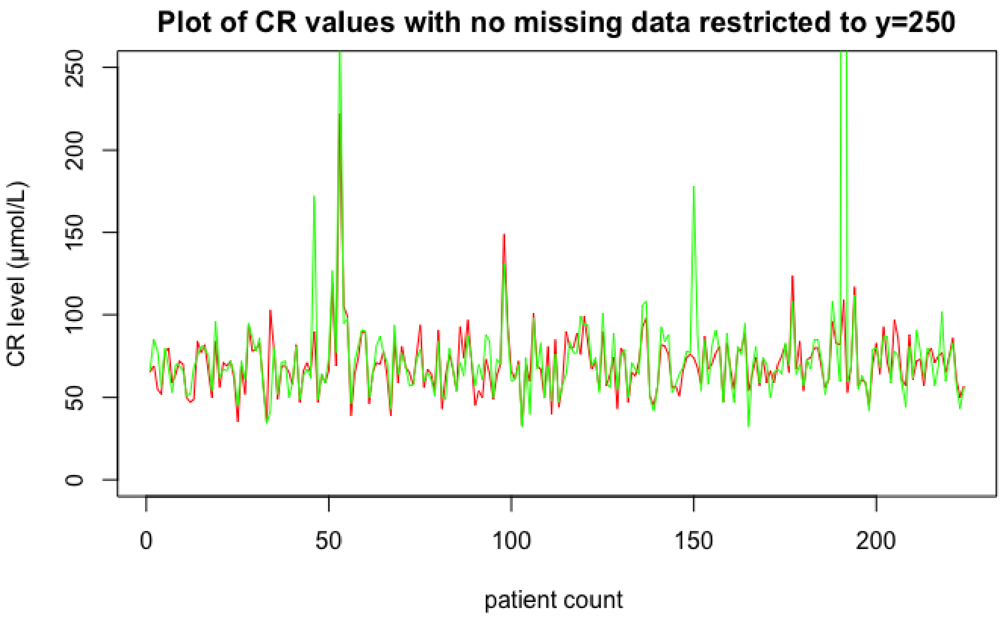


Figure B.84: Same plot as Figure B.83 but with the y-axis limited to maximum value of 250 $\mu mol/L$.

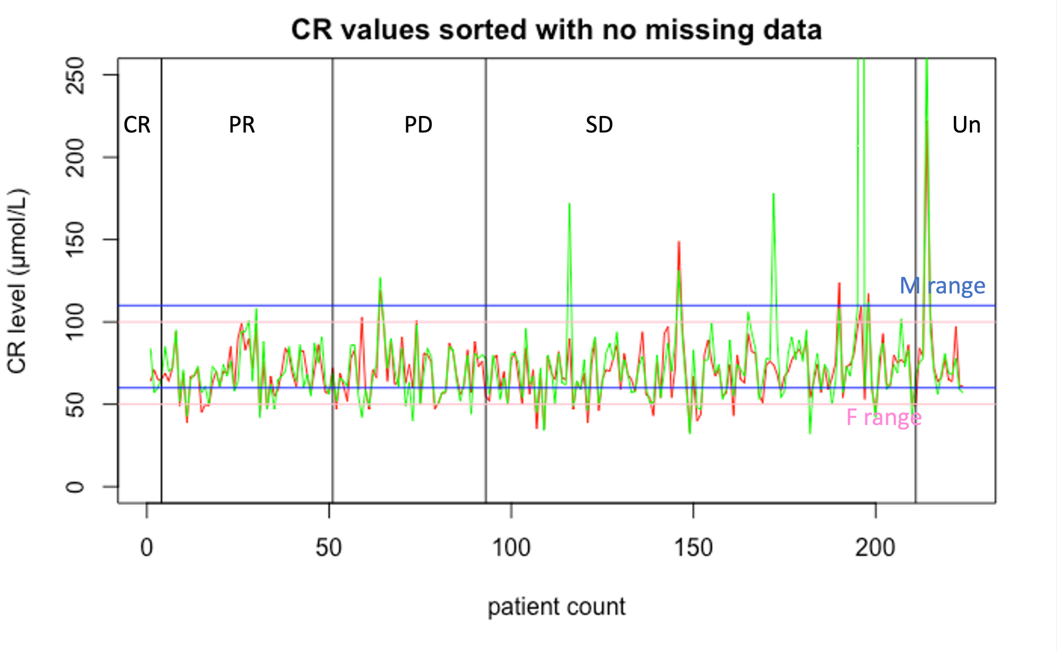


Figure B.85: Creatinin levels ($\mu mol/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=224. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 42), Stable Disease (SD, n=118), Unknown (Un, n=14). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = Healthy male range.

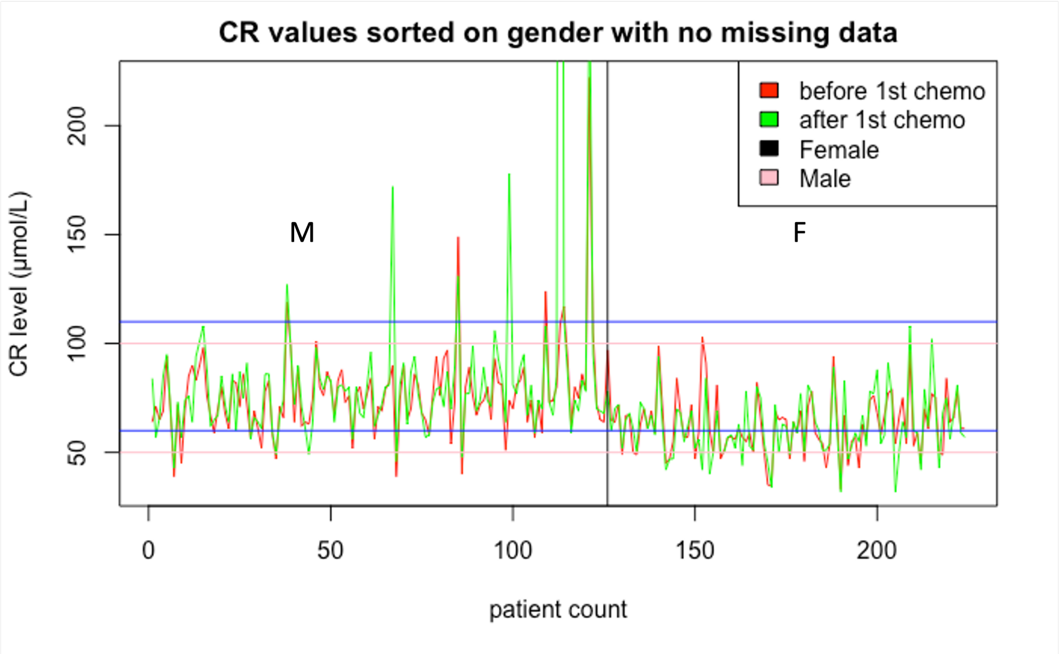


Figure B.86: Creatinin levels in ($\mu mol/L$) before and after the first chemotherapy cycle sorted by gender with no missing values, n=224 (126 male, 98 female). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

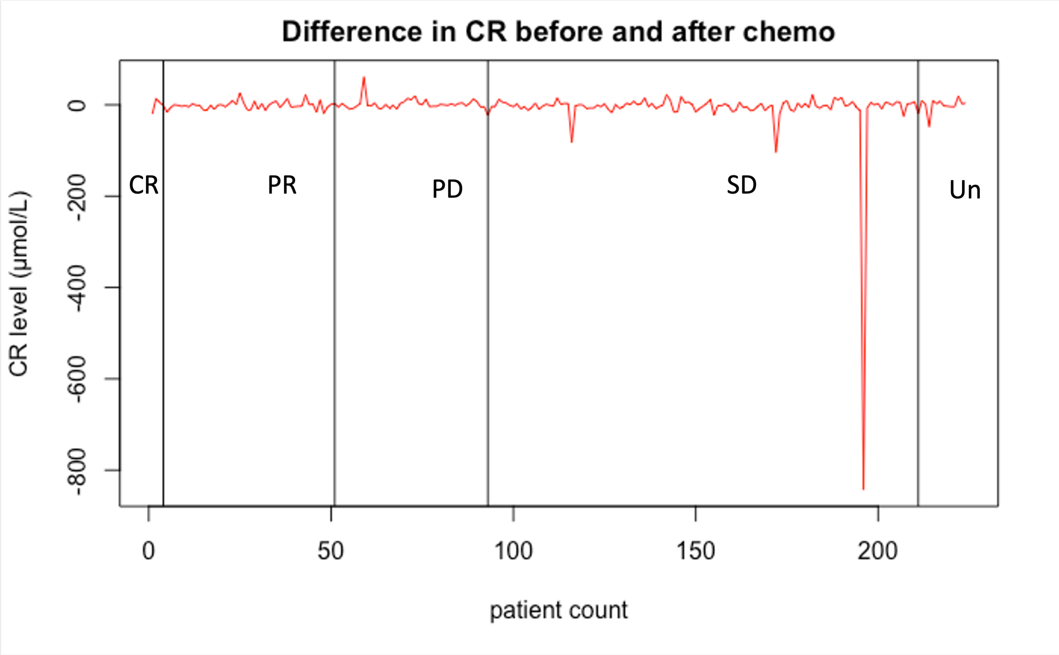


Figure B.87: Difference in creatinin levels ($\mu mol/L$) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=224. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 42), Stable Disease (SD, n=118), Unknown (Un, n=14).

B.1.3.7 Glomerular Filtration Rate (GFR)

The Glomerular Filtration Rate (GFR) is a measure of kidney function that indicates the rate at which the kidneys are filtering blood. Specifically, it is calculated as the sum of the filtration rates of the functional nephrons in the kidney. GFR is an important indicator of kidney health, along with other markers such as creatinin. In accordance with the National Kidney Foundation, a GFR value between 90-120 mL/min is considered healthy [106]. However, it should be noted that GFR typically decreases with age, resulting in lower values for older individuals. In addition, men tend to have more muscle mass compared to females, hence, females will also tend to have a lower GFR value compared to men. In the context of the provided dataset (n=247 (137 male, 110 female)), with an average age of 64 years, most GFR values fall within the healthy range or slightly below it. Interestingly, the median and 3rd quartile values for GFR remain constant at 90.0 mL/min before and after chemotherapy, suggesting that the treatment does not have a significant impact on GFR values. The exact values and other summary statistics can be found in Table B.24. Analysis of the distribution of GFR values before and after chemotherapy in Figure B.88 and Figure B.89, respectively, using both a normality test (results in Table B.25) and a QQplot in Figure B.90 indicate that these values are not normally distributed.

Once any missing observations have been eliminated, the remaining dataset comprises of n = 223 subjects, of which 125 were male and 98 were female. Of these subjects, 3 showed a complete response (CR), 27 partial response (PR), 41 progressive disease (PD), 118 stable disease (SD), and 14 values were unknown. In particular, removal of the missing values did not significantly alter the summary statistics of the dataset provided in Table B.26. Visual inspection of the plots in Figure B.91 and Figure B.92 reveals that the majority of GFR values in the dataset fall below the healthy range. This can be explained by the fact that most patients are elderly and thus have lower levels of GFR in general. Moreover, the most notable discrepancies between GFR values before and after chemotherapy appear to be observed in the PD and SD groups, as depicted in Figure B.94, Figure B.91 and Figure B.92.

GFR Value	*Before 1st Chemo*	*After 1st Chemo*
Min	27.0	22.0
1st Quartile	81.0	78.0
Median	90.0	90.0
Mean	85.1	84.3
3rd Quartile	90.0	90.0
Max	126.0	125.0
NA	7	14

Table B.24: Summary statistics values of the GFR values of the entire dataset in mL/min, n=247.

GFR value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.891	0.884
SW-test pvalue	3.82e-12	2.12e-12
KS test pvalue	Invalid	Invalid

Table B.25: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the GFR values, n=247. The KS-test is invalid due to the presence of ties in the data.

GFR Values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	27.0	22.0	-61.0
1st Quartile	81.0	79.0	-1.0
Median	90.0	90.0	0.0
Mean	85.5	84.7	0.8
3rd Quartile	90.0	90.0	3.0
Max	126.0	125.0	54.0

Table B.26: Summary statistics of the GFR values in mL/min before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=223.

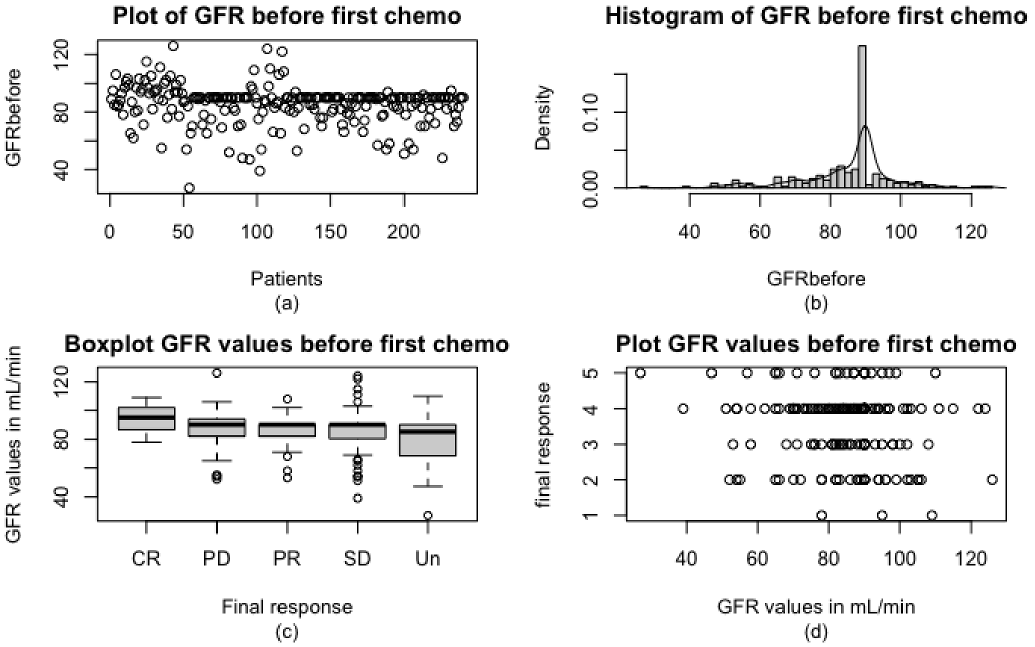


Figure B.88: GFR distribution before the first chemotherapy cycle (mL/min) for the entire dataset: (a) scatterplot of GFR values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of GFR values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

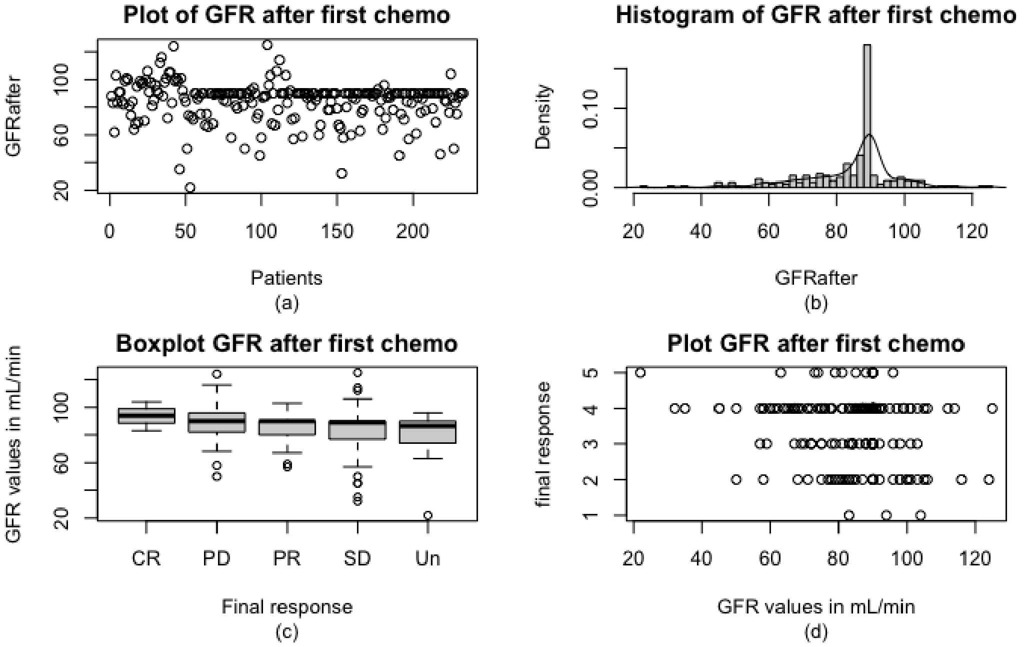


Figure B.89: GFR distribution after the first chemotherapy cycle (mL/min) for the entire dataset: (a) scatterplot of GFR values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of GFR values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

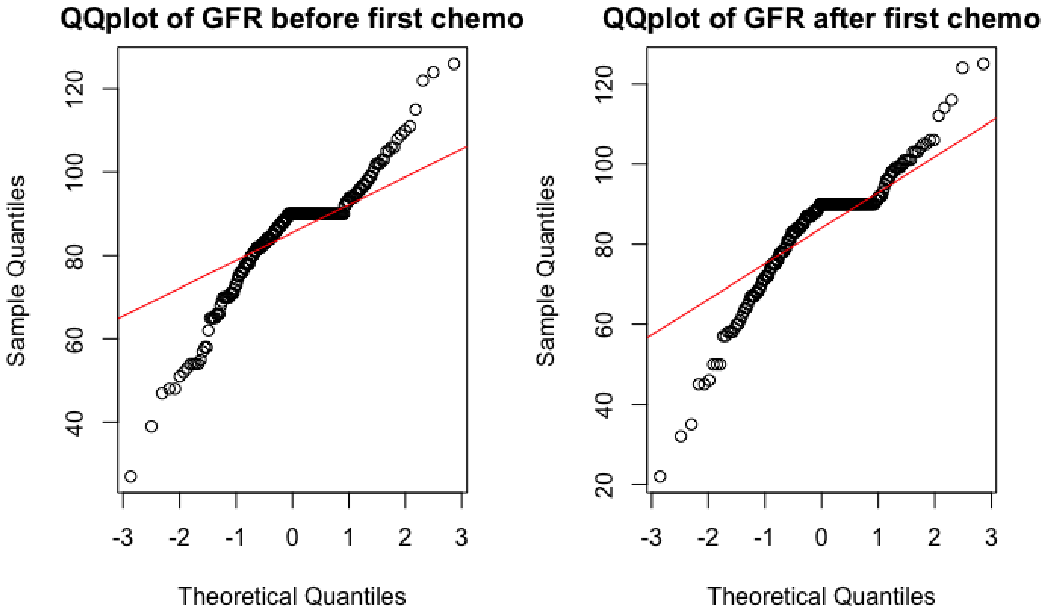


Figure B.90: QQplot of the GFR values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

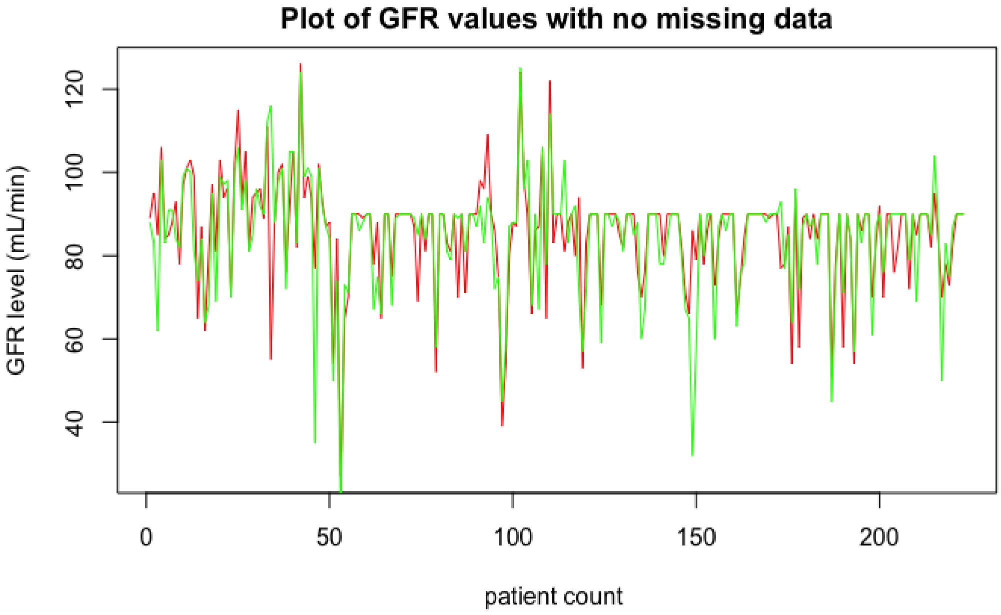


Figure B.91: GFR values (mL/min) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=223.

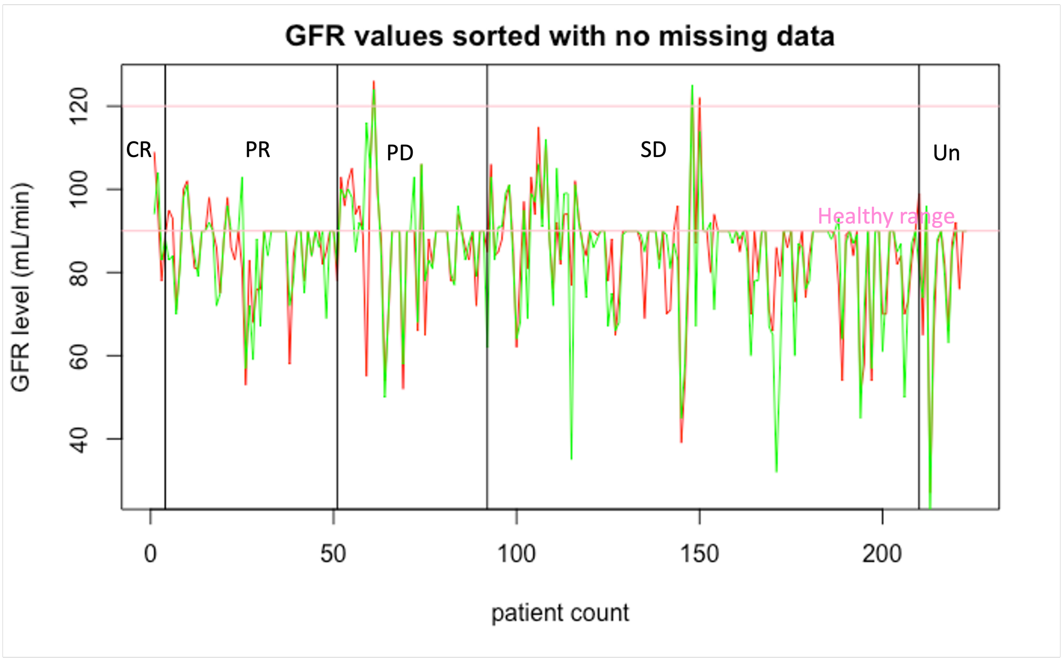


Figure B.92: GFR values (mL/min) before and after the first chemotherapy cycle sorted by final response with no missing values, n=223. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 41), Stable Disease (SD, n=118), Unknown (Un, n=14). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

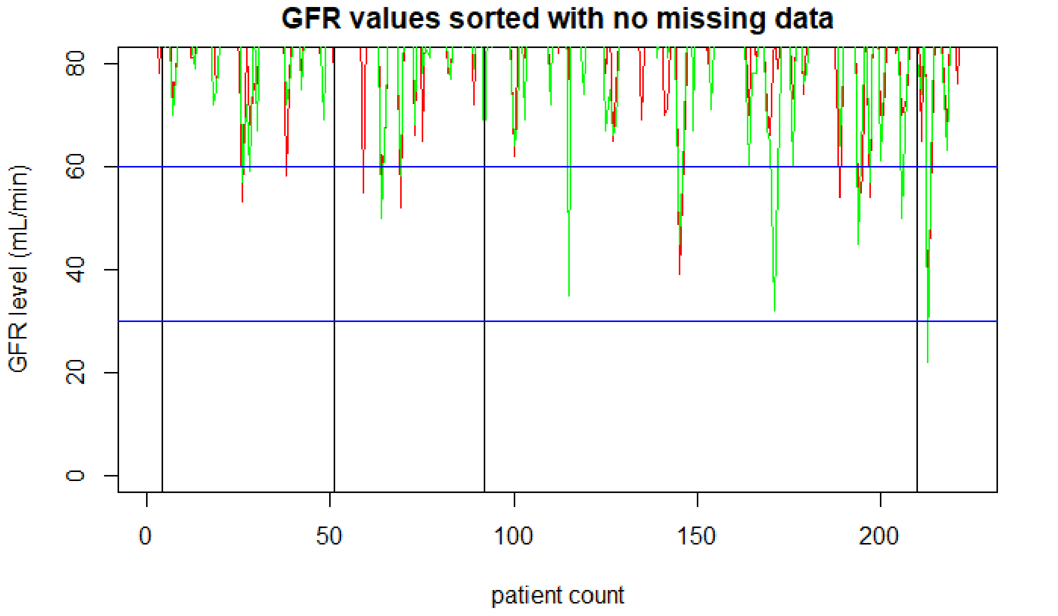


Figure B.93: Same plot as Figure B.92 but with GFR values between 0-80mL/min. The two blue lines are set a level 60 and 30 mL/min as these are two other levels that are critical to consider.

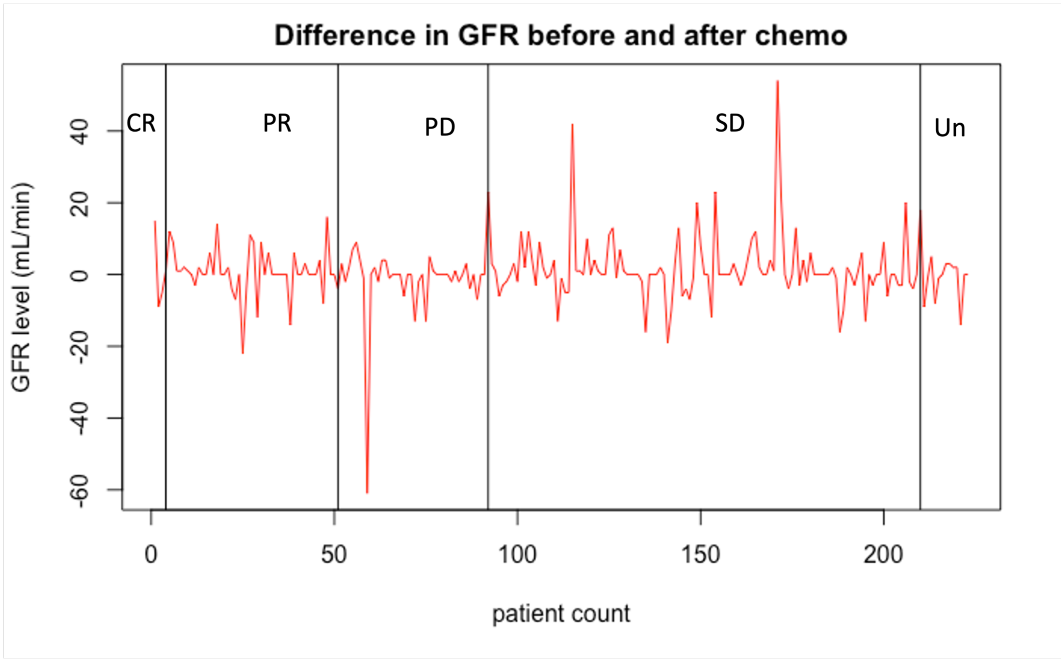


Figure B.94: Difference in GFR values (mL/min) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=223. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=47), Progressive Disease (PD, n= 41), Stable Disease (SD, n=118), Unknown (Un, n=14).

B.1.3.8 Sodium (Na)

Sodium (Na) is a crucial mineral in the human body due to its role in maintaining normal blood pressure, supporting nerve and muscle function, and regulating fluid balance. A healthy range of sodium concentration is typically considered to be between 136-145 mmol/L. However, a low level of sodium (hyponatremia), defined as a concentration below 135 mmol/L, can be hazardous because it can lead to an influx of excess water into cells, causing them to swell. This swelling can be especially dangerous in the brain.

Based on the values presented in Table B.27, the majority of sodium concentrations fall within the healthy range, as indicated by the first quartile up to the maximum values. Moreover, the sodium concentrations before and after chemotherapy appear to be largely unchanged, with similar values observed. However, the histograms depicted in Figure B.95 and Figure B.96, as well as the QQplot in Figure B.97, suggest that the sodium concentrations are not normally distributed, but more skewed towards the right. There is no clear trend to be seen between the final response to chemotherapy and the Na values before or after the first chemotherapy session. Upon discarding any data points with missing values, a total of 160 observations remain, comprising 84 males and 76 females. Among these, no patient exhibited complete response (CR) as a final response, while 36 displayed partial response (PR), 32 progressive disease (PD), 79 stable disease (SD), and 13 were classified as Unknown. Visual inspection of the plots presented in Figure B.98 and Figure B.99 indicates that the majority of sodium concentrations fall within the healthy range, similar to what was seen before removing the missing values. The difference between the values before and after chemotherapy can be seen in Figure B.100, showing no clear distinction between the final response outcomes.

Na Value	*Before 1st Chemo*	*After 1st Chemo*
Min	122.0	128.0
1st Quartile	137.0	137.0
Median	139.0	139.0
Mean	138.5	138.8
3rd Quartile	141.0	141.0
Max	145.0	146.0
NA	59	56

Table B.27: Summary statistics values of the sodium values of the entire dataset in $mmol/L$, n=247.

Na value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.935	0.954
SW-test pvalue	1.96e-7	7.10e-6
KS test pvalue	Invalid	Invalid

Table B.28: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the sodium values, n=247. The KS-test is invalid due to the presence of ties in the data.

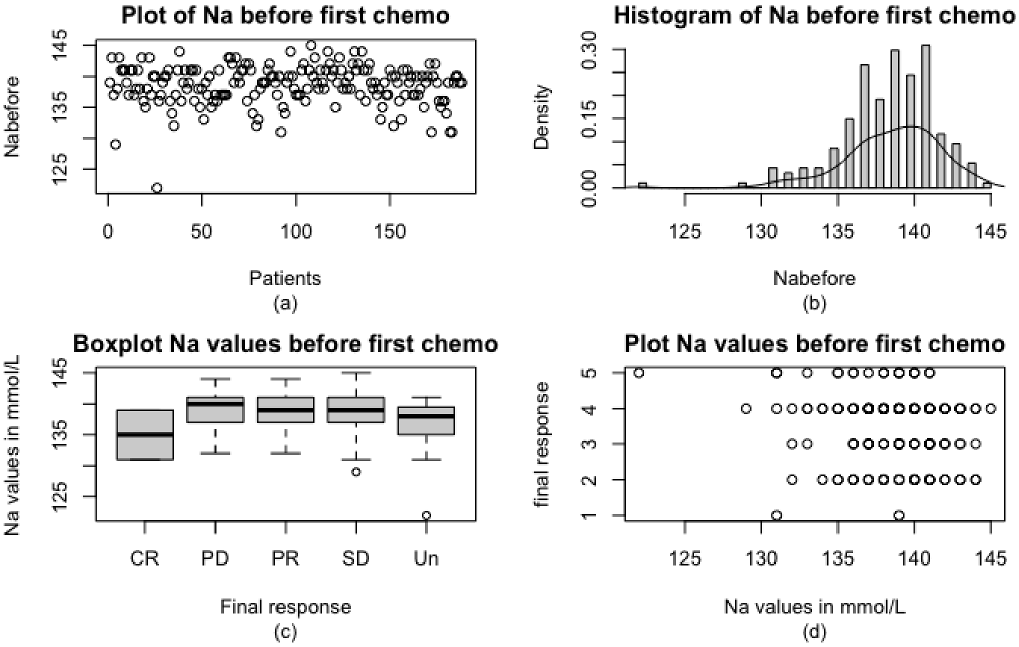


Figure B.95: Sodium distribution before the first chemotherapy cycle ($mmol/L$) for the entire dataset: (a) scatterplot of sodium values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of sodium values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

Na Values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	122	129	-9.0
1st Quartile	137	138	-2.0
Median	139	139	0.0
Mean	139	139	-0.1
3rd Quartile	141	141	2.0
Max	145	146	8.0

Table B.29: Summary statistics of the sodium values in $mmol/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=160.

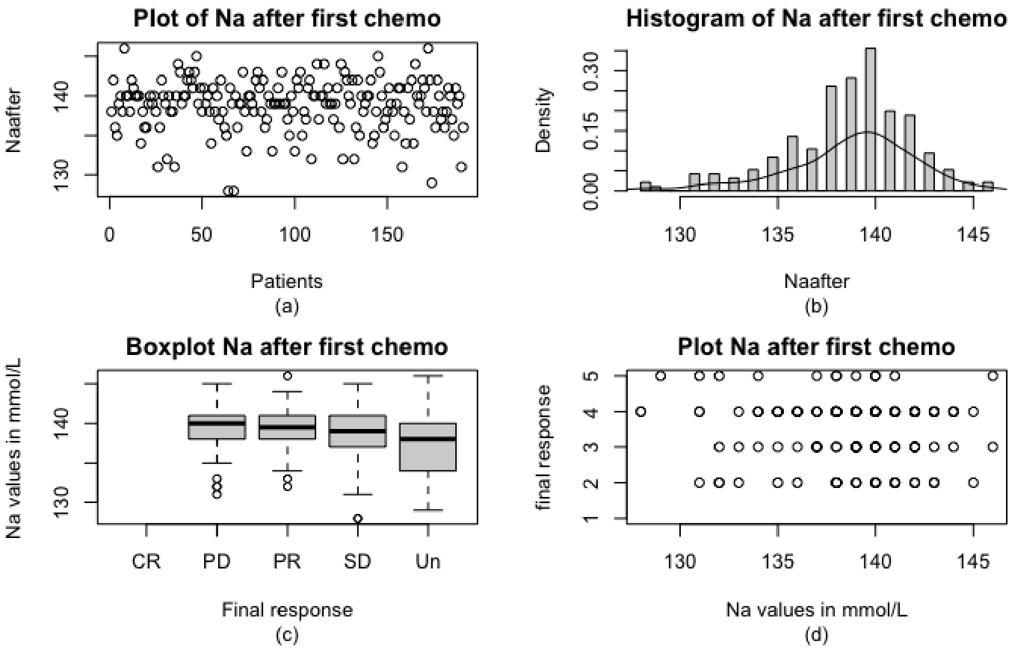


Figure B.96: Sodium distribution after the first chemotherapy cycle ($mmol/L$) for the entire dataset: (a) scatterplot of sodium values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of sodium values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

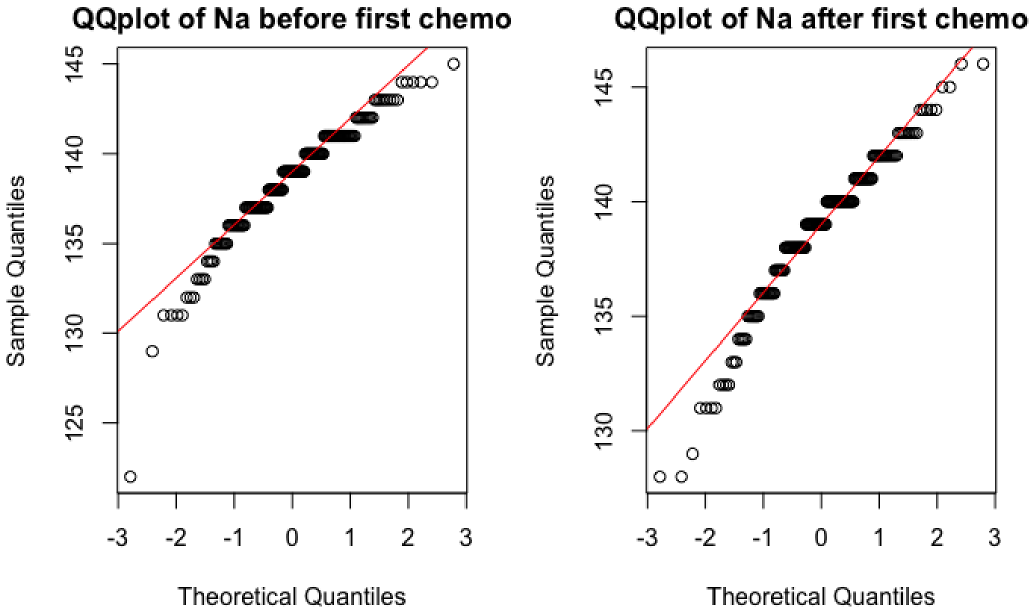


Figure B.97: QQplot of the sodium values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

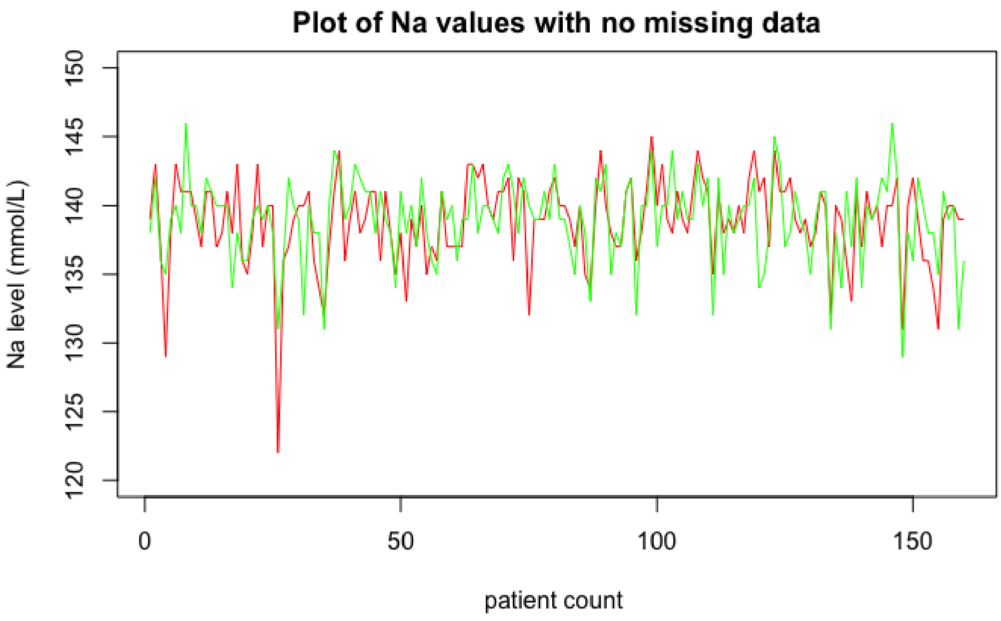


Figure B.98: Sodium values in $(mmol/L)$ before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=160.

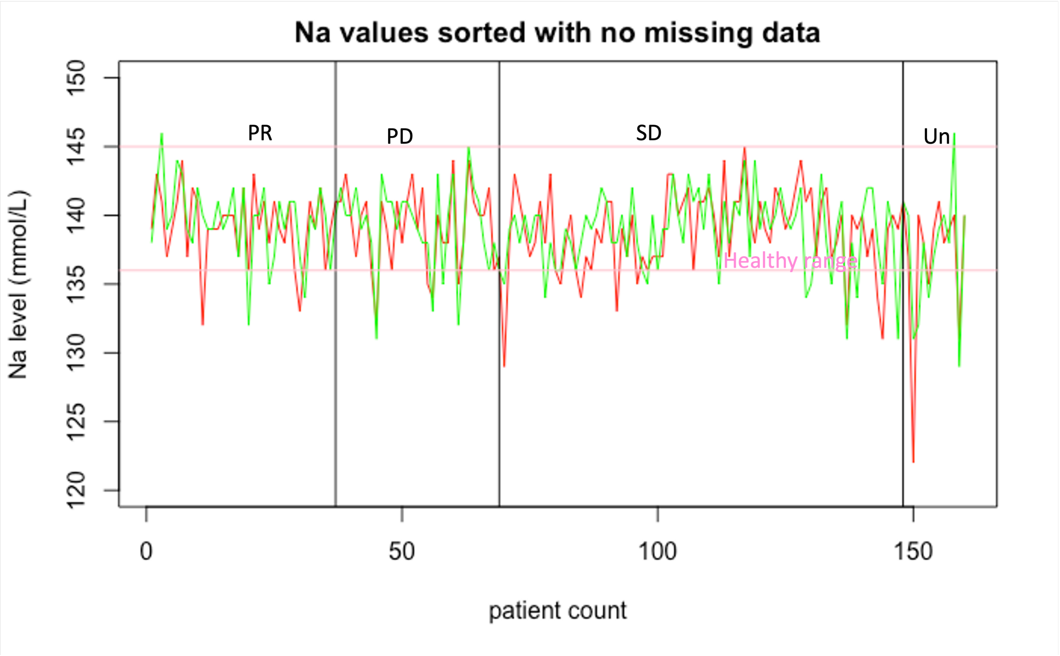


Figure B.99: Sodium levels $(mmol/L)$ before and after the first chemotherapy cycle sorted by final response with no missing values, n=160. The final response is classified as Complete Response (CR, n=0), Partial response (PR, n=36), Progressive Disease (PD, n=32), Stable Disease (SD, n=79), Unknown (Un, n=13). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

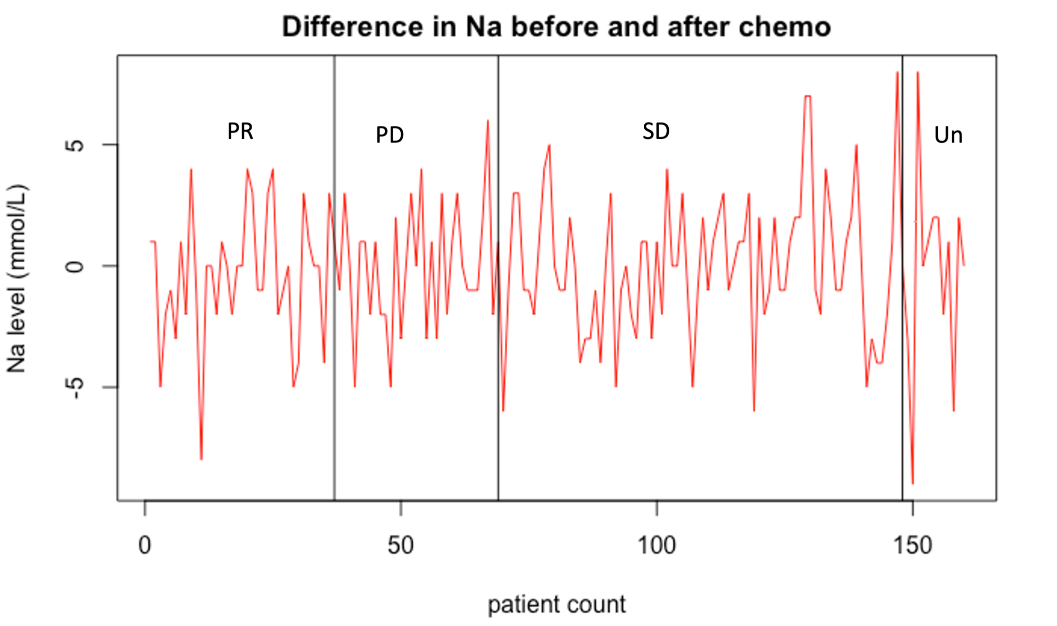


Figure B.100: Difference in sodium values (mmol/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=160. The final response is classified as Complete Response (CR, n=0), Partial response (PR, n=36), Progressive Disease (PD, n=32), Stable Disease (SD, n=79), Unknown (Un, n=13).

B.1.3.9 Potassium (K)

Potassium is an essential mineral in the human body that plays a vital role in nerve and muscle function, cellular nutrient exchange, waste removal, and the maintenance of normal cardiac function. The healthy range for potassium in blood is between 3.5 and 5.1 mmol/L according to medical experts from the Erasmus Medical Centre Rotterdam. One patient in the dataset (patient ID 002PP20005) had a recorded potassium value of 402.0 mmol/L, which is dangerously high. However, it is highly probable that this was a typographical error, and the correct value should have been 4.02 mmol/L. Hence, the high potassium value has been corrected to this more sensible value. The summary statistics presented in Table B.30 indicate that almost all potassium values fall within the healthy range, with only a few values falling below or above the minimum and maximum values (after the correction of 402 to 4.02 mmol/L). The histograms in Figures B.101 and B.102 suggest normality, which is confirmed by both the Shapiro-Wilk test and the QQplots. These findings suggest that the potassium levels in PDAC patients appear to be normally distributed around the mean. Furthermore, there appears no clear distinguishment between potassium values before or after the first chemotherapy session and the final response outcome.

Consider the dataset after removing missing values. This dataset contains n = 158 observations, with 83 males and 75 females. Final response to chemotherapy includes 36 partial responders (PR), 31 progressive disease (PD), 78 stable disease (SD), and 13 Unknown cases, with no complete responders (CR). The data visualizations in Figure B.104 and Figure B.105, the latter of which is sorted based on the final response to chemotherapy, suggest an overall slight decrease in K values post-chemotherapy. While many values were within the healthy range prior to the first chemotherapy, some have fallen below the healthy range post-chemotherapy. This may be due to side effects such as chemotherapy-induced vomiting or diarrhea. However, no clear association can be established between the K values before or after chemotherapy and the final response to chemotherapy, nor in the difference between K values.

K Value	*Before 1st Chemo*	*After 1st Chemo*
Min	3.1	2.4
1st Quartile	4.0	3.6
Median	4.2	4.0
Mean	4.2	3.9
3rd Quartile	4.5	4.2
Max	5.3	5.4
NA	61	56

Table B.30: Summary statistics values of the potassium values of the entire dataset in $mmol/L$, n=247.

K value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.991	0.991
SW-test pvalue	0.27	0.28
KS test pvalue	Invalid	Invalid

Table B.31: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the potassium values, n=247. The KS-test is invalid due to the presence of ties in the data.

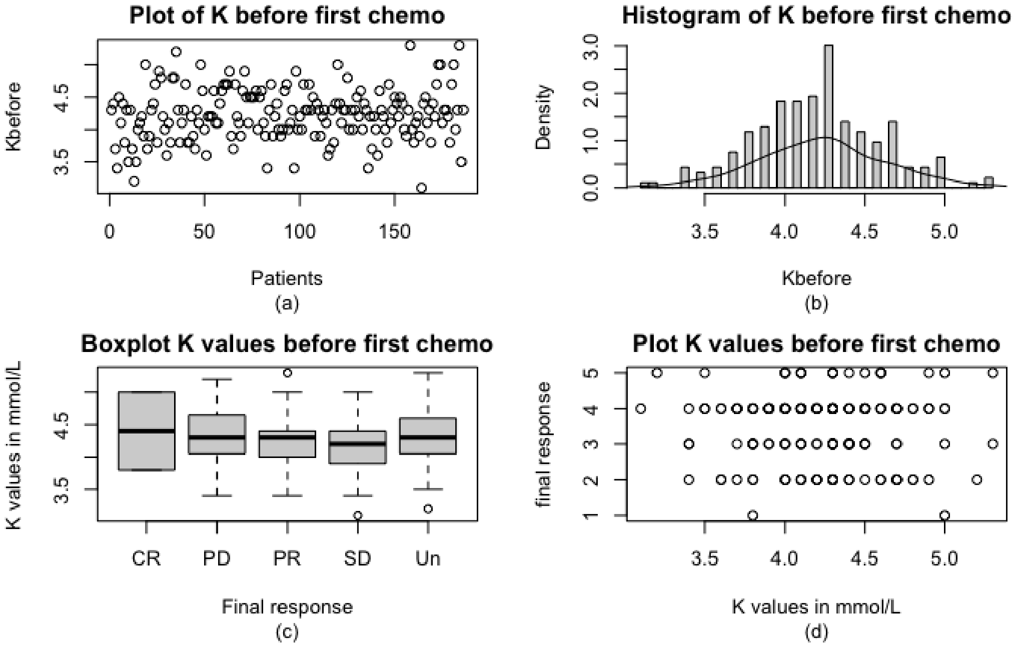


Figure B.101: Potassium distribution before the first chemotherapy cycle ($mmol/L$) for the entire dataset: (a) scatterplot of potassium values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of potassium values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

K Values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	3.2	2.4	-1.0
1st Quartile	4.0	3.6	-0.1
Median	4.3	4.0	0.3
Mean	4.2	4.0	0.3
3rd Quartile	4.5	4.3	0.5
Max	5.3	5.4	1.8

Table B.32: Summary statistics of the potassium values count in $mmol/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=158.

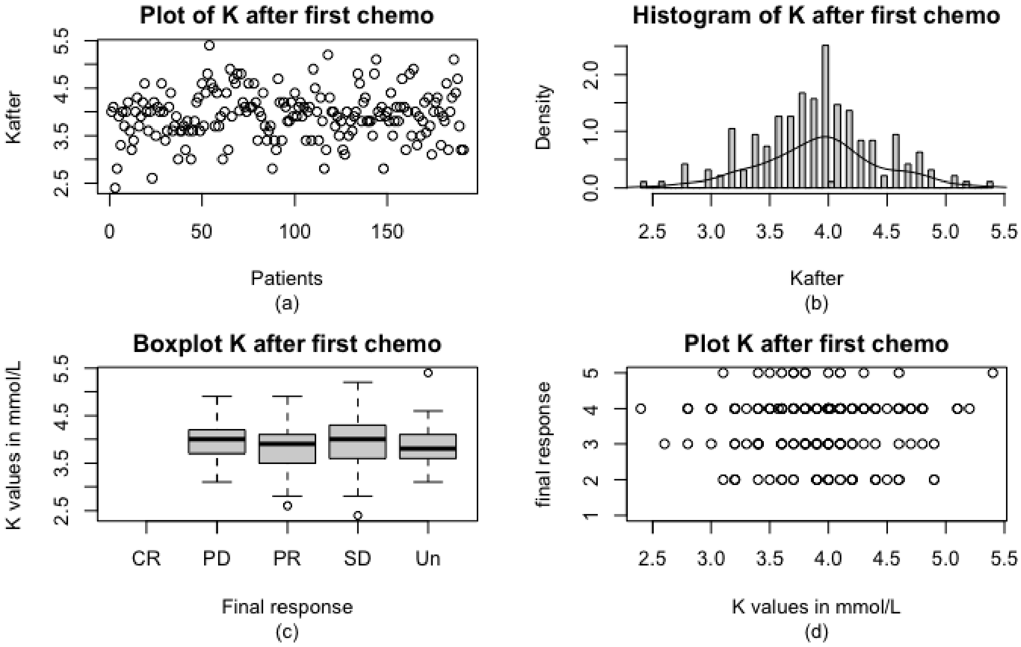


Figure B.102: Potassium distribution after the first chemotherapy cycle ($mmol/L$) for the entire dataset: (a) scatterplot of potassium values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of potassium values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

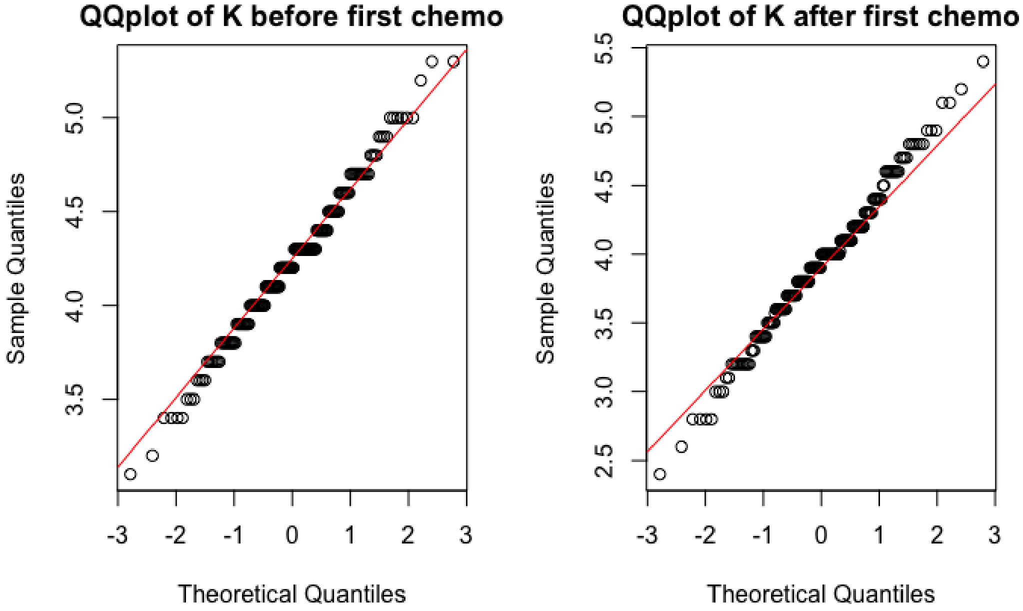


Figure B.103: QQplot of the potassium values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

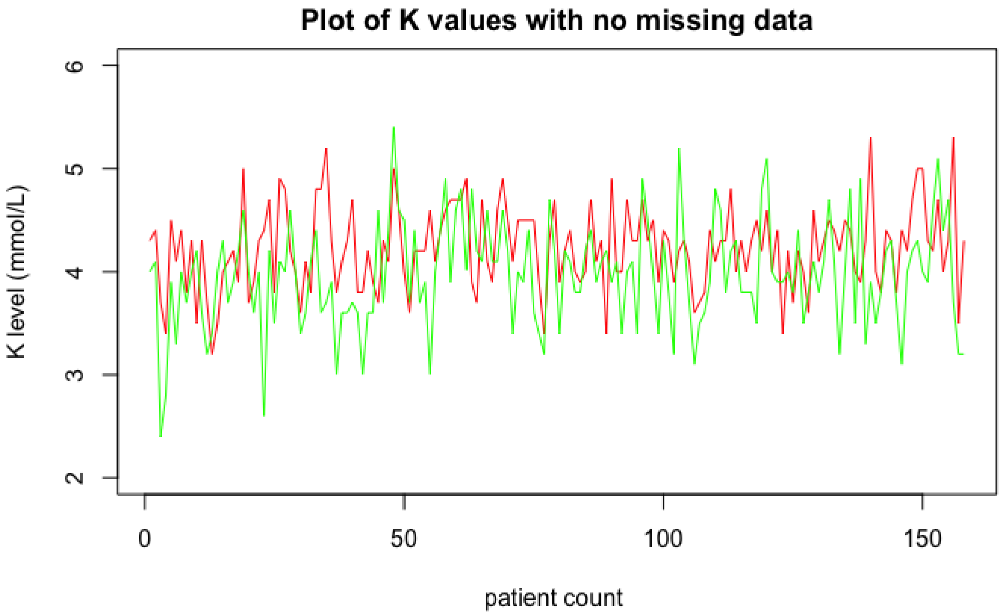


Figure B.104: Potassium levels ($mmol/L$)) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=158.

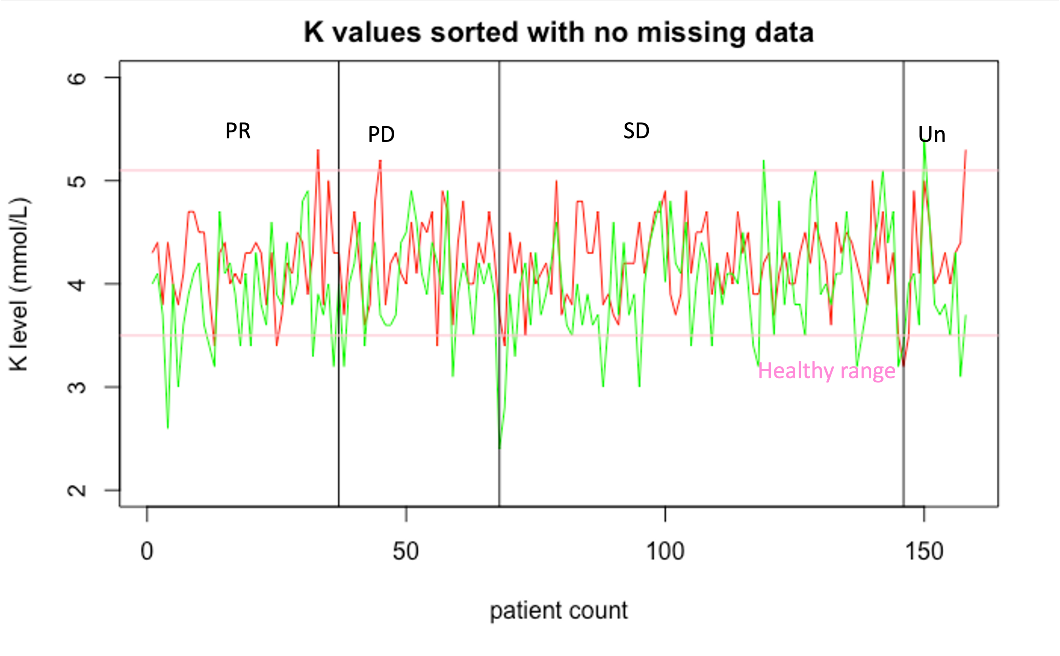


Figure B.105: Potassium levels ($mmol/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=158. The final response is classified as Complete Response (CR, n=0), Partial response (PR, n=36), Progressive Disease (PD, n= 31), Stable Disease (SD, n=78), Unknown (Un, n=13). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

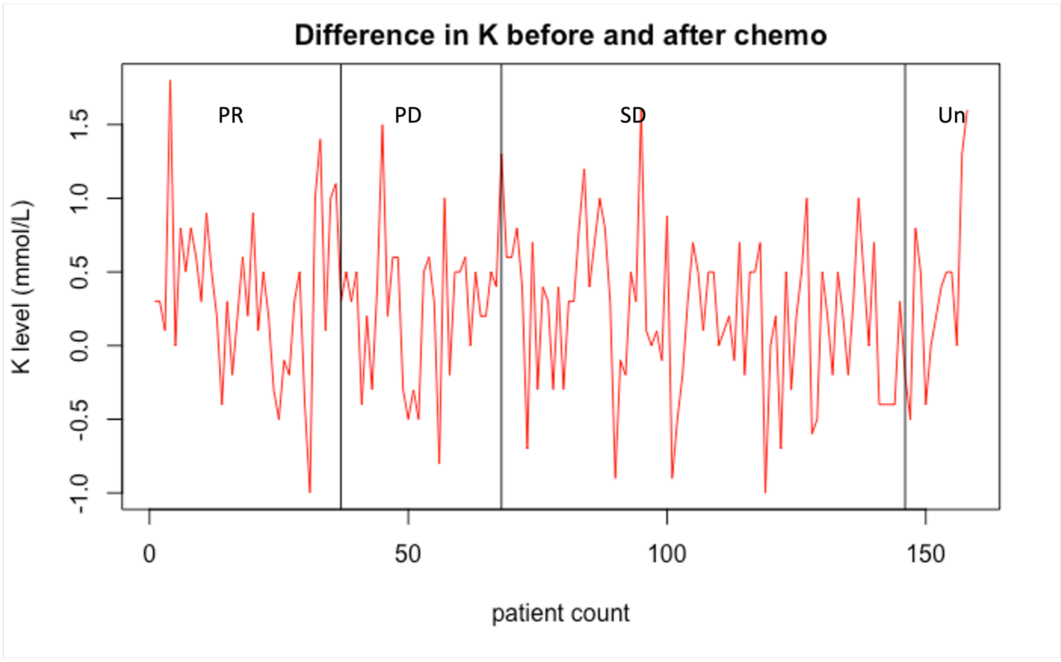


Figure B.106: Difference in potassium levels (mmol/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=158. The final response is classified as Complete Response (CR, n=0), Partial response (PR, n=36), Progressive Disease (PD, n= 31), Stable Disease (SD, n=78), Unknown (Un, n=13).

B.1.3.10 Aspartate Aminotransferase (ASAT)

Aspartate aminotransferase (ASAT) is an enzyme that provides information about liver function. It is present in various body tissues, particularly in the liver and muscles. The liver has several crucial functions, such as producing bile to aid digestion, removing toxins from the blood, producing clotting proteins, and processing drugs and alcohol. The ASAT test is commonly used to diagnose and monitor liver disease, with a healthy range of values typically below 25 U/L according to medical experts from the Erasmus Medical Centre Rotterdam. In cases of liver damage, ASAT levels in the blood will increase. With regards to the dataset provided, patient 001PANC0052 had an extremely high reported ASAT value of 1100 U/L before chemotherapy, which is likely a typographical error. The value is expected to be between 100-110 U/L. Therefore, it has been adjusted to 110 U/L before the analysis is performed.

Based on the histograms and QQplots in Figure B.107, Figure B.108, and Figure B.109, it is evident that the ASAT values are not normally distributed, mainly due to high values in the right tail. However, since the frequency of these large values is relatively small, it can be inferred that most values appear to be reasonably normal. This trend can be clearly seen back in the scatter plot as well as the boxplots, with no significant difference between the final response groups. When disregarding any values that are missing, the resulting dataset comprises 192 observations, with 102 males and 90 females. The final response in this dataset includes 3 patients with CR, 41 with PR, 34 with PD, 102 with SD, and 12 unknown values. From the plot in Figure B.110, it is evident that the ASAT values seem to have decreased after chemotherapy, particularly in patients with high pre-chemotherapy ASAT values. In particular, the highest spikes in ASAT values occur in the group of patients with stable disease (SD) as the final response. However, no solid hypothesis can be drawn from Figure B.112. Many patients' ASAT values seem not be influenced by chemotherapy, with some patients showing big differences, as seen in Figure B.113. These patients most likely have liver failure or other liver conditions, causing big spikes.

ASAT Value	*Before 1st Chemo*	*After 1st Chemo*
Min	10.0	9.0
1st Quartile	20.8	22.0
Median	27.0	28.0
Mean	46.5	32.7
3rd Quartile	39.0	35.0
Max	1100.0	222.0
NA	23	41

Table B.33: Summary statistics values of the ASAT values of the entire dataset in U/L, n=247.

ASAT value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.299	0.531
SW-test pvalue	<2.2e-16	<2.2e-16
KS test pvalue	Invalid	Invalid

Table B.34: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the ASAT values, n=247. The KS-test is invalid due to the presence of ties in the data.

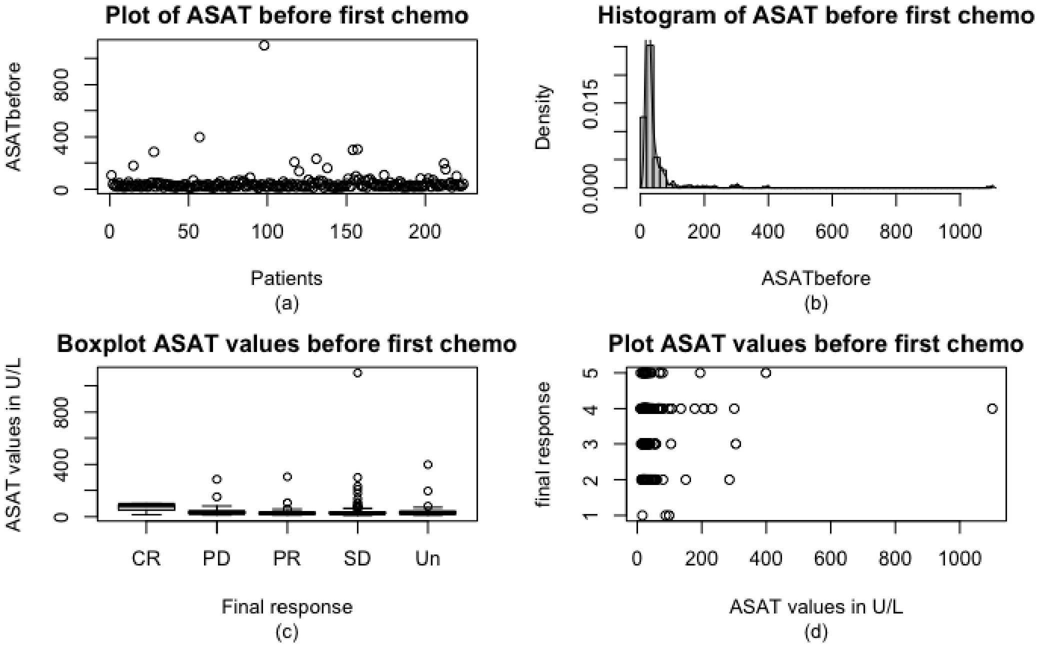


Figure B.107: ASAT distribution before the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of ASAT values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of ASAT values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

ASAT Values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	10.0	9.0	-191.0
1st Quartile	20.0	22.0	-7.0
Median	26.5	27.5	1.0
Mean	44.9	32.8	12.2
3rd Quartile	37.3	34.3	9.0
Max	1100.0	222.0	1079.0

Table B.35: Summary statistics of the ASAT values (U/L) before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=192.

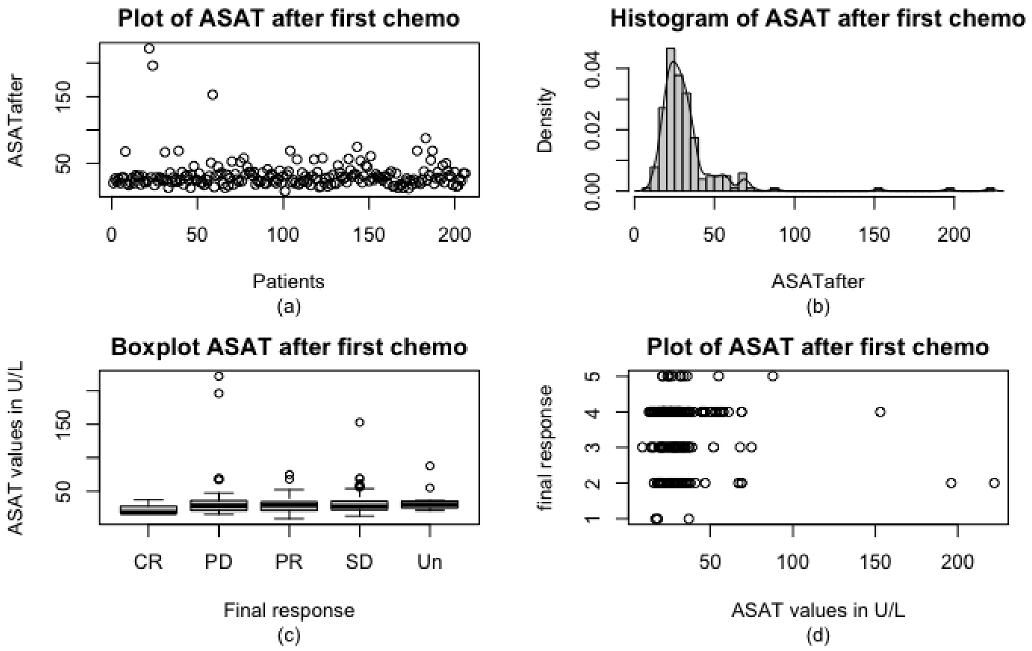


Figure B.108: ASAT distribution after the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of ASAT values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of ASAT values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

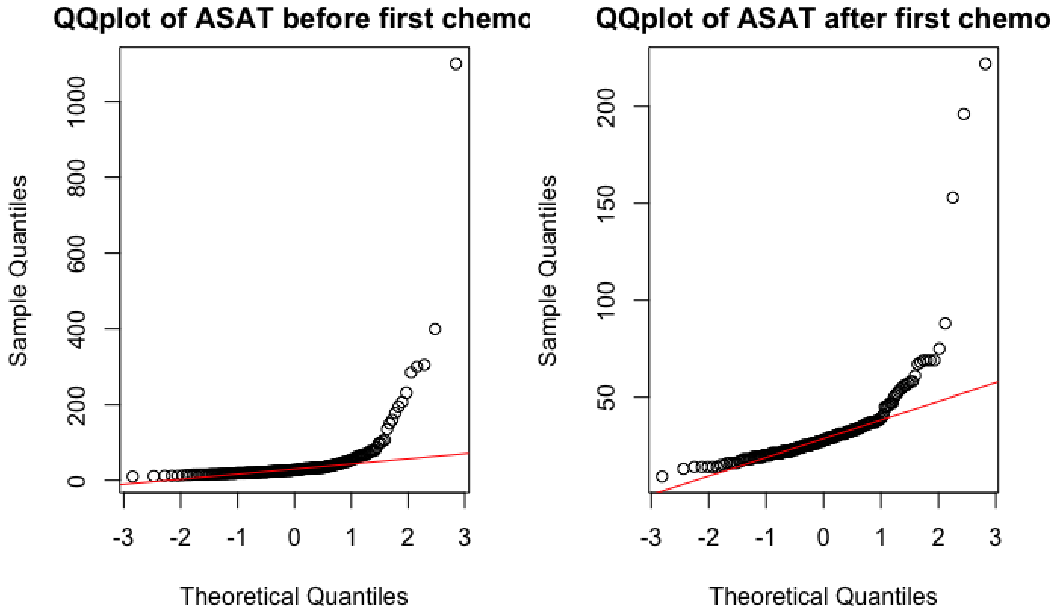


Figure B.109: QQplot of the ASAT values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

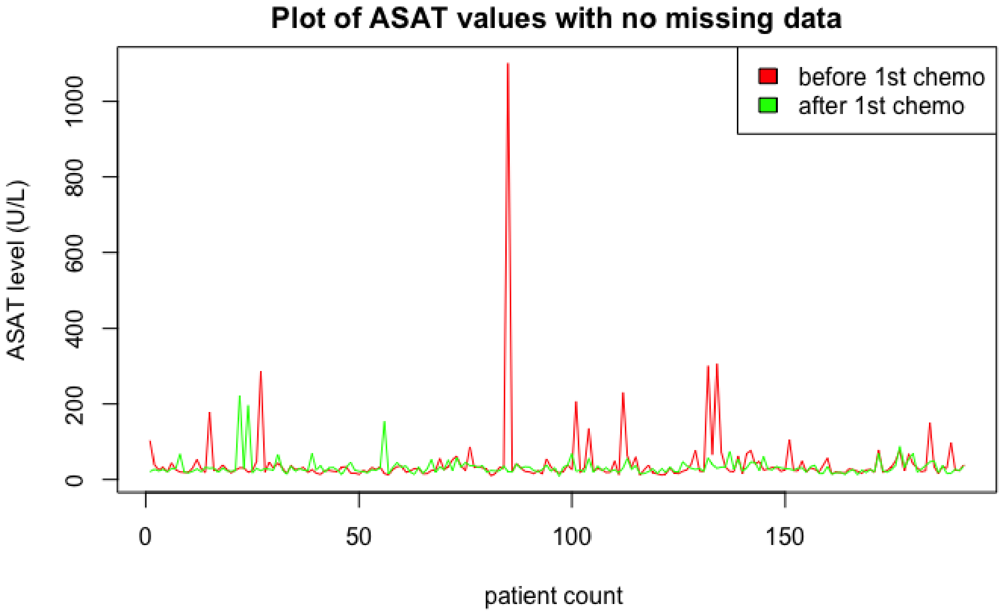


Figure B.110: ASAT values(U/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=192.

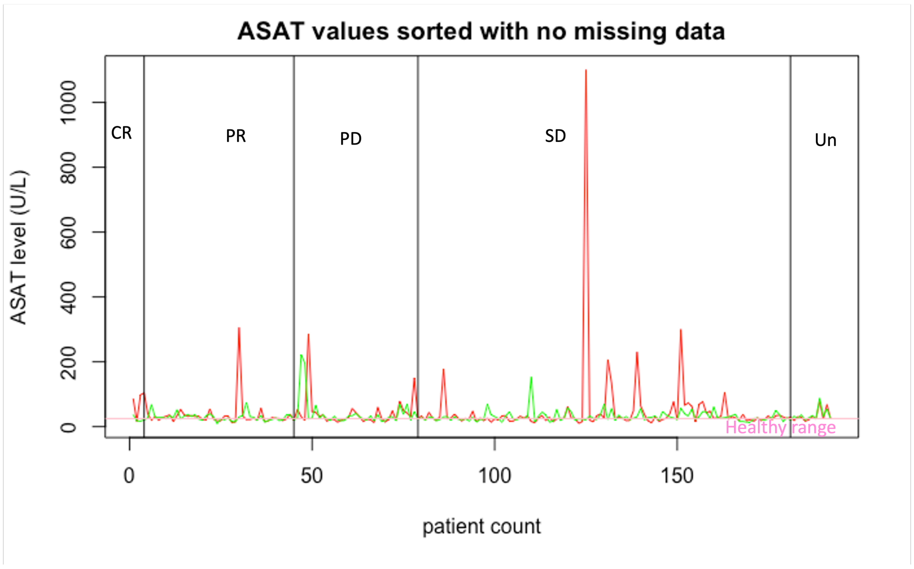


Figure B.111: ASAT values (U/L) before and after the first chemotherapy cycle sorted by final response with no missing values, n=192. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=41), Progressive Disease (PD, n= 34), Stable Disease (SD, n=102), Unknown (Un, n=12). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

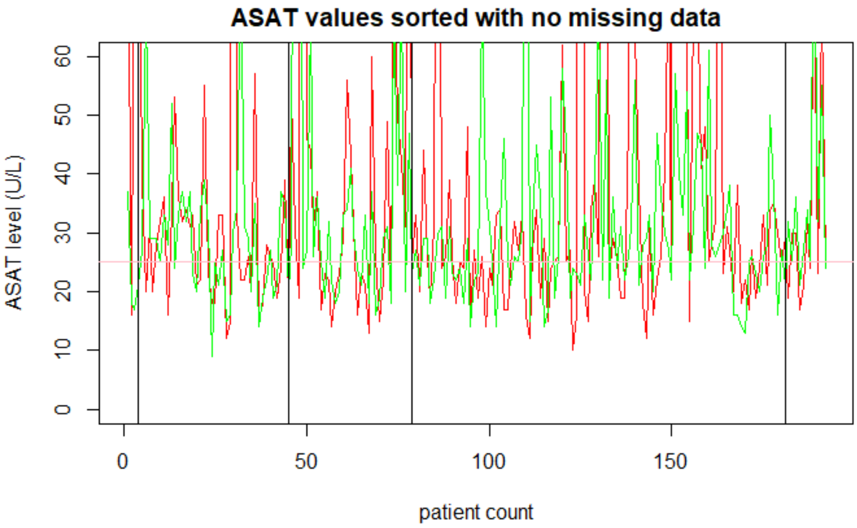


Figure B.112: Same plot as Figure B.112 but with ASAT values restricted from 0 to 60 U/L.

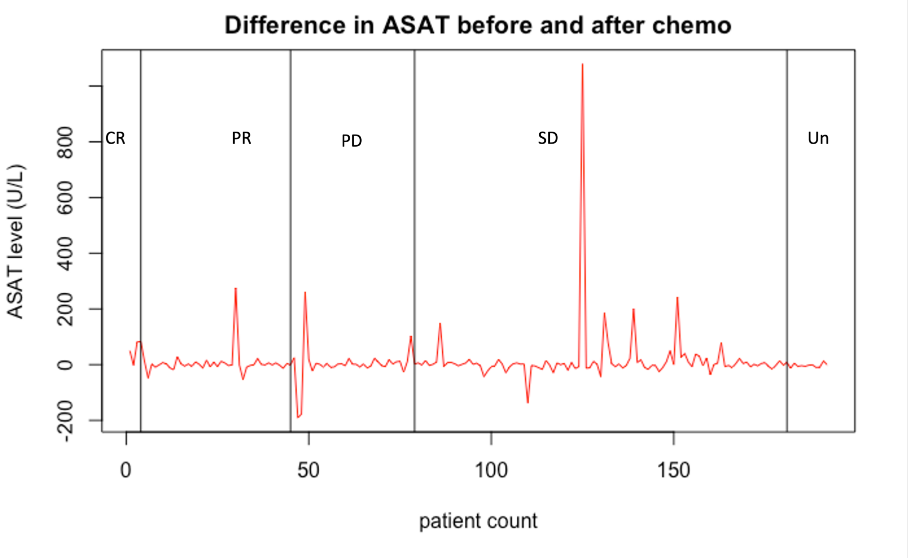


Figure B.113: Difference in ASAT values (U/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=192. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=41), Progressive Disease (PD, n= 34), Stable Disease (SD, n=102), Unknown (Un, n=12).

B.1.3.11 Alanine Transaminase (ALAT)

The ALAT blood test is an additional blood test that is used to assess liver function. ALAT, or Alanine Transaminase, is an enzyme primarily found in the liver, just like ASAT. The release of ALAT into the bloodstream is indicative of liver cell damage, making it a useful marker for identifying liver failure or disease. Experts from the Erasmus Medical Centre Rotterdam suggest that healthy ALAT values should be below 45 U/L for men and below 35 U/L for women.

Based on the information presented in Table B.36, it can be seen that the mean ALAT values before and after chemotherapy were 63.4 U/L and 43.7 U/L, respectively. Most of the ALAT values are above the healthy range, with a maximum value of 1241.0 U/L, which is extremely high considering that a healthy value below 45 U/L is recommended for men and below 35 U/L for women. The histograms in Figure B.114 and Figure B.115 show that the distributions are left-skewed, with a long right tail, caused by outlier values, similar to the ASAT values. The QQplot in Figure B.116 and the values presented in Table B.37 reject normality for ALAT values. However, this is most likely the result of the big outlier values caused by patients with a worsened liver functionality. After filtering out any incomplete data, the dataset contains 218 observations, with 121 males and 97 females. Out of these, 3 showed CR, 45 PR, 39 PD, 116 SD, and 15 values were unknown. The plots in Figure B.117, Figure B.118, and Figure B.120 indicate that many of the ALAT values remain above the healthy range even after chemotherapy, with no clear increasing or decreasing trend. The plots in Figure B.119 and Figure B.121 show a zoomed-in plot of the ALAT values, restricting the y-values from 5-120 U/L. Finally, the plot in Figure B.122 depicts the difference between the ALAT values before and after chemotherapy, sorted by the final response to treatment. All together, these plots show no clear distinction between final response to chemotherapy and the ALAT values.

ALAT Value	*Before 1st Chemo*	*After 1st Chemo*
Min	8.0	9.0
1st Quartile	23.0	25.0
Median	24.0	24.0
Mean	63.4	43.7
3rd Quartile	57.3	49.0
Max	1241.0	264.0
NA	11	15

Table B.36: Summary statistics values of the ALAT values of the entire dataset in U/L, n=247.

ALAT value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.393	0.706
SW-test pvalue	<2.2e-16	<2.2 e-16
KS test pvalue	Invalid	Invalid

Table B.37: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the ALAT values, n=247. The KS-test is invalid due to the presence of ties in the data.

ALAT values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	8.0	9.0	-221.0
1st Quartile	23.0	25.0	-11.8
Median	34.0	34.0	1.0
Mean	59.2	44.4	14.9
3rd Quartile	54.8	49.0	13.0
Max	1241.0	264.0	1183.0

Table B.38: Summary statistics of the ALAT values in U/L before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=218.

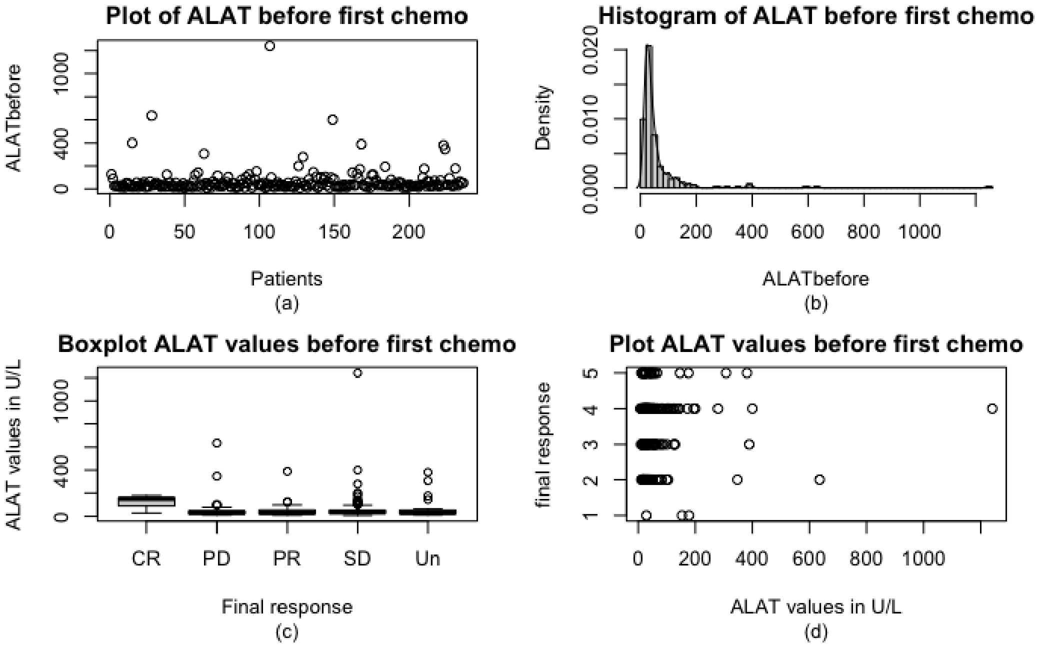


Figure B.114: ALAT distribution before the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of ALAT values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of ALAT values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

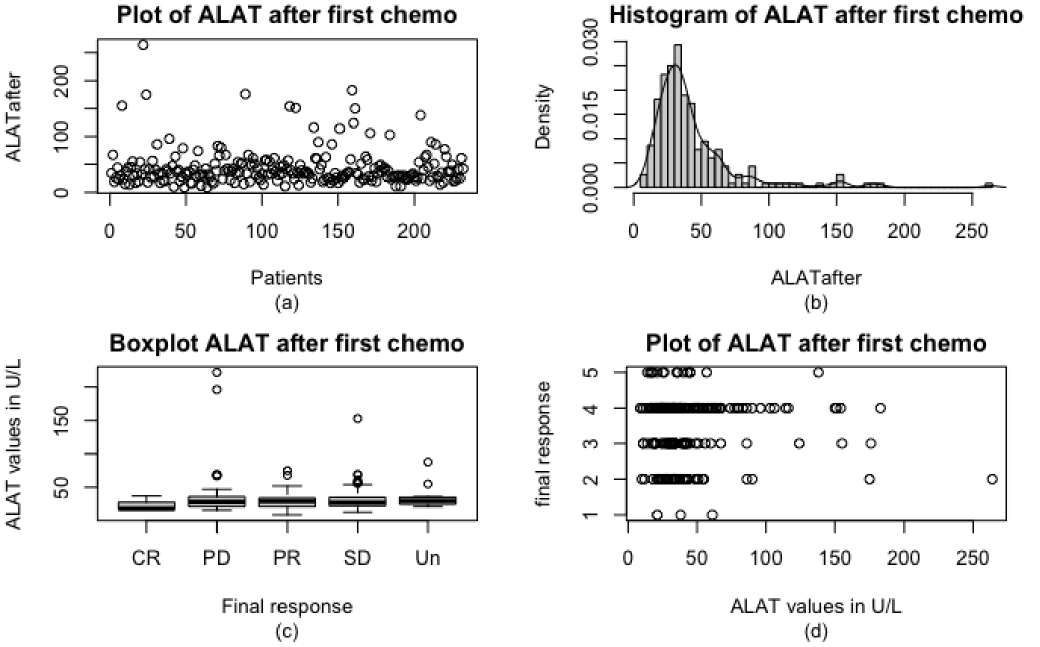


Figure B.115: ALAT distribution after the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of ALAT values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of ALAT values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

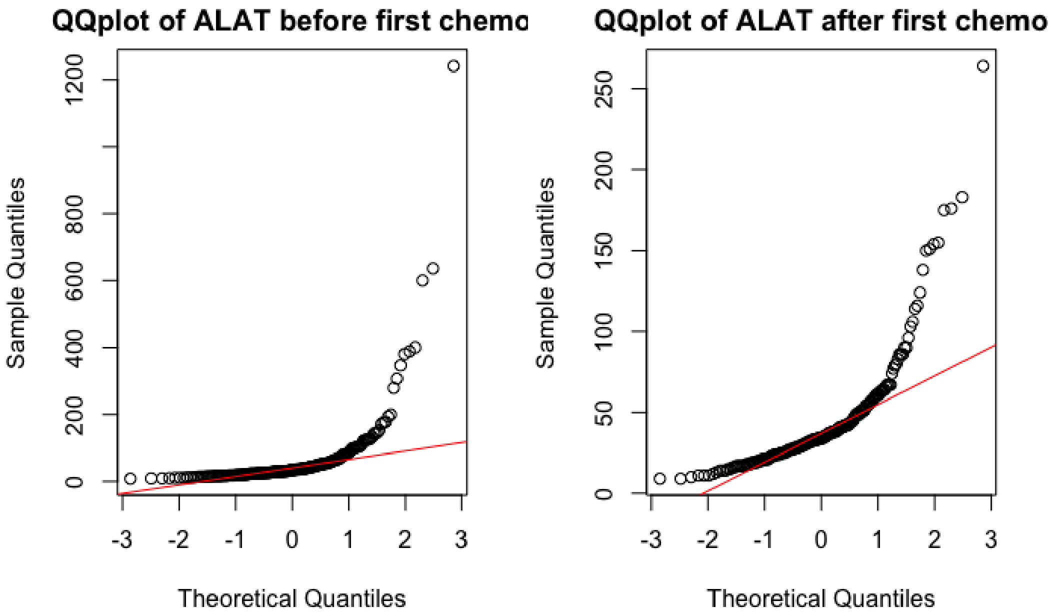


Figure B.116: QQplot of the ALAT values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

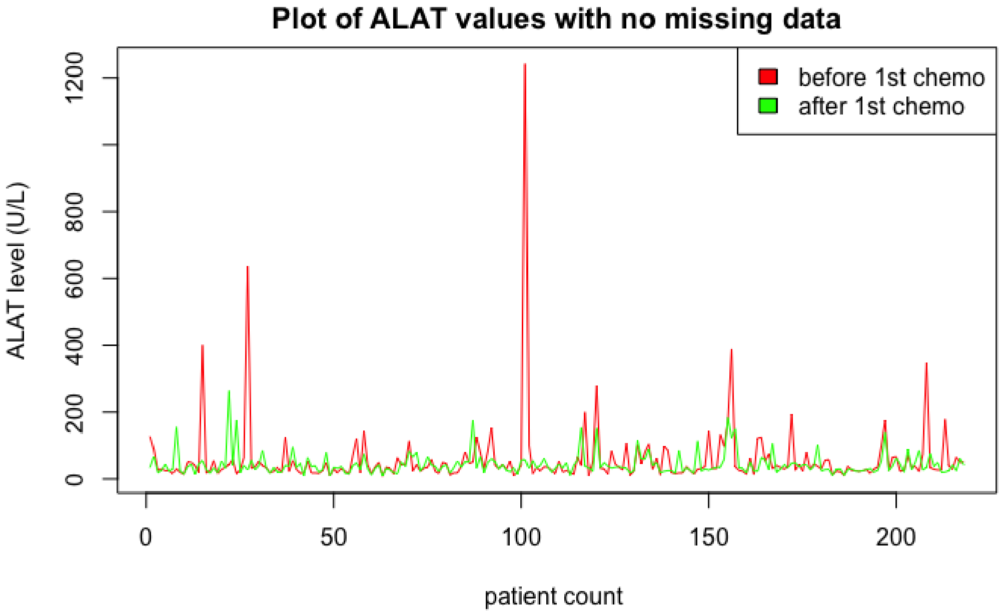


Figure B.117: ALAT levels count (U/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=218.

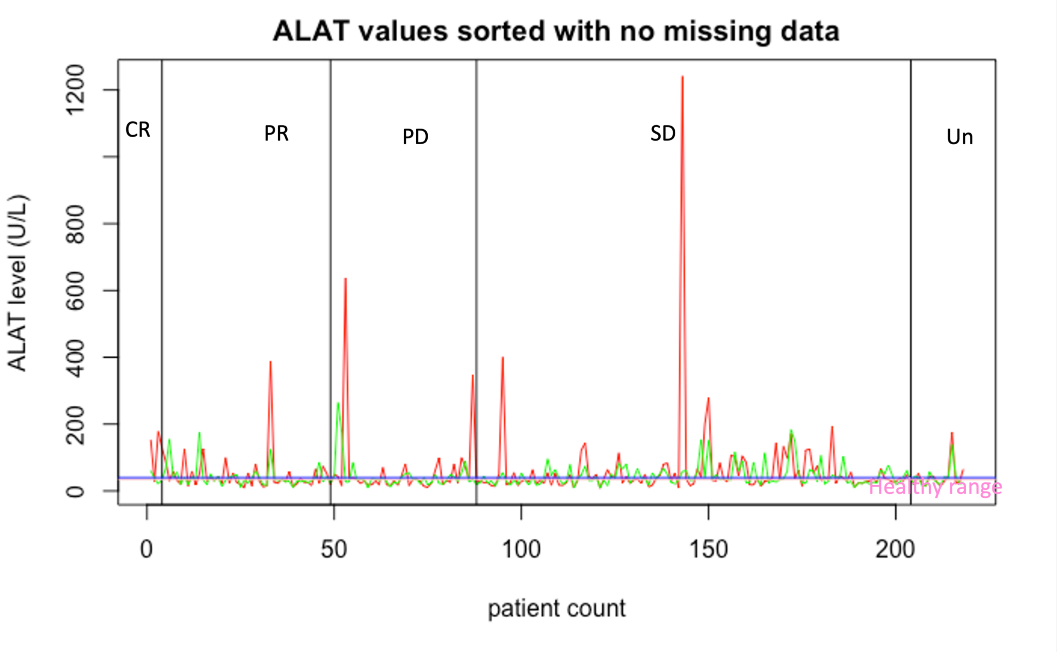


Figure B.118: ALAT levels (U/L) before and after the first chemotherapy cycle sorted by final response with no missing values, n=218. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=45), Progressive Disease (PD, n=39), Stable Disease (SD, n=116), Unknown (Un, n=15). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

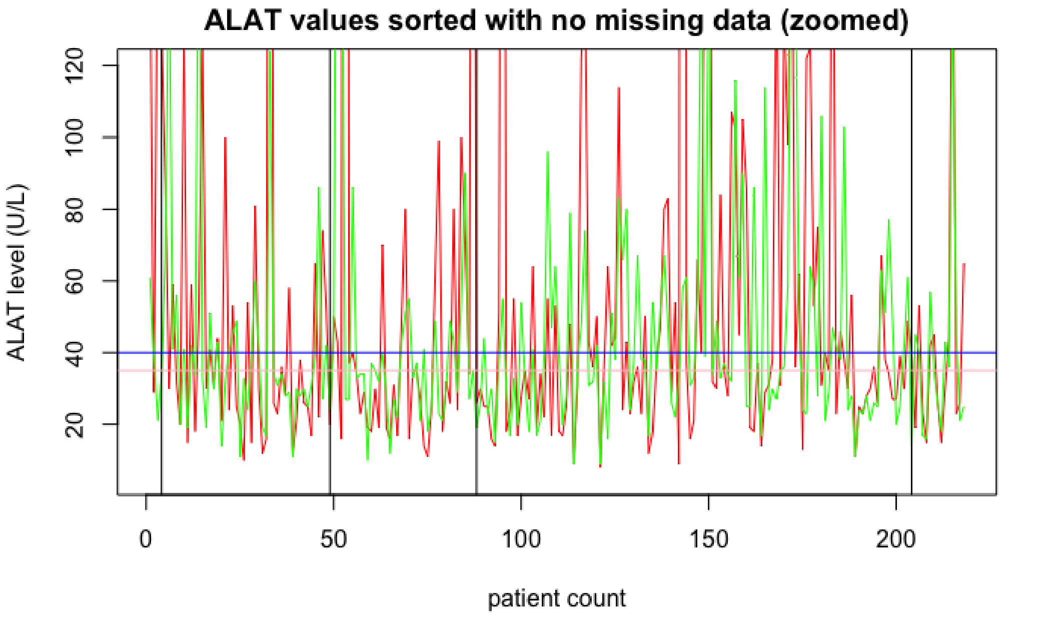


Figure B.119: Same plot as Figure B.118 but the ALAT levels are limited from 5-120 U/L.

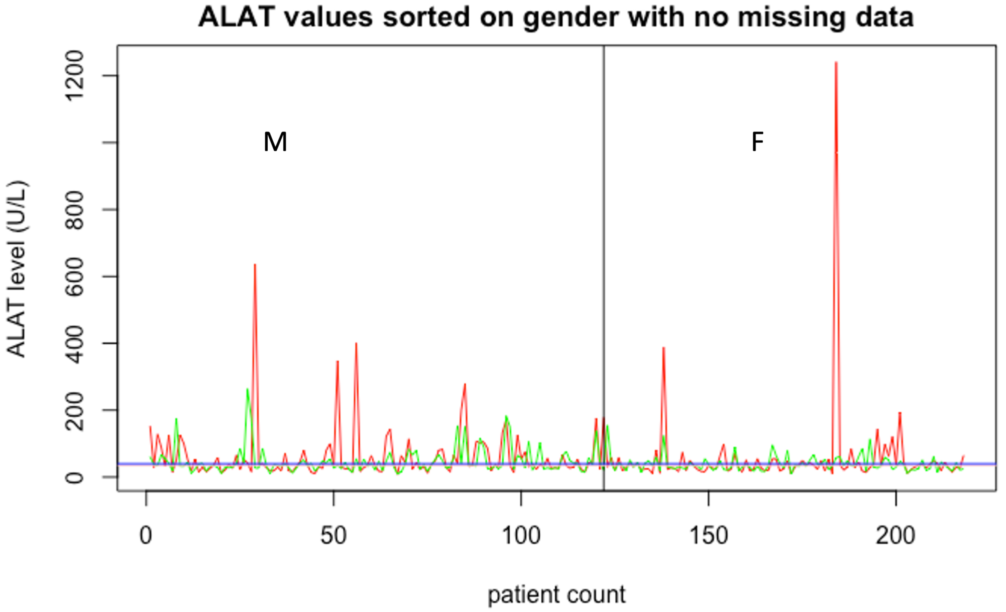


Figure B.120: ALAT levels (U/L) before and after the first chemotherapy cycle sorted by gender with no missing values, n=218 (121 male, 97 female). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

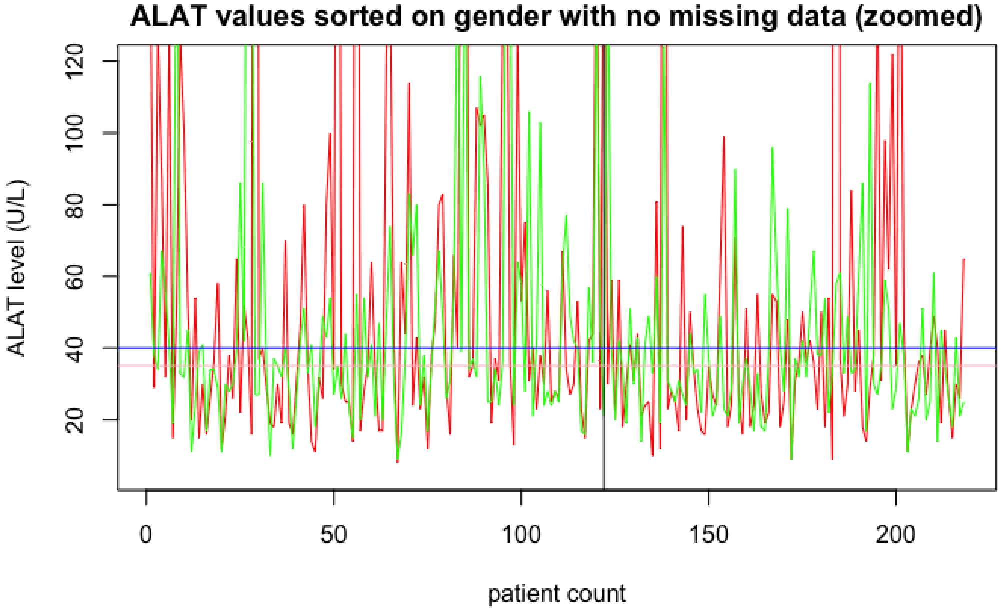


Figure B.121: Same plot as Figure B.120 but the y-axis is restricted from 5-120 U/L.

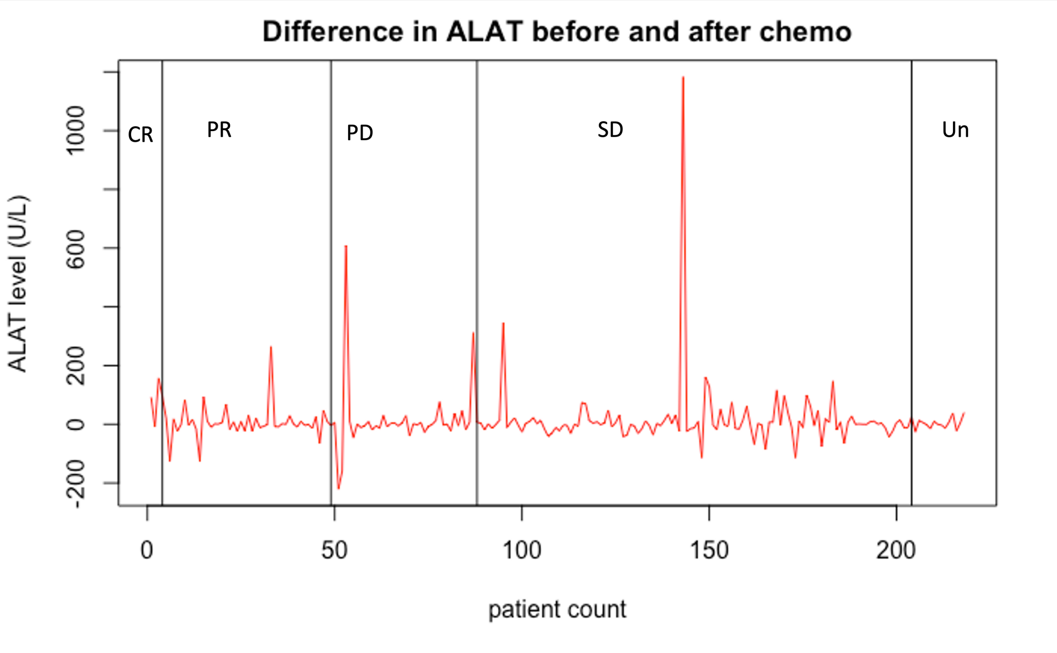


Figure B.122: Difference in ALAT values (U/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=218. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=45), Progressive Disease (PD, n=39), Stable Disease (SD, n=116), Unknown (Un, n=15).

B.1.3.12 Alkaline Phosphatase (AF)

Alkaline Phosphatase (AF) is another enzymatic measure that provides insight into liver function, similarly to ASAT and ALAT. AF is primarily found in the liver, as well as in bones, intestines, and kidneys. High levels of AF may indicate liver or gallbladder conditions, or other health issues. A healthy AF value is considered to be below 125 U/L.

In a similar manner to the ALAT and ASAT values, elevated AF values are also commonly observed in patients with PDAC, as evidenced by the plots and tables presented previously. The distribution of AF values before and after chemotherapy is similarly left-skewed with a long right tail, dominated by high AF values. The QQplot, histograms, and p-values depicted in Figure B.125, Figure B.123, Figure B.124, and Table B.40 further highlight a clear deviation from normality caused by the outlier values. After removal of missing values, the resulting dataset includes 214 observations, with 118 males and 96 females. The final response comprises 3 patients with complete response (CR), 45 with partial response (PR), 39 with progressive disease (PD), 113 with stable disease (SD), and 14 unknown cases. The plot in Figure B.126 clearly shows that the AF values exceed the healthy range, while the plot in Figure B.127 indicates no apparent correlation between these values and the final response to chemotherapy. In general, the plot in Figure B.129 shows that AF values tend to decrease slightly after chemotherapy. However, it varies significantly from patient to patient.

AF Value	*Before 1st Chemo*	*After 1st Chemo*
Min	25.0	25.0
1st Quartile	91.0	111.0
Median	128.0	149.0
Mean	171.5	176.4
3rd Quartile	197.5	200.8
Max	947.0	901.0
NA	9	21

Table B.39: Summary statistics values of the alkaline phosphatase values of the entire dataset in U/L, n=247.

AF value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.660	0.707
SW-test pvalue	<2.2e-16	<2.2 e-16
KS test pvalue	Invalid	Invalid

Table B.40: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the alkaline phosphatase values, n=247. The KS-test is invalid due to the presence of ties in the data.

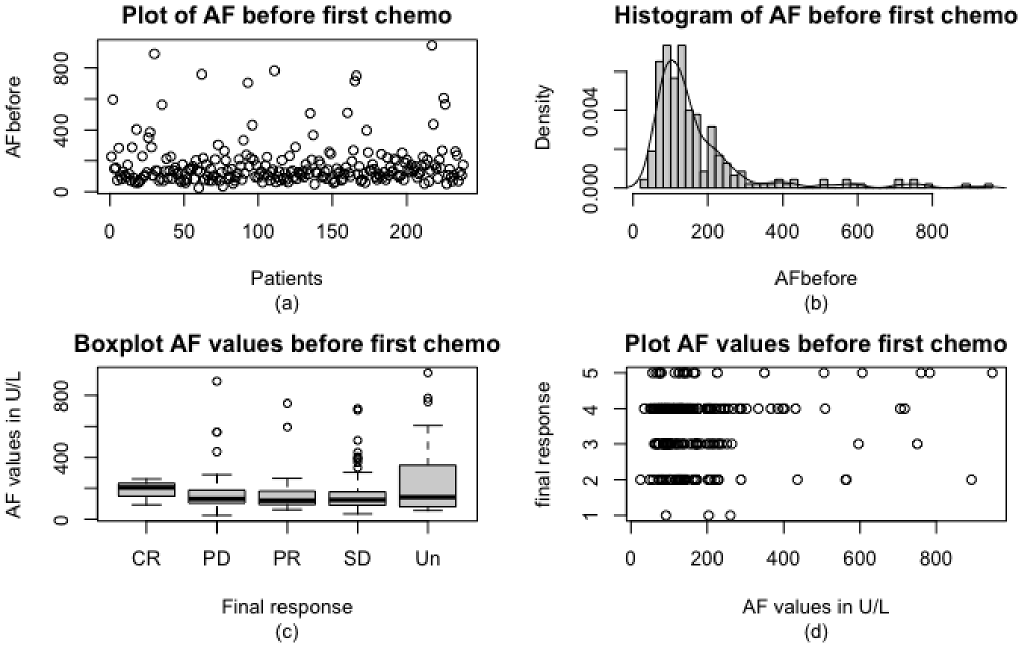


Figure B.123: Alkaline Phosphatase distribution before the first chemotherapy cycle ($10^9/L$) for the entire dataset: (a) scatterplot of alkaline phosphatase values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of alkaline phosphatase values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

AF values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	25.0	25.0	-485.0
1st Quartile	91.0	111.0	-65.0
Median	126.0	148.5	-17.0
Mean	162.4	176.8	-14.4
3rd Quartile	175.8	199.5	22.0
Max	947.0	901.0	362.0

Table B.41: Summary statistics of the alkaline phosphatase levels in U/L before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=214.

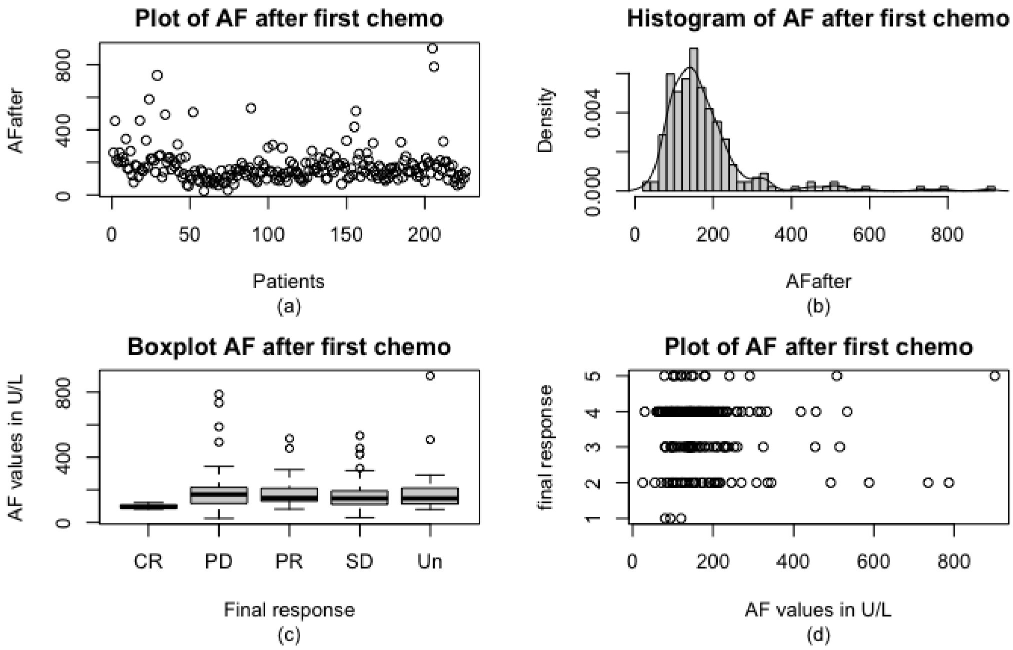


Figure B.124: Alkaline Phosphatase distribution after the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of alkaline phosphatase values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of alkaline phosphatase values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

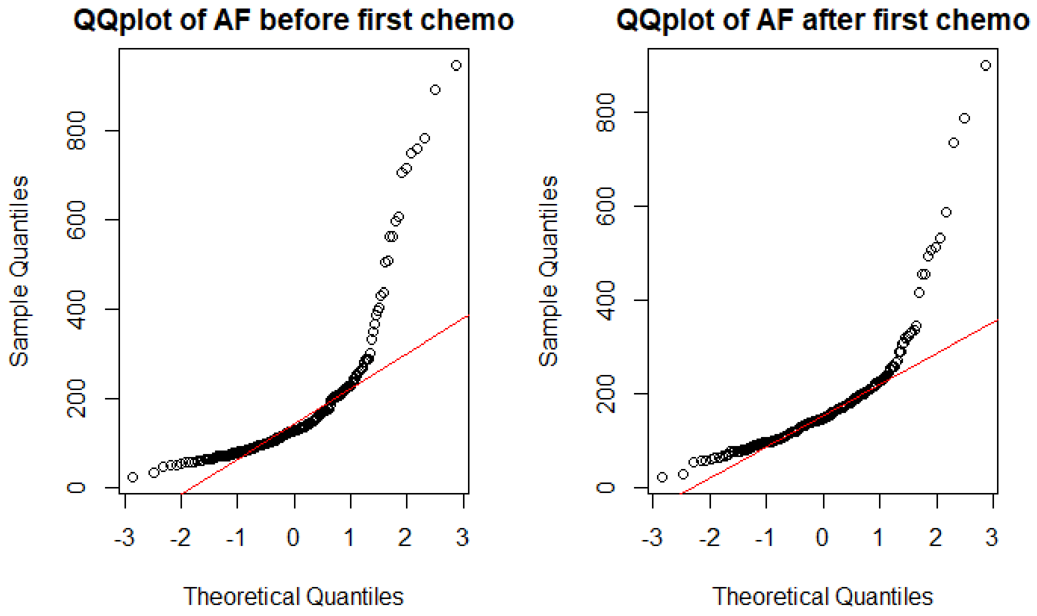


Figure B.125: QQplot of the alkaline phosphatase values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

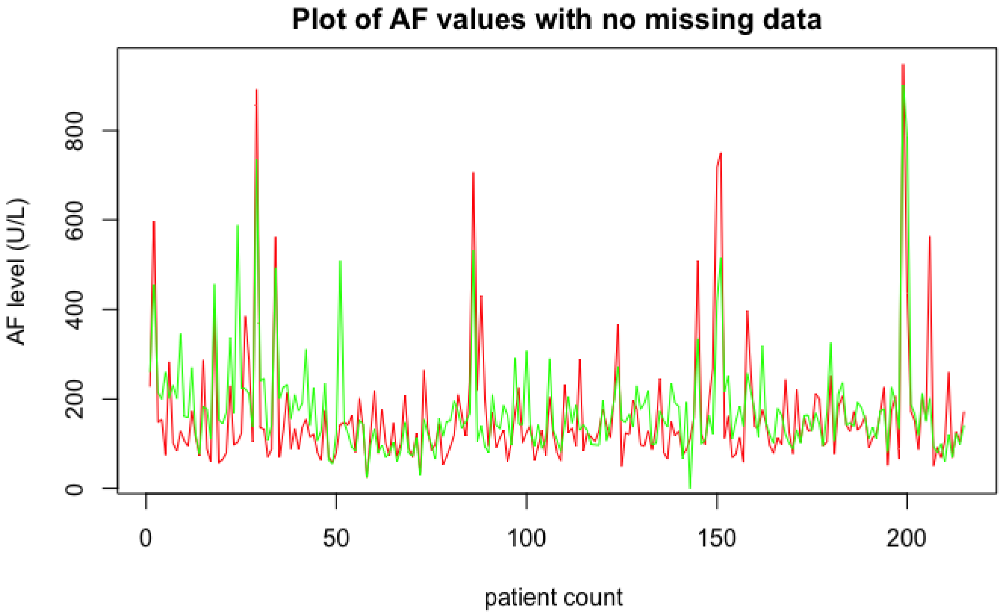


Figure B.126: Alkaline Phosphatase levels (U/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=214.

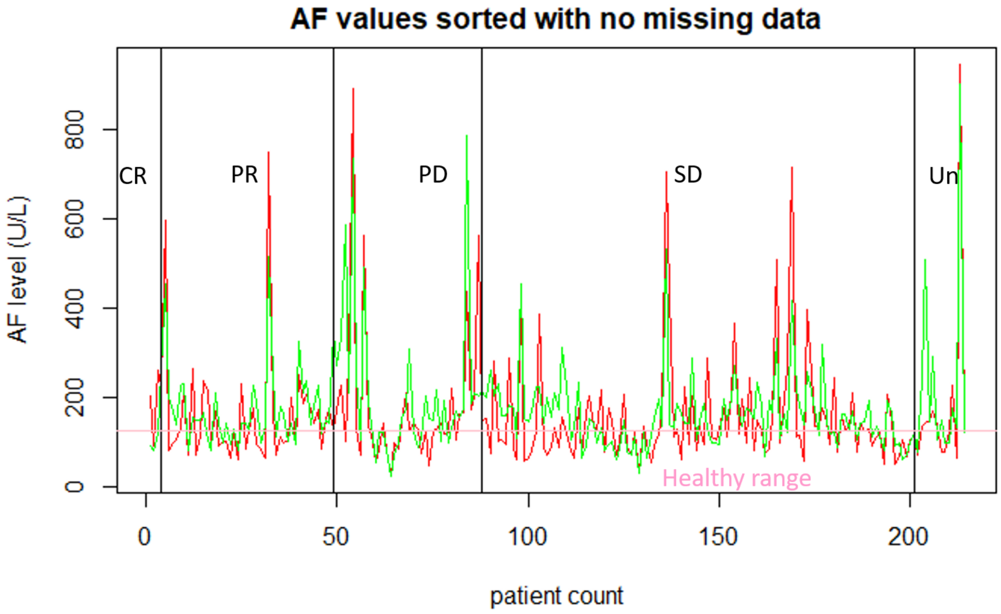


Figure B.127: Alkaline phosphatase levels (U/L) (U/L) before and after the first chemotherapy cycle sorted by final response with no missing values, n=214. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=45), Progressive Disease (PD, n= 39), Stable Disease (SD, n=113), Unknown (Un, n=14). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

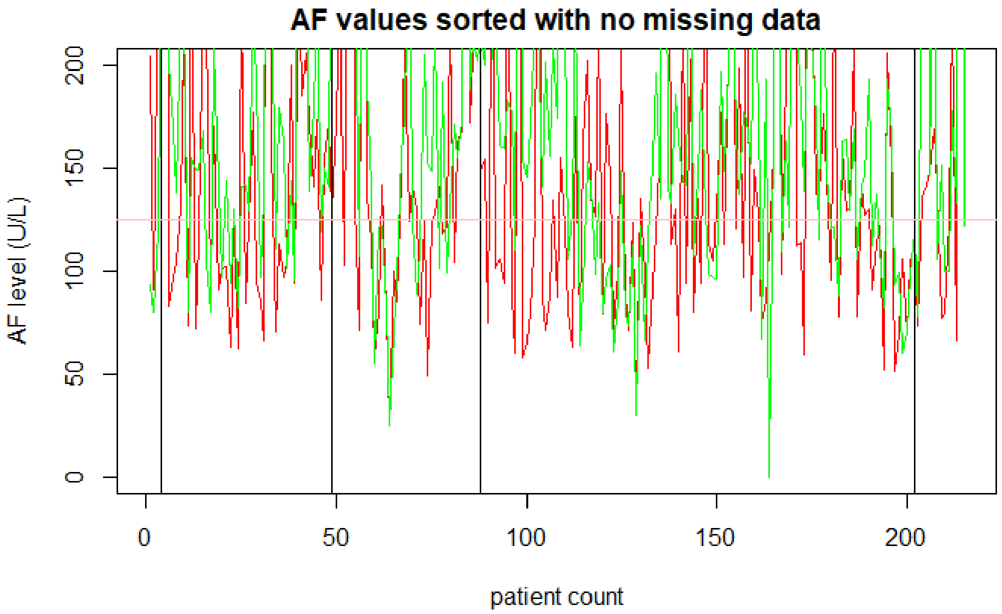


Figure B.128: Same plot as Figure B.127 but with AF values restricted between 0 and 200 U/L.

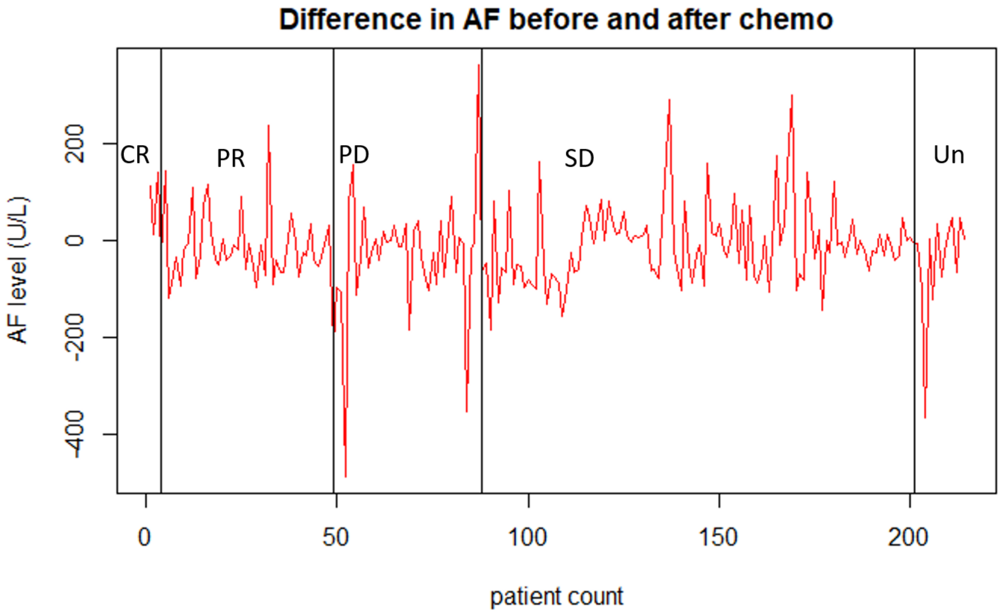


Figure B.129: Difference in alkaline phosphatase values (U/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=214. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=45), Progressive Disease (PD, n= 39), Stable Disease (SD, n=113), Unknown (Un, n=14).

B.1.3.13 Gamma-Glutamyltransferase (GGT)

In line with the previous liver function tests, namely ALAT, ASAT, and AF, Gamma-glutamyltransferase (GGT) also provides information on liver function. GGT is a liver enzyme that is involved in the metabolism of glutathione, a molecule that helps protect cells from damage. It is also responsible for breaking down food, drinks, and waste materials. Elevated levels of GGT in the blood are a marker of liver dysfunction and can be an indicator of various liver diseases, including alcoholic liver disease, nonalcoholic fatty liver disease, and viral hepatitis. However, high levels of GGT can also be caused by other factors such as obesity, diabetes, and certain medications. According to medical experts from the Erasmus Medical Centre Rotterdam, for males, a healthy value is below 40 U/L, while for females, a healthy value is below 25 U/L. As a side note, patient 002PP20052 in the dataset had a reported GGT value of 10229.0 U/L, which is likely a typographical error and should be 1029 U/L. It is highly improbable to observe a value above 10000 U/L. Therefore, this high value has been adjusted to 1029 U/L before the start of the analysis.

Alcohol consumption is a well-known cause of elevated GGT levels. Chronic alcohol use can lead to alcoholic liver disease, which is characterized by inflammation and scarring of the liver tissue. GGT levels are commonly used as a marker of alcohol consumption, and elevated levels can indicate excessive drinking even in the absence of liver damage. However, it should be noted that other factors, such as obesity and certain medications, can also cause elevated GGT levels, so GGT alone cannot be used as a definitive marker of alcohol use. In the context of PDAC the relationship between GGT, alcohol use, and treatment is complex. While chronic alcohol use is a risk factor for the development of PDAC, it is unclear whether alcohol use has a direct effect on GGT levels in PDAC patients. Additionally, the effects of chemotherapy on GGT levels in PDAC patients are not well understood, and it is possible that chemotherapy could cause temporary elevations in GGT levels due to liver damage [107].

Similar to the previous observations of the liver enzymes, most GGT values seem to be scattered within the same range with a couple of outliers causing a left-skewed histogram and an extreme long tail, as seen in Figure B.130 and Figure B.131. There seems to be no distinction between GGT values and the final response either. Furthermore, the QQplot as well as the values reported in Table B.43 reject normality of the GGT-values. When the NA values are removed from the dataset, a total of 191 observations remain, with 102 male and 89 female patients. Among them, 3 patients have complete response as the final outcome, 39 have partial response, 34 have progressive disease, 104 have stable disease, and 11 have unknown values. From the plots given in Figure B.133, Figure B.134 and Figure B.135, it can be inferred that PDAC patients generally have GGT values above the healthy range for both males and females. In addition, there seems to be no clear distinguishment across the final response groups. Sorting the GGT-values based on gender, as can be seen in Figure B.136 and the zoomed plot in Figure B.137, no separation between the two genders can clearly be detected based on the GGT-values. The difference in GGT-values is also highly dependent on the individual, as seen in Figure B.138.

Given the significant influence of alcohol on GGT levels, it might be insightful to examine the relationship between alcohol use and GGT levels among PDAC patients. This analysis is presented in Figure B.139 and Figure B.140. However, the results do not indicate any discernible trend between these variables.

GGT Value	*Before 1st Chemo*	*After 1st Chemo*
Min	13.0	15.0
1st Quartile	37.0	38.8
Median	92.0	74.0
Mean	230.7	127.1
3rd Quartile	206.0	142.5
Max	2856.0	1230.0
NA	26	39

Table B.42: Summary statistics values of the GGT values of the entire dataset in U/L, n=247.

GGT value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.213	0.602
SW-test pvalue	<2.2e-16	<2.2 e-16
KS test pvalue	Invalid	Invalid

Table B.43: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the GGT values, n=247. The KS-test is invalid due to the presence of ties in the data.

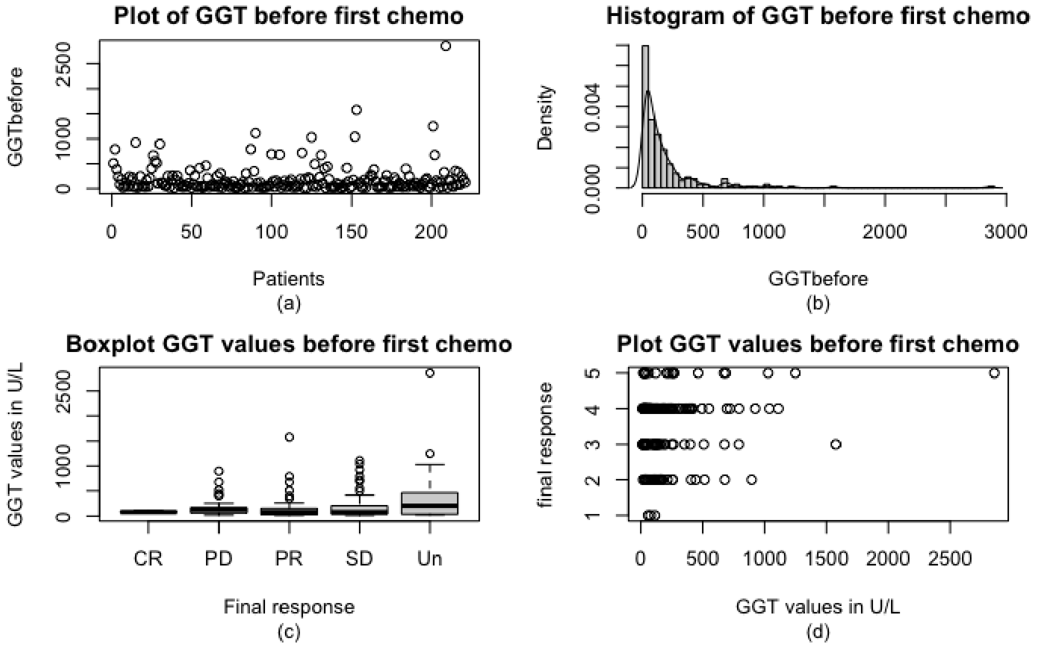


Figure B.130: GGT distribution before the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of GGT values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of GGT values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

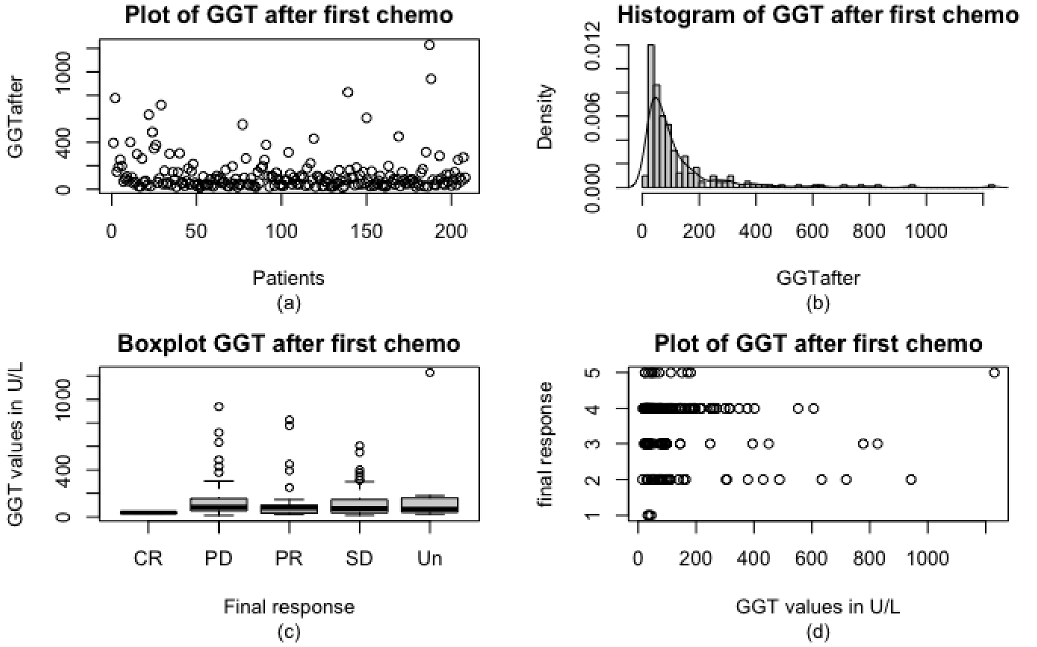


Figure B.131: GGT distribution after the first chemotherapy cycle (U/L) for the entire dataset: (a) scatterplot of GGT values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of GGT values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

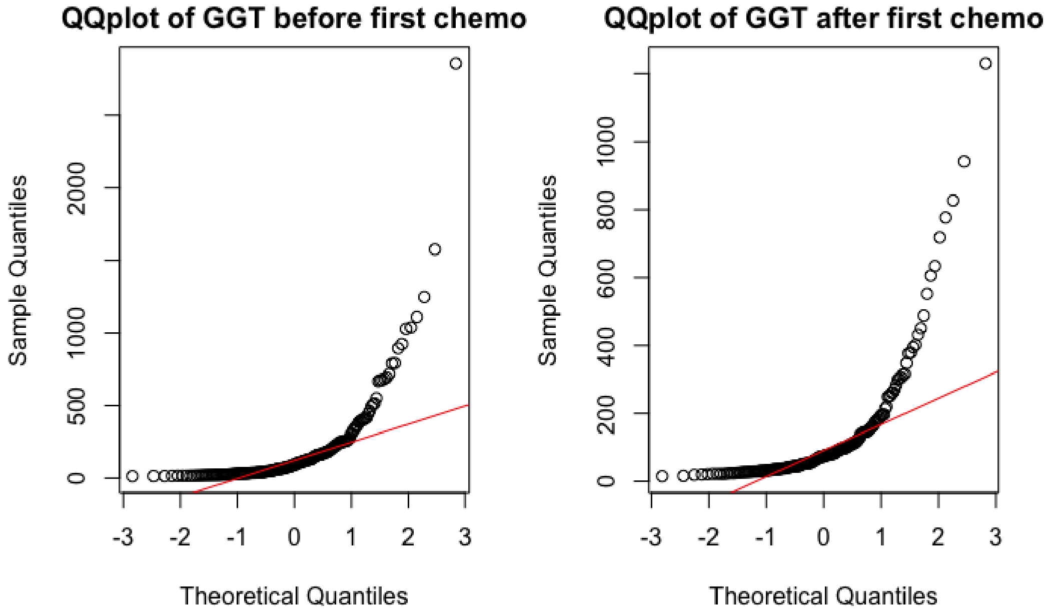


Figure B.132: QQplot of the GGT values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

GGT values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	13.0	15.0	-463.0
1st Quartile	36.5	38.0	-16.0
Median	88.0	75.0	8.0
Mean	164.9	129.8	35.1
3rd Quartile	187.0	145.0	54.5
Max	1576.0	1230.0	848.0

Table B.44: Summary statistics of the GGT levels in U/L before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=191.

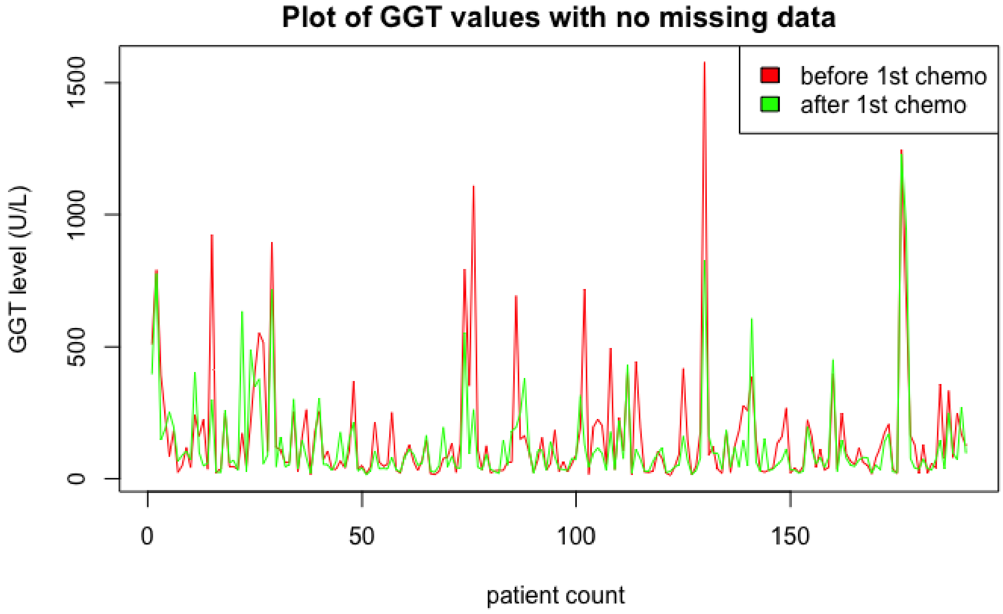


Figure B.133: GGT values (U/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=191.

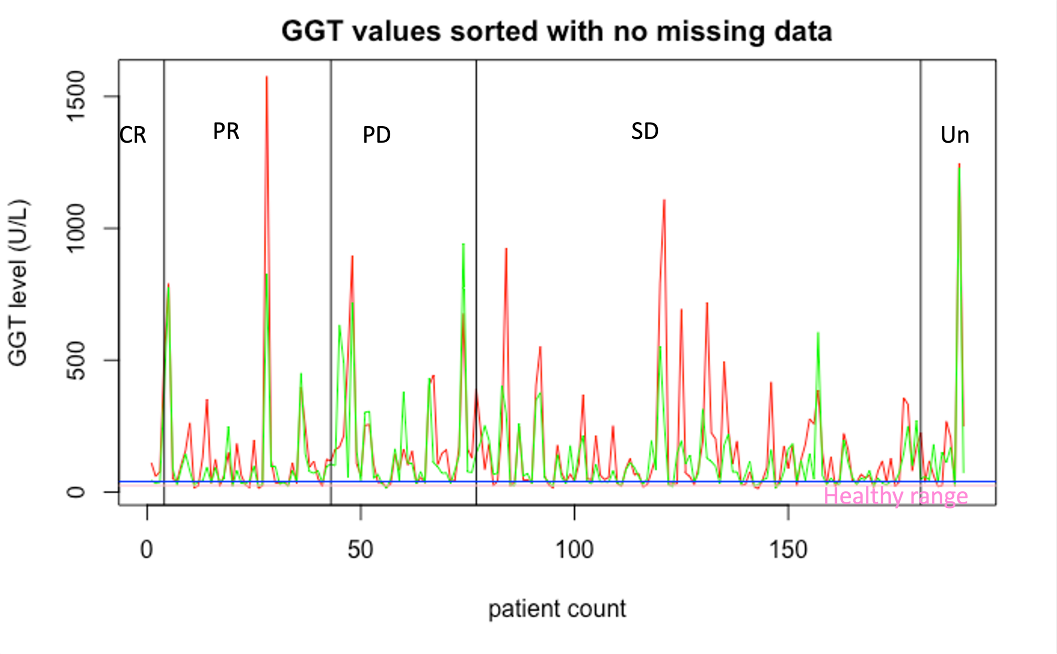


Figure B.134: GGT levels (U/L) before and after the first chemotherapy cycle sorted by final response with no missing values, n=191. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=39), Progressive Disease (PD, n= 34), Stable Disease (SD, n=104), Unknown (Un, n=11). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

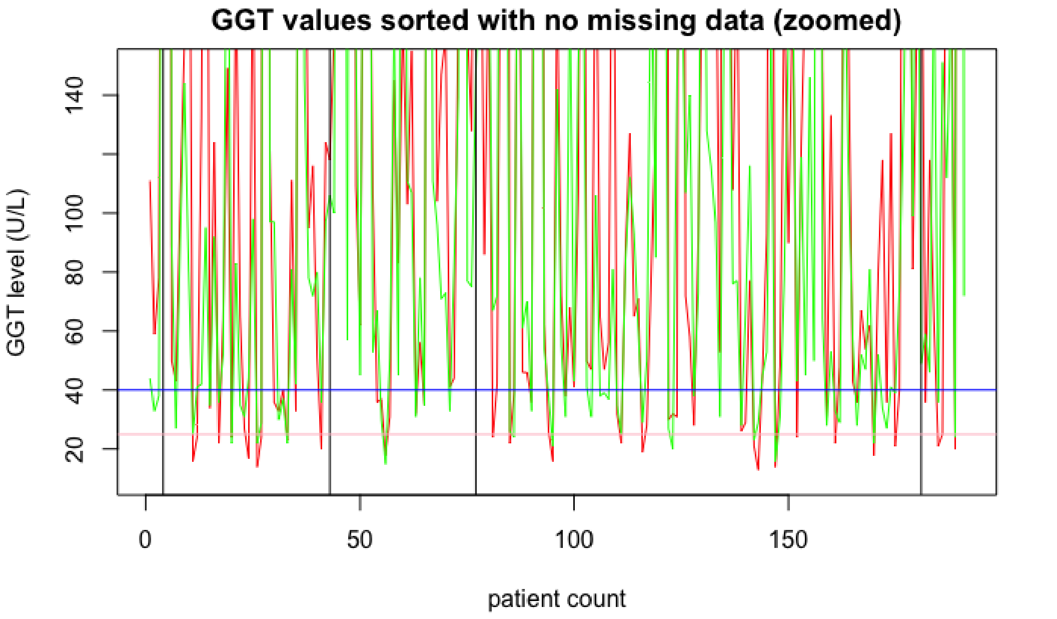


Figure B.135: Same plot as Figure B.134 but the y-axis limited from 10-150 U/L.

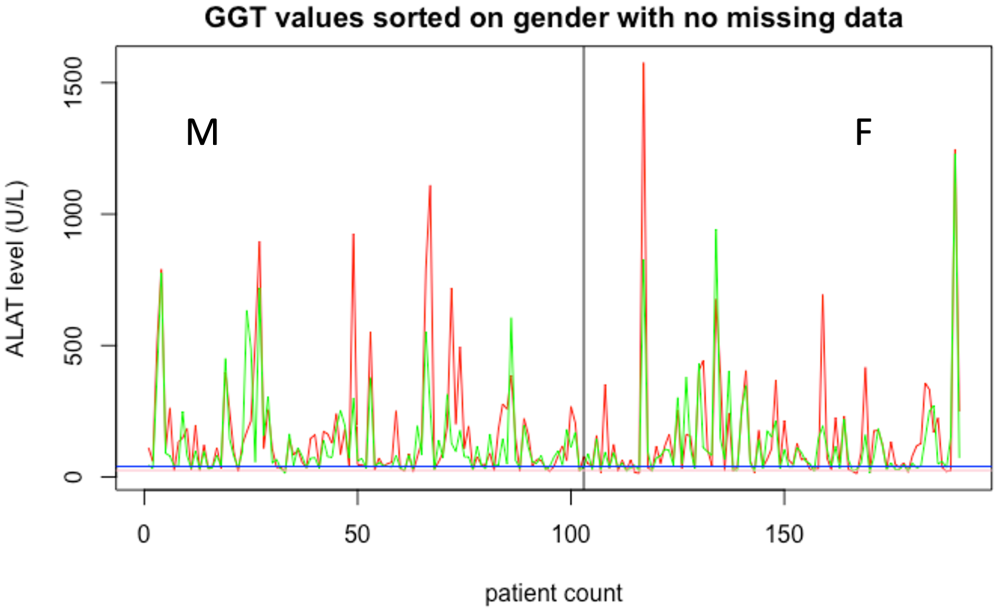


Figure B.136: GGT levels (U/L) before and after the first chemotherapy cycle sorted by gender with no missing values, n=191 (102 male, 89 female). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

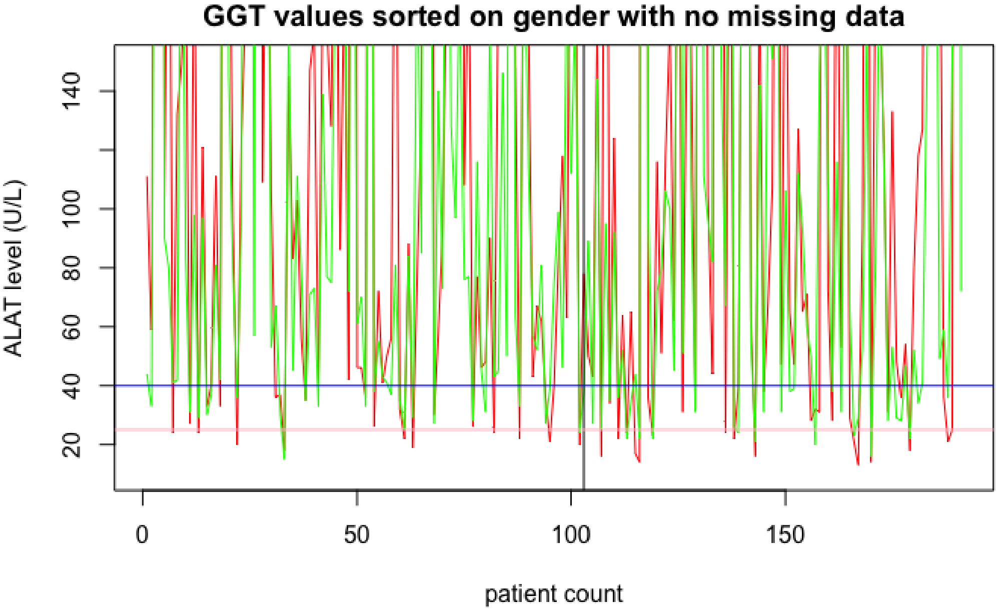


Figure B.137: Same plot as Figure B.136 but the y-values restricted from 10-150 U/L.

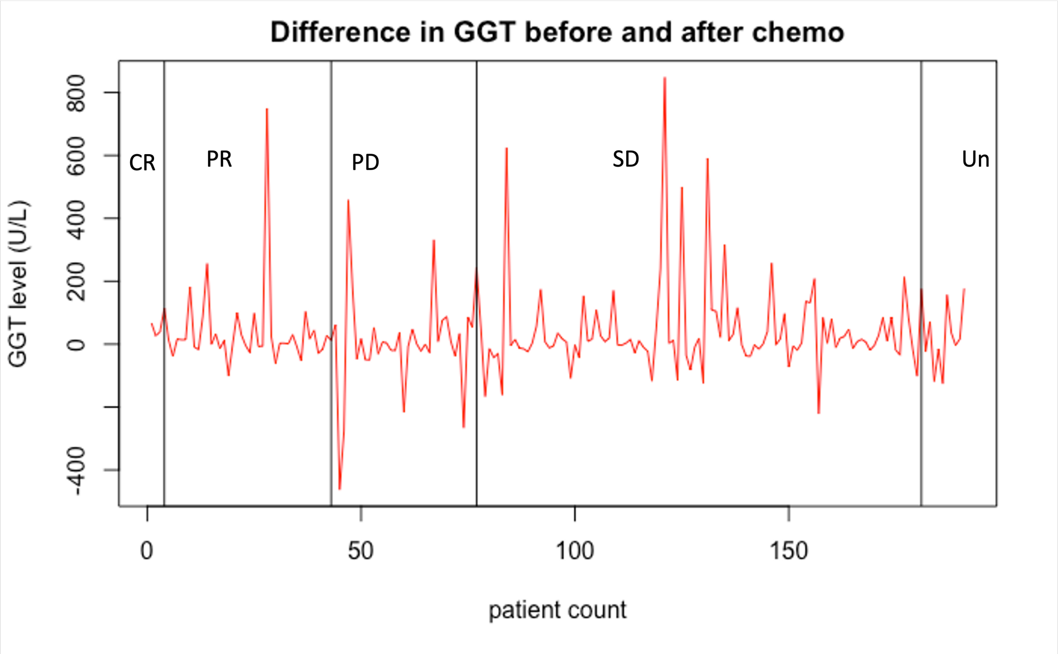


Figure B.138: Difference in GGT values (U/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response,n=191. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=39), Progressive Disease (PD, n= 34), Stable Disease (SD, n=104), Unknown (Un, n=11).

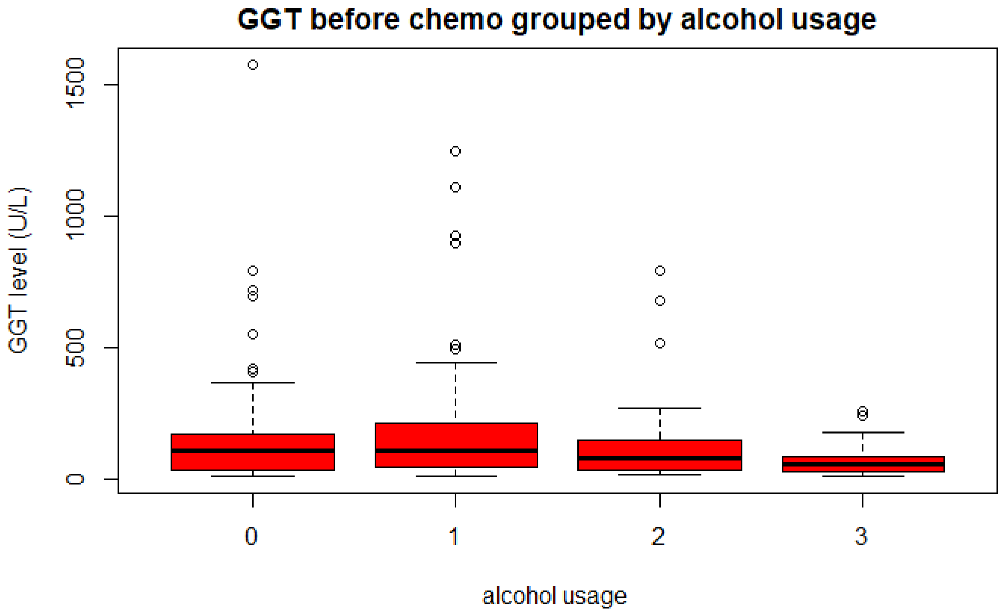


Figure B.139: Boxplot of GGT levels (U/L) before the first chemotherapy cycle sorted by alcohol with no missing values, n=191. The alcohol usage is classified as 0= No (n=63), 1 = Yes (n=83), 2 = Stopped (n=31), 3 = Unknown (n=14)

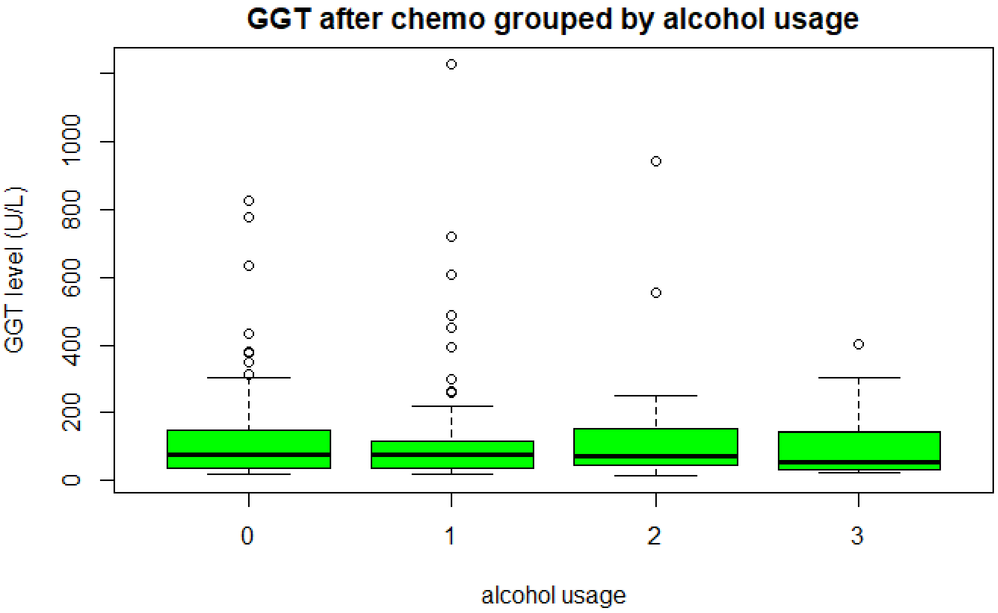


Figure B.140: Boxplot of GGT levels (U/L) after the first chemotherapy cycle sorted by alcohol with no missing values, n=191. The alcohol usage is classified as 0= No (n=63), 1 = Yes (n=83), 2 = Stopped (n=31), 3 = Unknown (n=14)

B.1.3.14 Bilirubin (BR)

Bilirubin is a byproduct of the breakdown of red blood cells and is present in bile, which is synthesized by the liver to aid in digestion. Elevated levels of bilirubin can be indicative of liver or bile duct dysfunction, as a healthy liver is capable of efficiently removing most bilirubin from the body. Although CA19-9 is a well-studied tumor marker in PDAC patients, roughly 10% of the population is incapable of producing this marker, making it necessary to also measure bilirubin levels. When assessing the relationship between CA19-9 and bilirubin levels, a high level of bilirubin alone or in conjunction with a high level of CA19-9 does not provide significant information about the cancer. Only individuals with normal CA19-9 levels and elevated bilirubin levels can have conclusions drawn regarding cancer. A bilirubin level under 17 μmol/L is generally considered healthy for adults according to medical experts from the Erasmus Medical Centre Rotterdam.

The summary statistics presented in Table B.45 show that the bilirubin levels tend to be high before the first chemotherapy treatment, with an average of 18.0 μ mol/L and an extremely high maximum value of 132.0 μ mol/L. However, after the first chemotherapy treatment, a significant decrease in the bilirubin values is observed, as evidenced by all statistical measures, with the maximum value being 30.0 μ mol/L. This indicates that chemotherapy has a significant effect on bilirubin values in PDAC patients. The scatterplots, histogram, and boxplots presented in Figure B.141 and Figure B.142 illustrate that the distribution of bilirubin values is left-skewed and non-normal. This finding is further supported by the p-values presented in Table B.46 and the QQplots provided in Figure B.143. Additionally, the boxplots suggest that partial responders to chemotherapy tend to have lower bilirubin values in general, although the difference from the other groups is minimal and conclusive statements cannot be made.

To continue, examine the dataset without missing values, consisting of 219 observations, with 123 male and 96 female patients. The final response variable is distributed among 3 patients with complete response (CR), 45 with partial response (PR), 41 with progressive disease (PD), 115 with stable disease (SD), and 15 Unknown values. The visual representations of the data, given in Figure B.144 and Figure B.145 suggest that the majority of bilirubin (BR) values before chemotherapy are above healthy levels, whereas after chemotherapy, they significantly decrease towards a more optimal range. Moreover, the difference in bilirubin values, shown in Figure B.146, seem to be smaller in the partial responders to chemotherapy compared to the other groups. Nevertheless, clear conclusions cannot be drawn.

BR Value	*Before 1st Chemo*	*After 1st Chemo*
Min	2.0	2.0
1st Quartile	7.0	5.0
Median	14.0	7.0
Mean	18.0	8.8
3rd Quartile	24.5	11.8
Max	132.0	30.0
NA	8	17

Table B.45: Summary statistics values of the bilirubin values of the entire dataset in $\mu mol/L$, n=247.

BR value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.731	0.901
SW-test pvalue	<2.2e-16	3.641e-11
KS test pvalue	Invalid	Invalid

Table B.46: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the bilirubin values, n=247. The KS-test is invalid due to the presence of ties in the data.

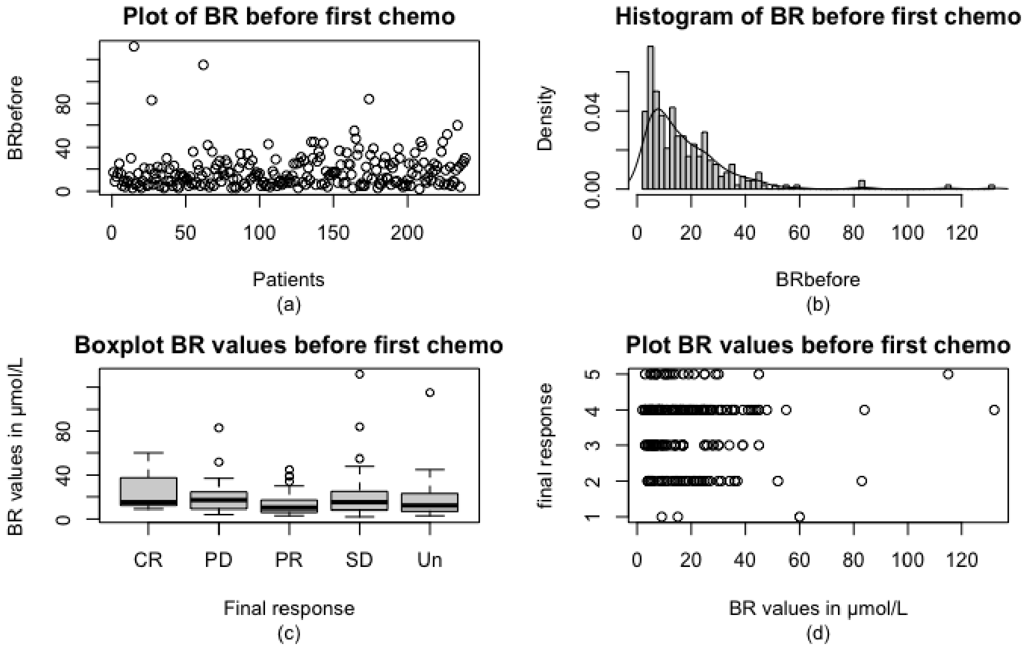


Figure B.141: Bilirubin distribution before the first chemotherapy cycle ($\mu mol/L$) for the entire dataset: (a) scatterplot of bilirubin values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of bilirubin values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

BR values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	2.0	2.0	-21.0
1st Quartile	7.0	5.0	1.0
Median	14.0	8.0	5.0
Mean	17.7	9.0	8.7
3rd Quartile	24.5	12.0	12.8
Max	132.0	30.0	118.0

Table B.47: Summary statistics of the bilirubin values in $\mu mol/L$ before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=219.

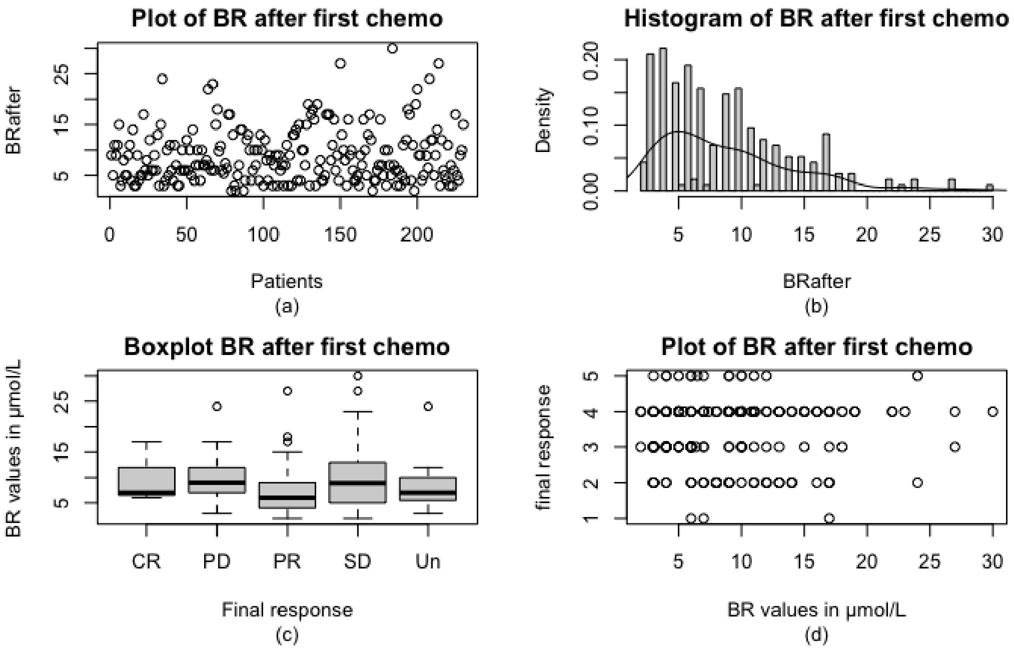


Figure B.142: Bilirubin distribution after the first chemotherapy cycle ($\mu mol/L$) for the entire dataset: (a) scatterplot of bilirubin values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of bilirubin values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

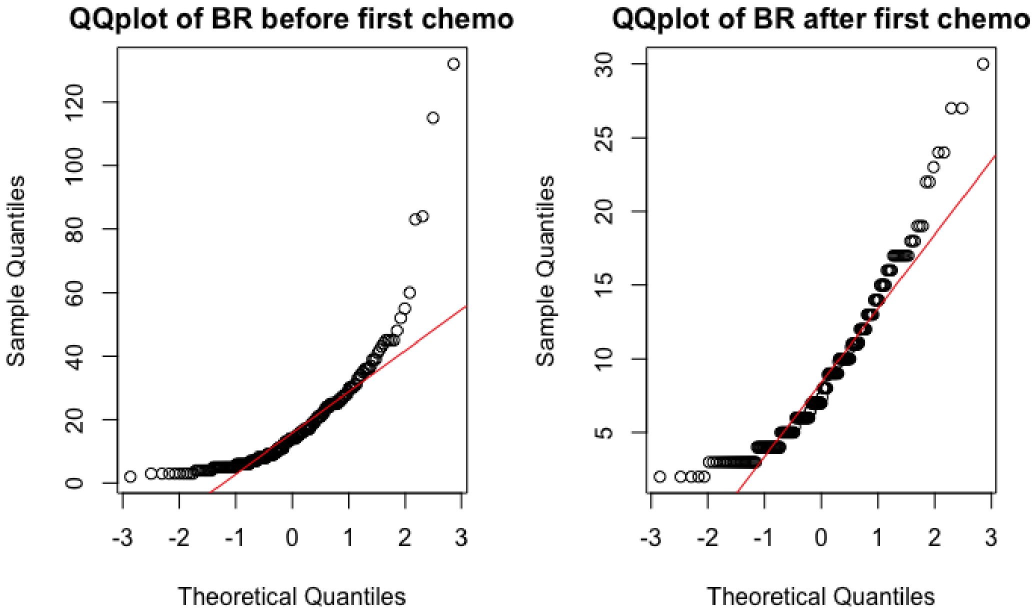


Figure B.143: QQplot of the bilirubin values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

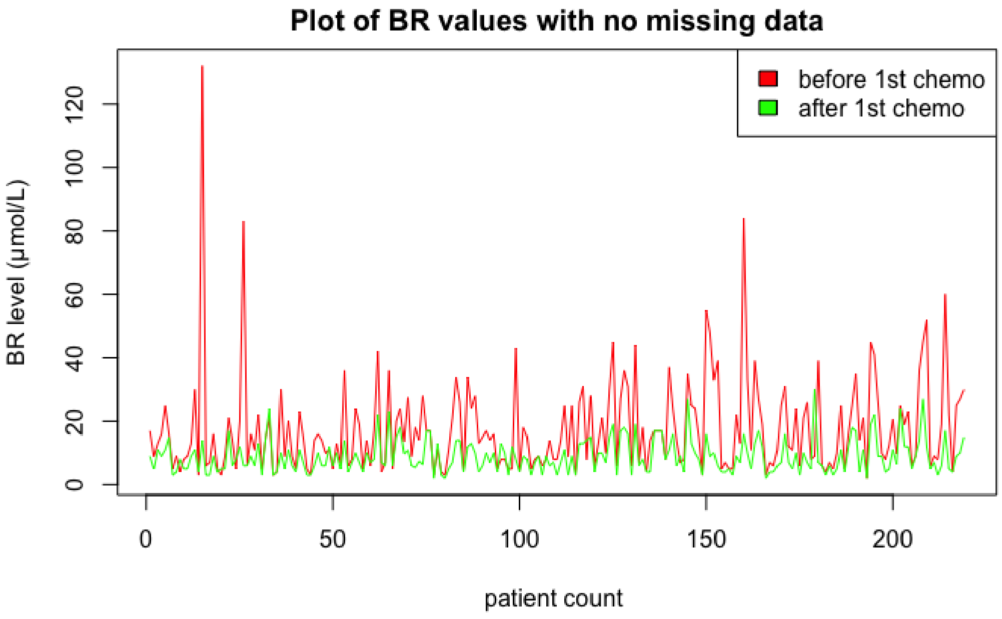


Figure B.144: Bilirubin levels ($\mu mol/L$) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=219.

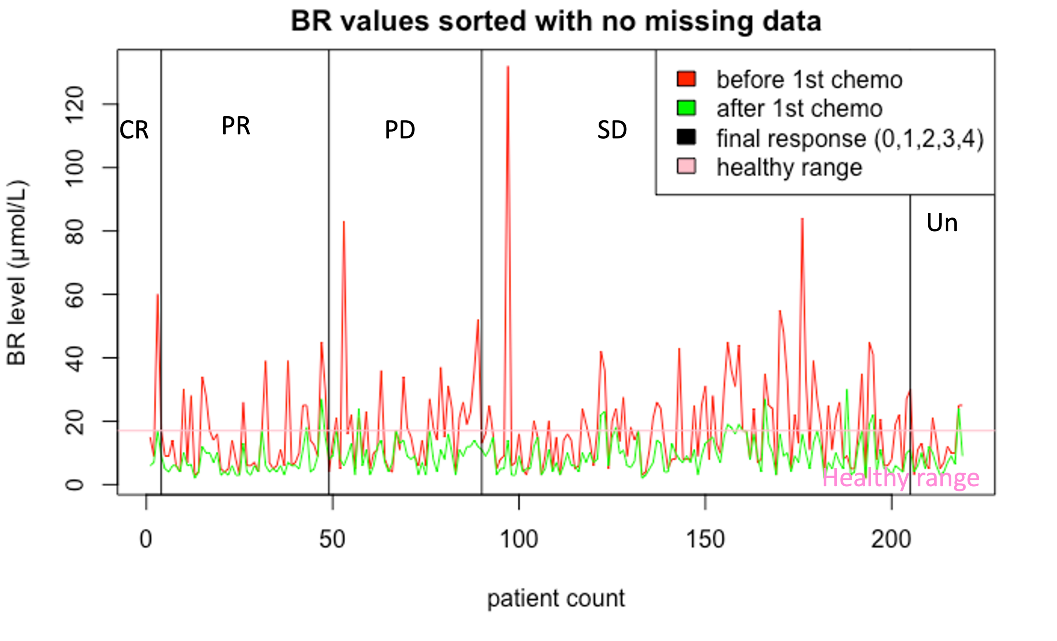


Figure B.145: Bilirubin levels ($\mu mol/L$) before and after the first chemotherapy cycle sorted by final response with no missing values, n=219. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=45), Progressive Disease (PD, n= 41), Stable Disease (SD, n=115), Unknown (Un, n=15). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

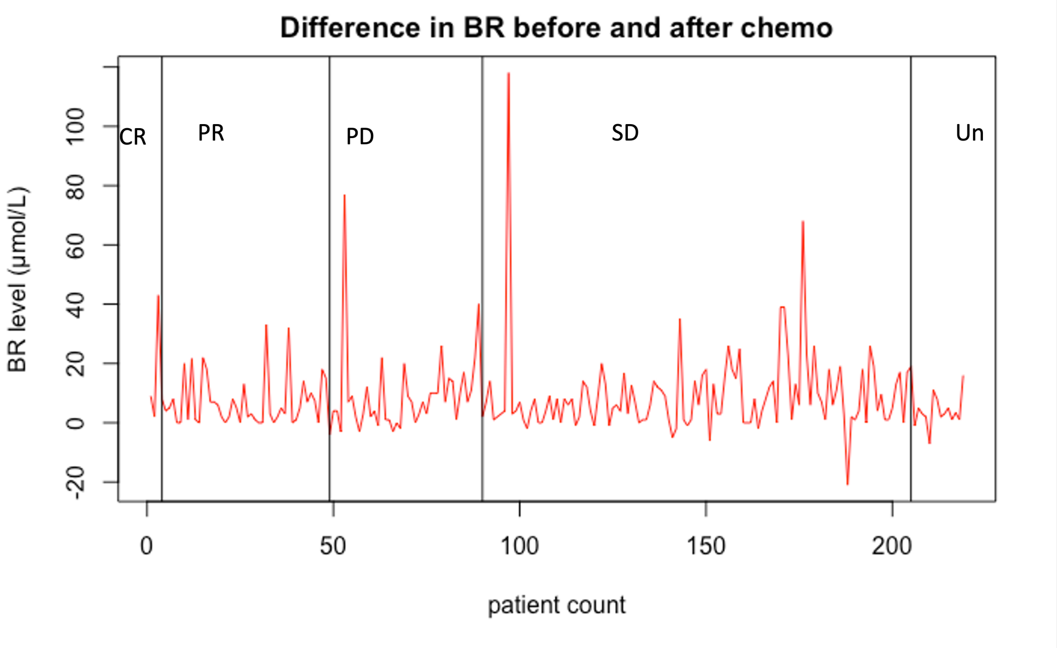


Figure B.146: Difference in bilirubin values ($\mu mol/L$)between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=219. The final response is classified as Complete Response (CR, n=3), Partial response (PR, n=45), Progressive Disease (PD, n= 41), Stable Disease (SD, n=115), Unknown (Un, n=15).

B.1.3.15 Albumin Alb

Albumin is a crucial protein synthesized by the liver, with two primary functions. Firstly, it contributes to the colloid osmotic pressure, which aids in maintaining fluid balance between the blood and tissues. Secondly, albumin is involved in the transportation, distribution and metabolism of endogenous and exogenous molecules, such as fatty acids, thyroid hormones, metals, peptides, among others. Inadequate levels of albumin in the blood can lead to fluid leakage from the blood into the body's cavities, such as the lungs or abdomen. Low albumin levels can signify health conditions affecting the liver, kidneys, or other organs, while excessively high levels are often a result of dehydration. The healthy range for albumin levels is typically between 35 and 55 g/L according to medical experts. The dataset used for analysis included several entries with an albumin level of 0.0 g/L, which is likely to be missing data. Therefore, all albumin levels with a value of 0.0 g/L were converted to NA for further analysis.

The tabulated summary statistics presented in Table Table B.48 indicate a slight decrease in albumin values following the first chemotherapy treatment. The observed decline is approximately 3.0g/L across most statistics, with the exception of the maximum value, which increased from 49.0 g/L to 64.0 g/L. However, this increase may be attributed to a patient with pre-existing liver disease or other health conditions. Further insights into the distribution of albumin values before and after chemotherapy can be obtained from the scatter plots, histograms, and boxplots depicted in Figures B.147 and B.148. Prior to chemotherapy, albumin values were mostly concentrated within the 25-45 g/L range, while post-treatment, a wider range of values were observed, including some extremely high and low outliers. Nevertheless, there appears to be no discernible pattern in the ultimate response to chemotherapy based on albumin levels before or after the initial session. Additionally, the QQplots displayed in Figure B.149 suggest that albumin levels exhibit some normality, albeit with deviations from normality in the tails due to the presence of outliers. Particularly, albumin levels following the first chemotherapy treatment appear to be reasonably normally distributed, with two clear outlier observations. However, the Shapiro-Wilk test results presented in Table B.49 reject normality due to the aforementioned outliers. KS-test remains invalid due to the presence of ties in the dataset.

Based on the dataset without missing values, the study includes n= 161 observations (83 male, 78 female). The final response of the patients is categorized as 2 patients with CR, 36 with PR, 28 with PD, 84 with SD, and 11 unknown cases. The visualization of the dataset indicates that most of the albumin values lie within the healthy range, but there seems to be a trend towards the lower end of the healthy range. A distinct trend is observed in the plot depicted in Figure B.150 and Figure B.151, indicating that after chemotherapy, the albumin values tend to shift towards the lower healthier range as compared to before, with some patients falling below the healthy range and one patient above. The analysis of Albumin values before and after chemotherapy demonstrates the no clear differences in the final response groups, as illustrated in Figure B.152.

Alb Value	*Before 1st Chemo*	*After 1st Chemo*
Min	23.0	8.0
1st Quartile	36.0	33.0
Median	40.0	37.0
Mean	39.0	36.5
3rd Quartile	43.0	40.0
Max	49.0	64.0
NA	49	65

Table B.48: Summary statistics values of the albumin values of the entire dataset in g/L, n=247.

Alb value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.964	0.946
SW-test pvalue	5.36e-5	2.16e-6
KS test pvalue	Invalid	Invalid

Table B.49: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the albumin values, n=247. The KS-test is invalid due to the presence of ties in the data.

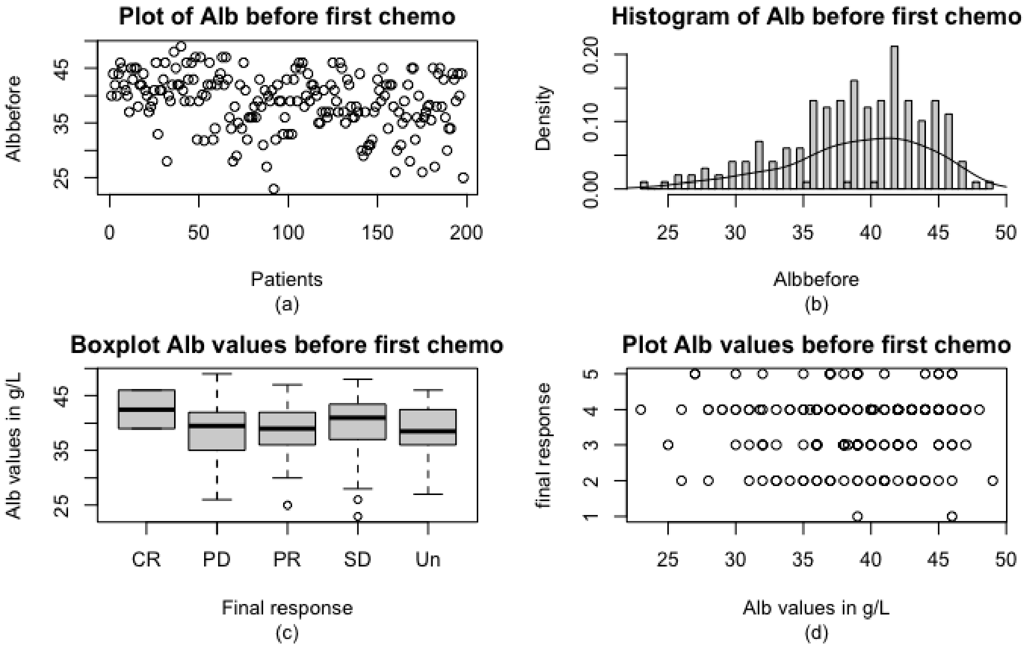


Figure B.147: Albumin distribution before the first chemotherapy cycle (g/L) for the entire dataset: (a) scatterplot of albumin values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of albumin values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

Alb values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	23.0	8.0	-23.0
1st Quartile	37.0	33.0	0.0
Median	40.0	37.0	2.8
Mean	39.3	36.5	3.5
3rd Quartile	43.0	40.0	5.0
Max	49.0	64.0	34.0

Table B.50: Summary statistics of the albumin levels in g/L before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=161.

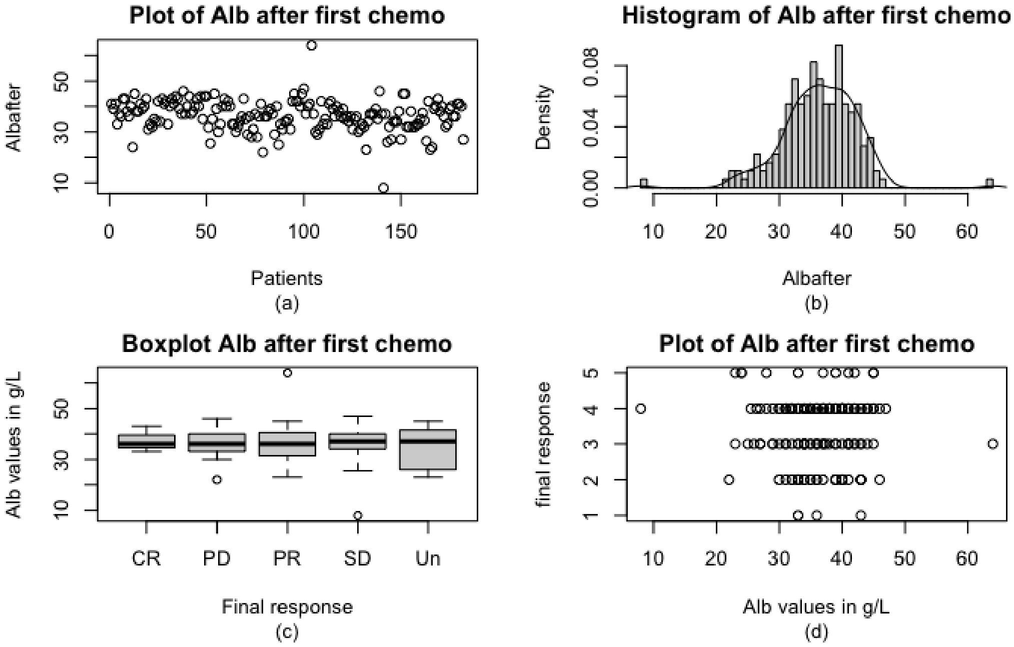


Figure B.148: Albumin distribution after the first chemotherapy cycle (g/L) for the entire dataset: (a) scatterplot of albumin values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of albumin values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

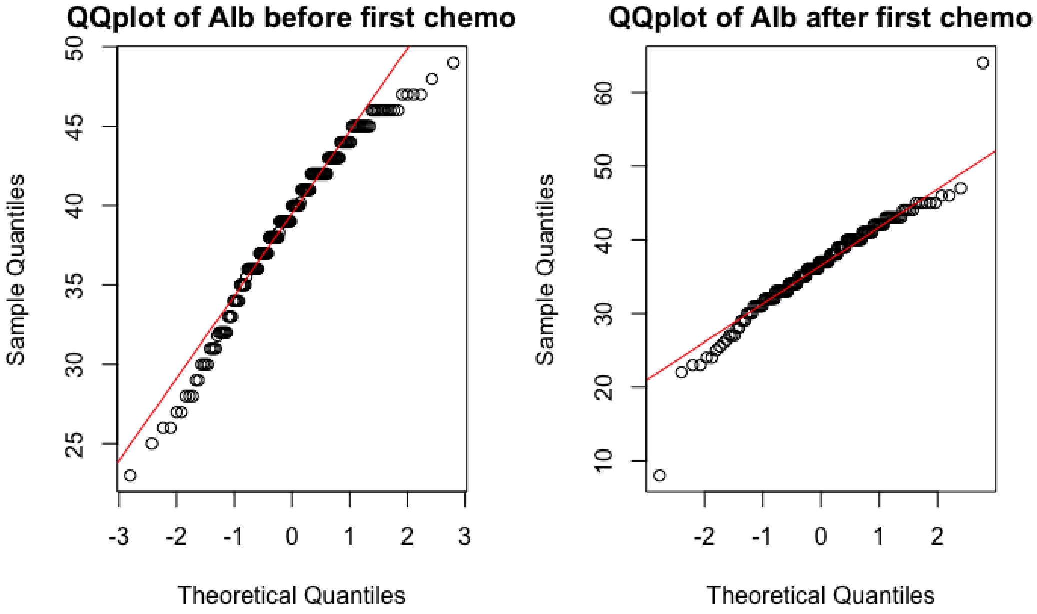


Figure B.149: QQplot of the albumin values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

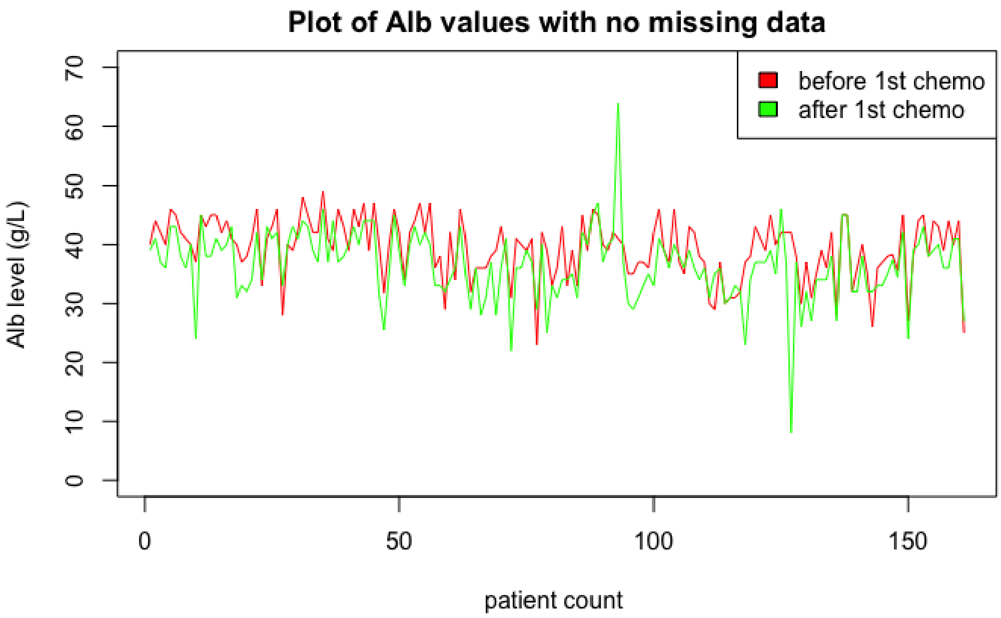


Figure B.150: Albumin levels (g/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=161.

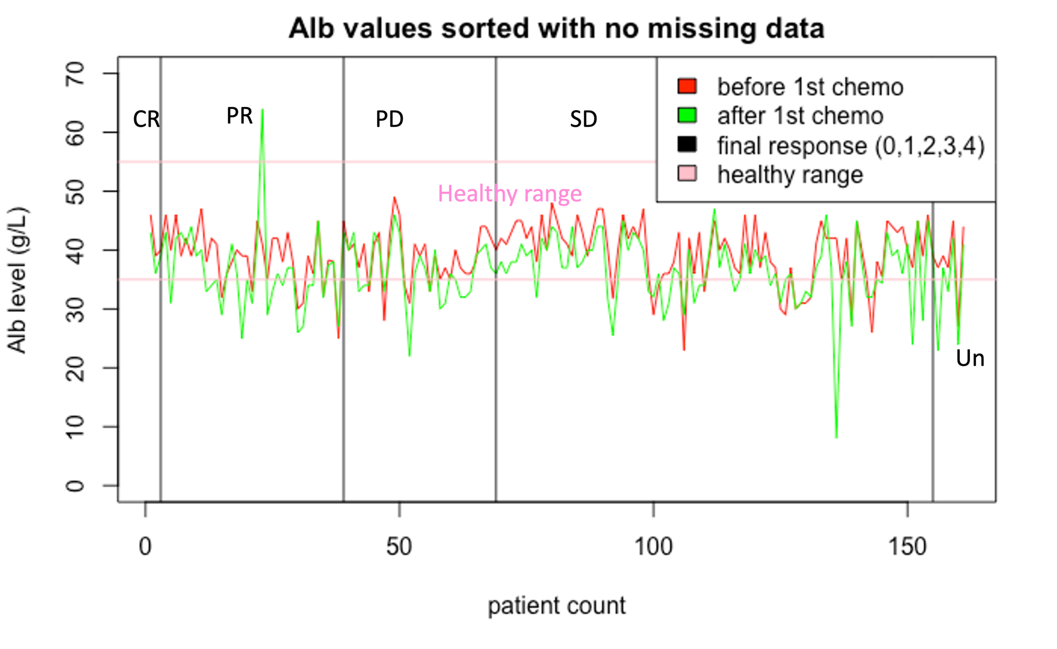


Figure B.151: Albumin levels (g/L) before and after the first chemotherapy cycle sorted by final response with no missing values, n=161. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=36), Progressive Disease (PD, n=28), Stable Disease (SD, n=84), Unknown (Un, n=11). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

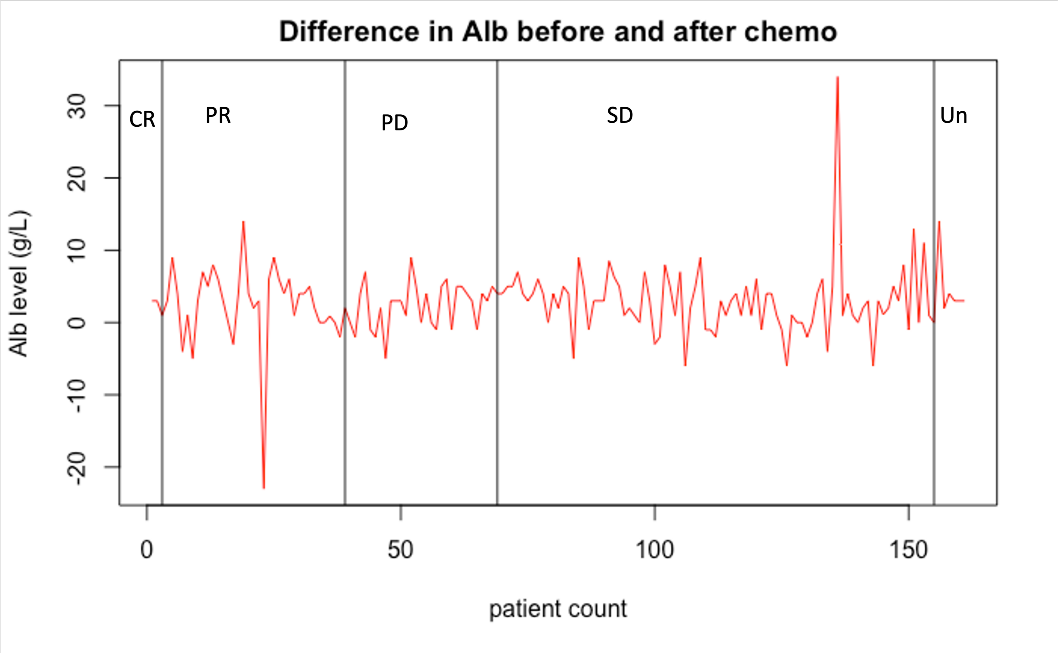


Figure B.152: Difference in albumin values (g/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=161. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=36), Progressive Disease (PD, n=28), Stable Disease (SD, n=84), Unknown (Un, n=11).

B.1.3.16 C-Reactive Protein (CRP)

C-Reactive Protein (CRP) is a hepatic protein that demonstrates an elevation in response to inflammation in the body. The presence of high levels of CRP is indicative of a significant underlying health condition that triggers inflammation, which in turn aids the body in safeguarding its tissues and facilitating their recuperation from injuries, infections, and other illnesses. A desirable CRP value should ideally be less than 1.0 mg/L according to medical experts from the Erasmus Medical Centre Rotterdam. The present study's dataset (n=247) revealed that the mean CRP value before the first chemotherapy treatment was 14.0 mg/L, whereas after the initial chemotherapy treatment, the mean increased to 22.0 mg/L. The statistical summary presented in Table B.51 demonstrated that all CRP metrics increased after the first chemotherapy session, with the maximum value exhibiting more than a twofold increment, indicating a substantial surge in inflammation among patients. The visual aids depicted in Figure B.153 and Figure B.154 corroborated the data by displaying a similar trend in which the majority of patients demonstrated low CRP values before the first chemotherapy session, followed by an increase afterward. Moreover, the scatter plot and histogram showed that the outlier values in both the before and after first chemotherapy CRP values increased. Nonetheless, there was no observable significant difference among the final response groups regarding the CRP values before or after the first chemotherapy treatment. The QQplots in Figure B.155 as well as the tests results given in Table B.52 also reject normality of the CRP values.

Omitting the missing values results in a dataset consisting of n= 127 observations, with 66 being male and 61 being female. The distribution of the final response variable indicates that only one patient achieved a complete response (CR), while 25 and 24 patients demonstrated partial response (PR) and progressive disease (PD), respectively. Additionally, 70 patients displayed stable disease (SD), while the remaining seven observations had unknown values. Moreover, the CRP statistics presented in Table B.53 exhibit an increasing trend in all values following the first chemotherapy session, similar to what has been observed earlier before removing the missing values. The graphs in Figure B.156 and Figure B.157 further support this trend, as the green curve predominantly lies above the red curve for a considerable number of patients, indicating an elevation in inflammation after the initial chemotherapy treatment. Moreover, the zoomed-in view of Figure B.158 reveals that almost all CRP values surpass the healthy threshold, an expected outcome, given that PDAC patients tend to have already elevated inflammation levels. Nevertheless, there was no definitive demarcation among the final response groups, consistent with previous observations. The sole observable pattern was that roughly half of the patients with partial response had lower CRP levels in general. However, no definitive conclusions could be drawn from this. The graph in Figure B.159 demonstrates that the most significant difference in CRP levels occurred in the PD and SD groups, with SD demonstrating the most prominent spikes.

CRP Value	*Before 1st Chemo*	*After 1st Chemo*
Min	0.0	0.0
1st Quartile	1.8	2.7
Median	5.0	8.0
Mean	14.0	22.0
3rd Quartile	12.0	24.2
Max	150.0	332.0
NA	78	100

Table B.51: Summary statistics values of the C-Reactive Protein values of the entire dataset in mg/L, n=247.

CRP value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.591	0.506
SW-test pvalue	<2.2e-16	<2.2e-16
KS test pvalue	Invalid	Invalid

Table B.52: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the C-reactive protein values, n=247. The KS-test is invalid due to the presence of ties in the data.

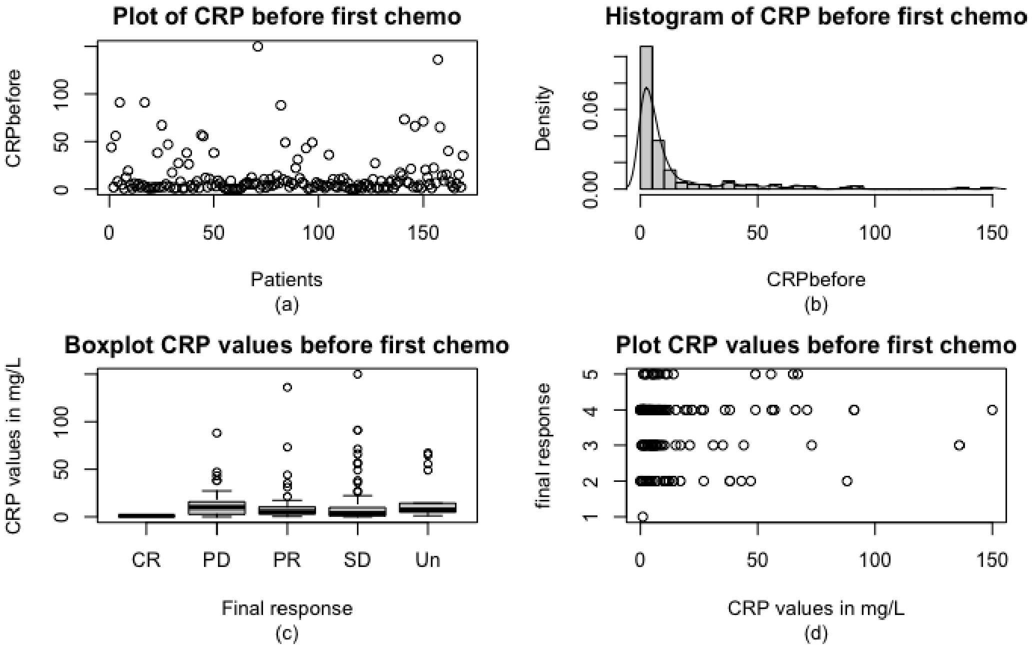


Figure B.153: C-reactive protein distribution before the first chemotherapy cycle (mg/L) for the entire dataset: (a) scatterplot of CRP values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CRP values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

CRP values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	0.0	0.0	-318.0
1st Quartile	1.3	2.7	-12.0
Median	4.7	8.0	-1.3
Mean	13.0	22.3	-9.3
3rd Quartile	10.0	23.0	2.0
Max	150.0	332.0	143.0

Table B.53: Summary statistics of the C-reactive protein levels in mg/L before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=127.

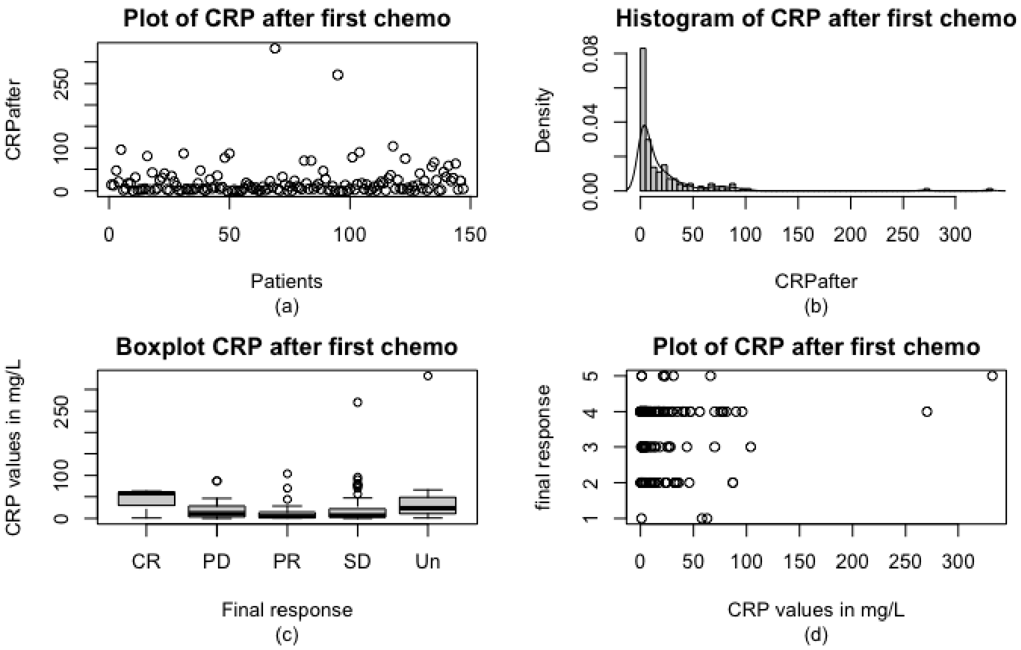


Figure B.154: C-reactive protein distribution after the first chemotherapy cycle (mg/L) for the entire dataset: (a) scatterplot of CRP values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of CRP values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

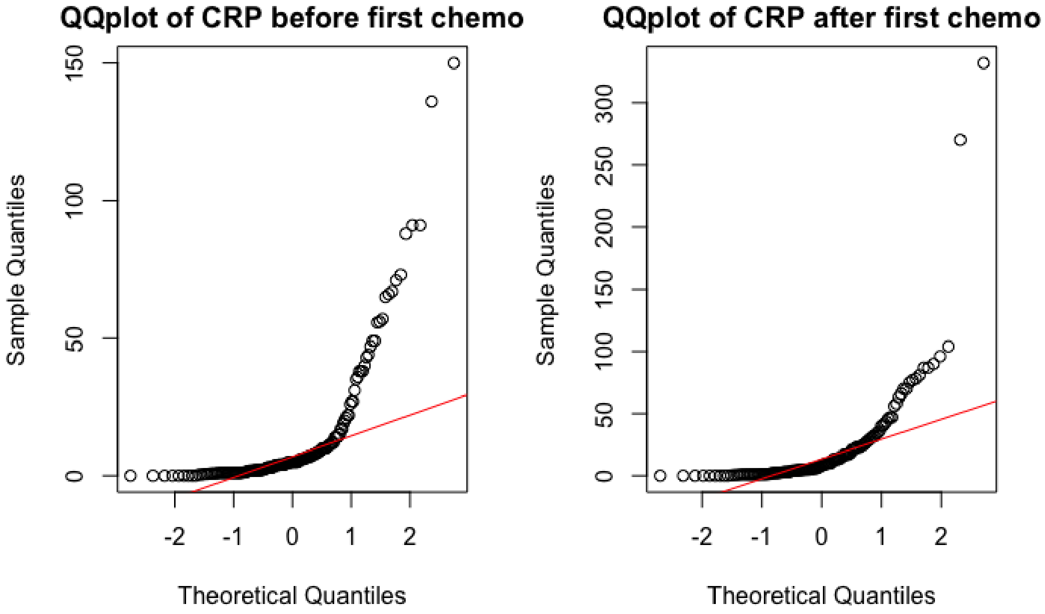


Figure B.155: QQplot of the thrombocyte values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

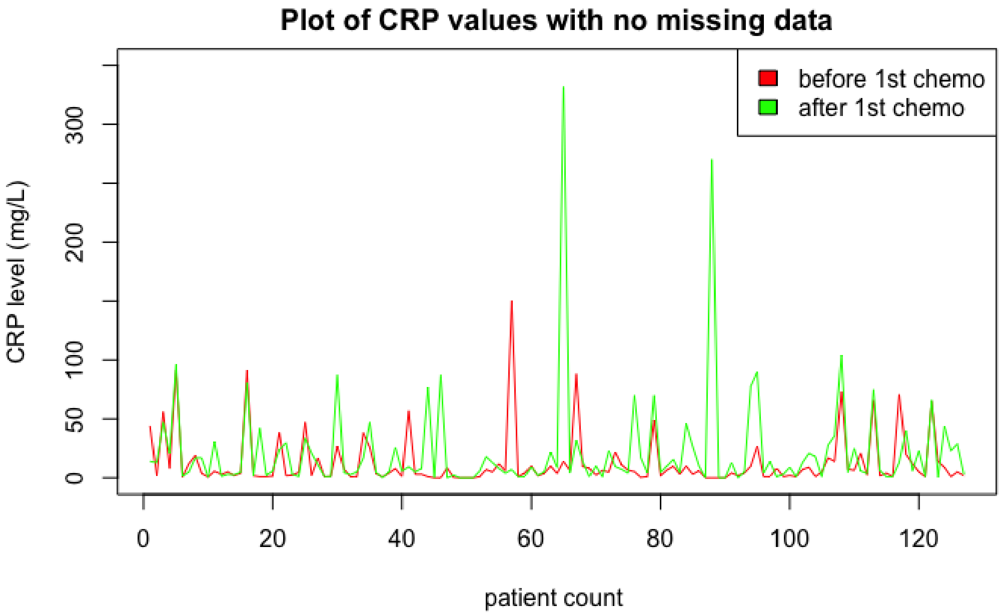


Figure B.156: C-reactive protein levels (mg/L) before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=127.

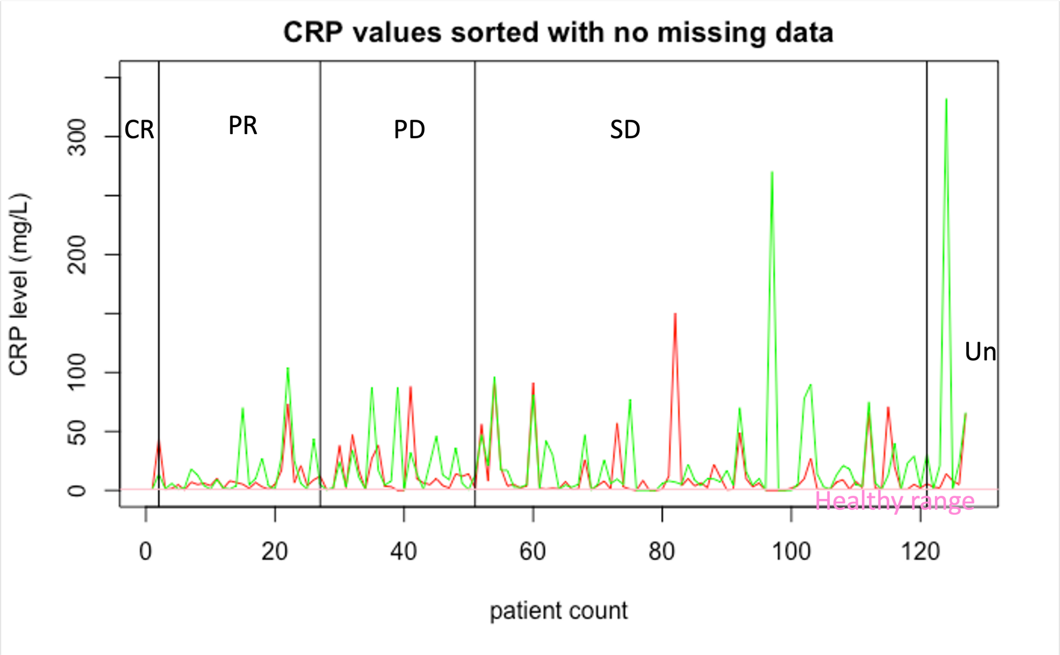


Figure B.157: C-reactive protein levels (mg/L) before and after the first chemotherapy cycle sorted by final response with no missing values, n=127. The final response is classified as Complete Response (CR, n=1), Partial response (PR, n=25), Progressive Disease (PD, n=24), Stable Disease (SD, n=70), Unknown (Un, n=7). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

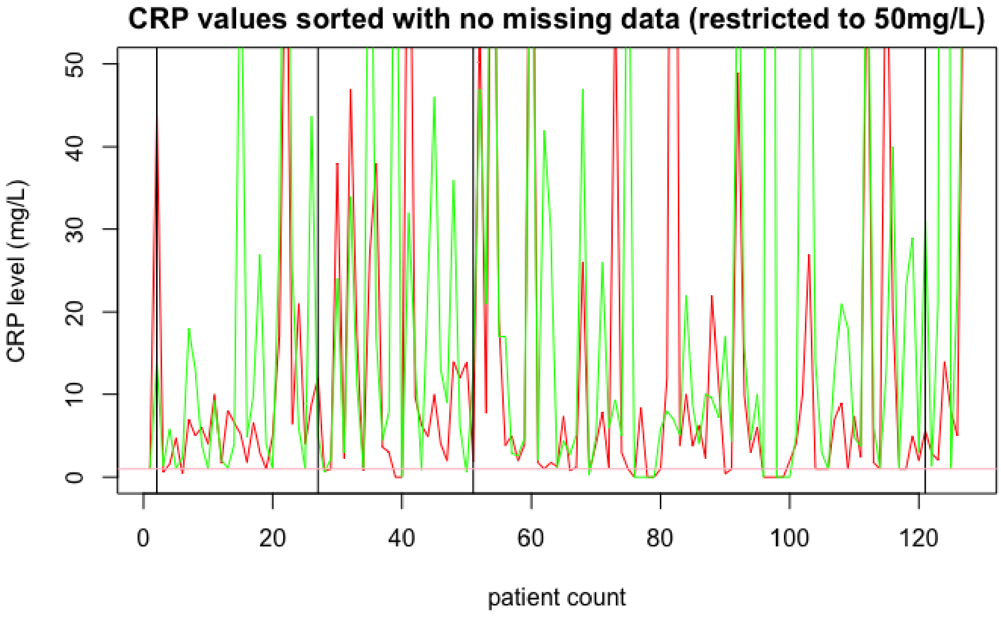


Figure B.158: Same plot as Figure B.157 but with the y-values restricted from 0 to 50 mg/L. The pink line indicates a CRP level of 1.0 mg/L.

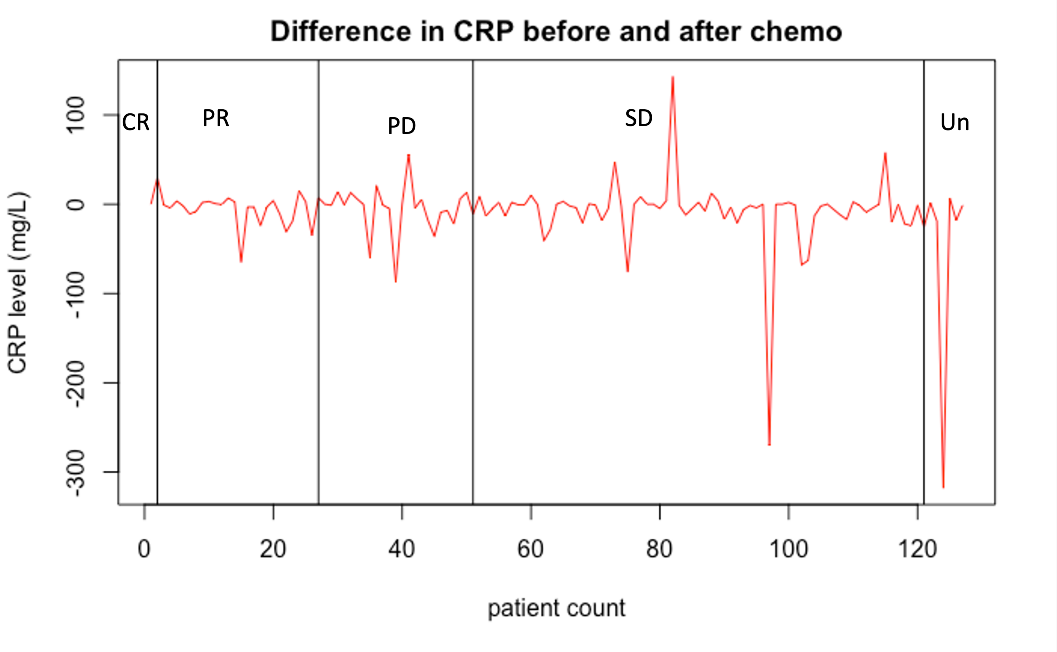


Figure B.159: Difference in hemoglobin values (mmol/L) between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=127. The final response is classified as Complete Response (CR, n=1), Partial response (PR, n=25), Progressive Disease (PD, n=24), Stable Disease (SD, n=70), Unknown (Un, n=7).

B.1.3.17 International Normalized Ratio (INR)

The International Normalized Ratio (INR) is a ratio commonly used to evaluate blood coagulation disorders by measuring the time for blood to clot, also known as Prothrombin time (PT). Prothrombin, a protein synthesized by the liver, is one of several clotting factors that assist in maintaining the appropriate blood consistency. INR values below 2.0 are considered healthy, with higher values indicating a worse condition according to medical experts from the Erasmus Medical Centre Rotterdam. The dataset provided comprises no more than 30 INR value observations, making it inadequate for drawing robust conclusions. Nevertheless, examining the available data may still be informative. Of particular interest is the high outlier value of 26.0 found in patient 018PP20003. The statistical analysis presented in Table B.54 indicates that the median INR value remains unchanged following the initial chemotherapy treatment, while the mean has increased. It should be noted, however, that the outlier value of 26.0 is likely responsible for this increase, as evidenced by the third quartile value of 1.2 after the first chemotherapy treatment versus 1.1 pre-treatment. Visualization of the data via scatterplots, histograms, and boxplots, as depicted in Figure B.160 and Figure B.161, indicates that the majority of patients exhibit INR values of 1.0, and all partial responders exhibit INR values of either 0.0 or 1.0 following the first chemotherapy session. Furthermore, the QQplots depicted in Figure B.162 and test values presented in Table B.55 suggest that the INR values are not normally distributed.

After filtering out any incomplete data, the dataset comprises only n=19 observations, with 8 male and 11 female patients. Of these, 2 patients have partial response (PR), 3 have progressive disease (PD), 13 have stable disease (SD), and 1 has an unknown final response. The statistical measures reported in Table B.56 are quite similar to those before. Specifically, the median INR value before and after the first chemotherapy session remains 1.0, while the mean value after the first session has increased somewhat due to the presence of the high outlier INR value of 26.0. The plots provided in Figure B.163 and Figure B.164 do not reveal any clear trends with respect to the final response groups. Likewise, Figure B.165 does not show any significant differences between the response groups, as the spike observed in the SD group is caused by a single patient.

INR Value	*Before 1st Chemo*	*After 1st Chemo*
Min	0.0	0.0
1st Quartile	0.99	0.0
Median	1.0	1.0
Mean	0.85	1.7
3rd Quartile	1.1	1.2
Max	2.5	26.0
NA	217	217

Table B.54: Summary statistics values of the INR values of the entire dataset, n=247.

INR value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.717	0.303
SW-test pvalue	2.8e-6	7.4e-11
KS test pvalue	6.6e-10	6.2e-10

Table B.55: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the INR values, n=30.

INR values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	0.0	0.0	-25.0
1st Quartile	0.0	0.0	-0.1
Median	1.0	1.0	0.0
Mean	0.74	2.1	-1.4
3rd Quartile	1.1	1.2	0.0
Max	2.5	26.0	2.5

Table B.56: Summary statistics of the INR values before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=19.

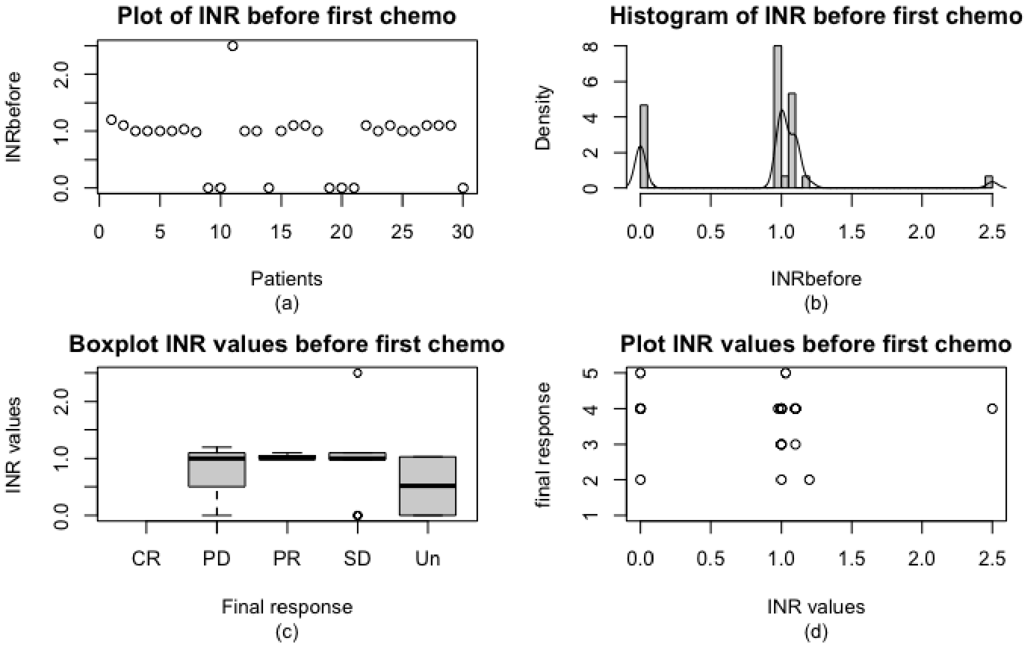


Figure B.160: INR distribution before the first chemotherapy cycle for the entire dataset: (a) scatterplot of INR values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of INR values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

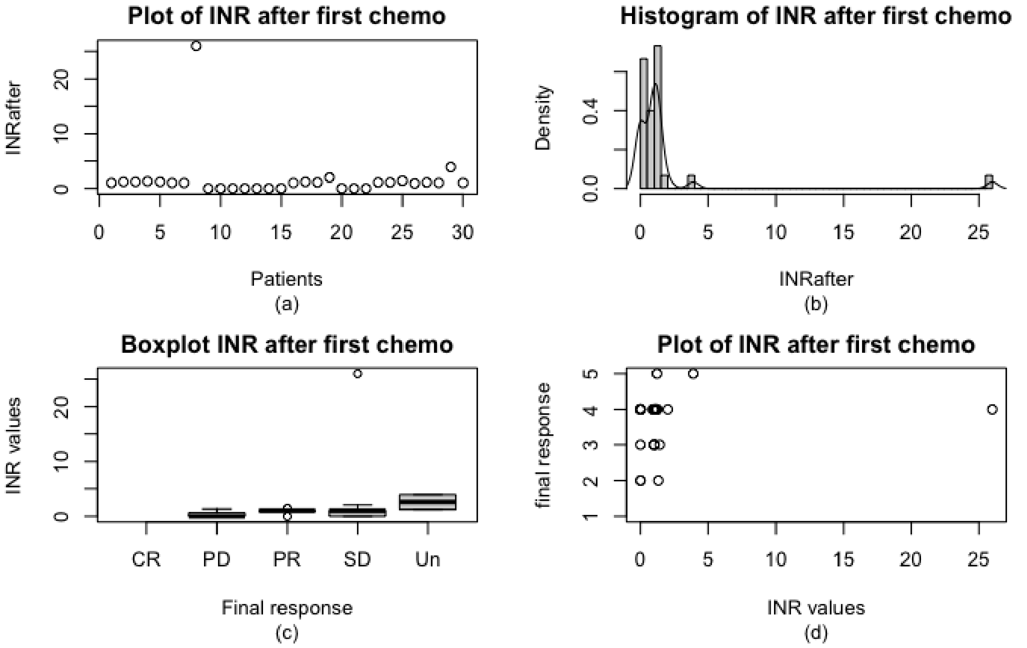


Figure B.161: INR distribution after the first chemotherapy cycle for the entire dataset: (a) scatterplot of INR values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of INR values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

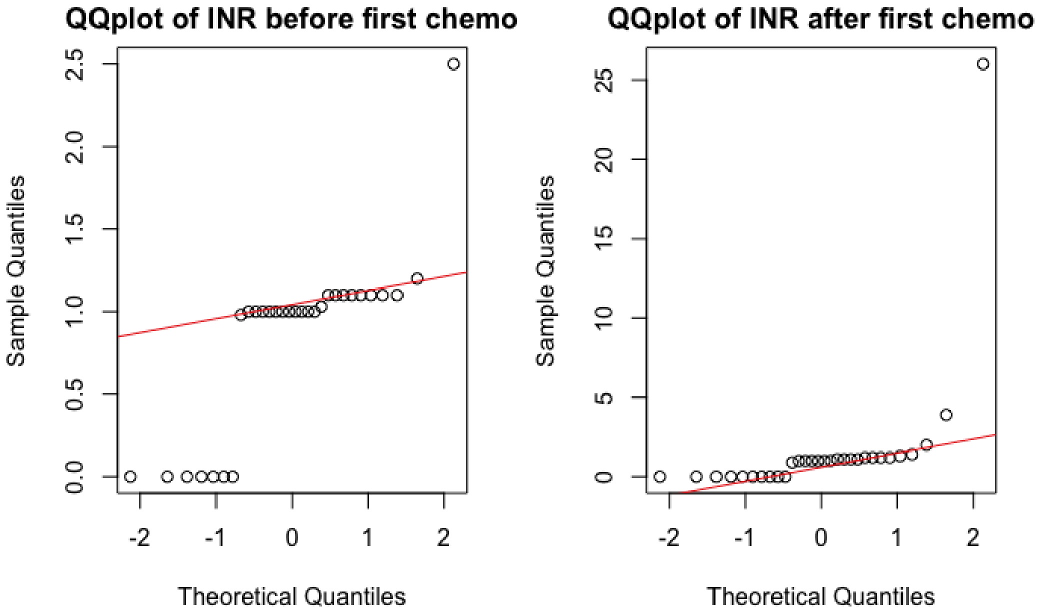


Figure B.162: QQplot of the INR values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

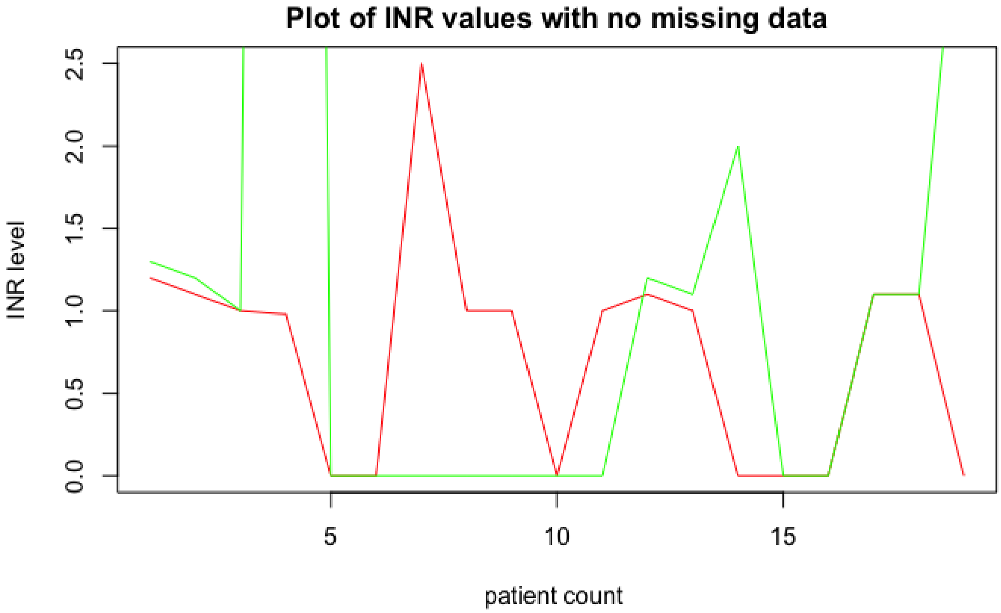


Figure B.163: INR levels before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=19.

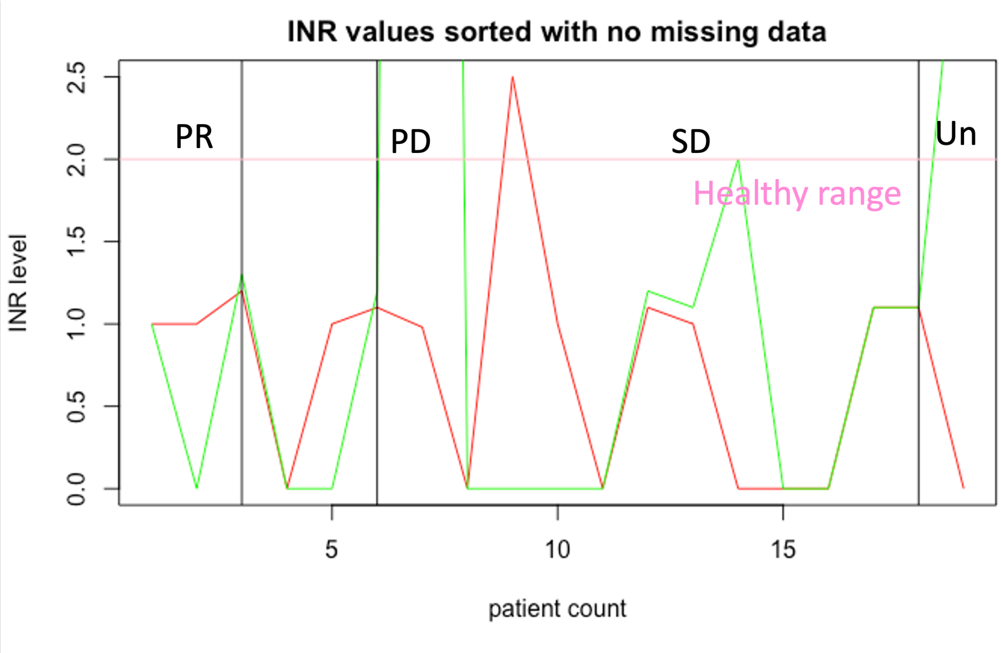


Figure B.164: INR values before and after the first chemotherapy cycle sorted by final response with no missing values, n=19. The final response is classified as Complete Response (CR, n=0), Partial response (PR, n=2), Progressive Disease (PD, n=3), Stable Disease (SD, n=13), Unknown (Un, n=1). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

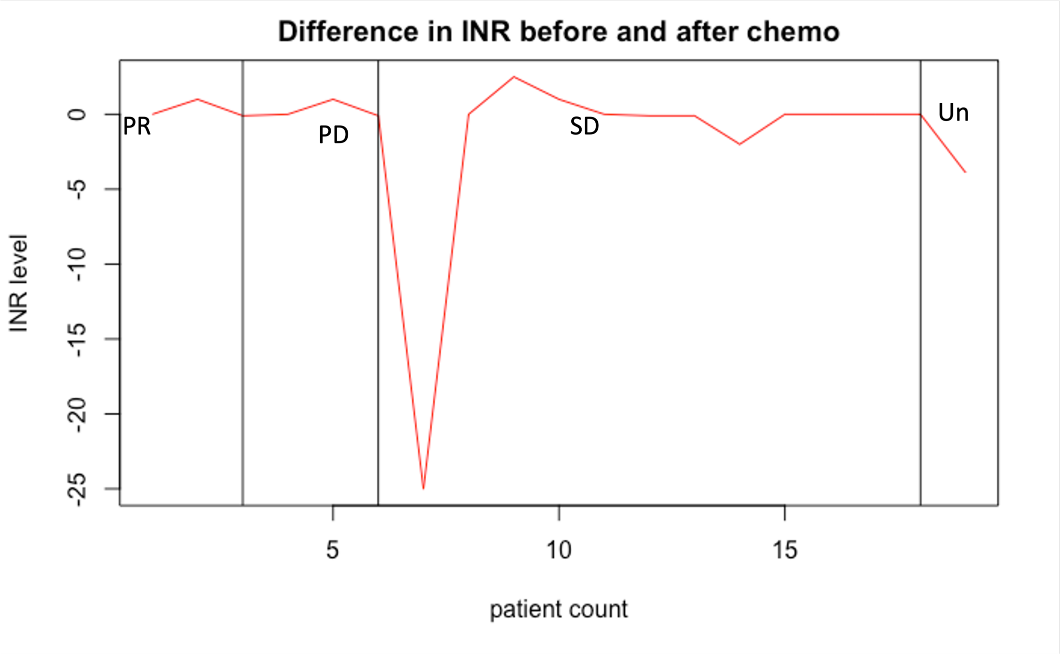


Figure B.165: Difference in INR values between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=19. The final response is classified as Complete Response (CR, n=0), Partial response (PR, n=2), Progressive Disease (PD, n=3), Stable Disease (SD, n=13), Unknown (Un, n=1).

B.1.3.18 Systemic Inflammation Index (SII)

The systemic inflammation index (SII) is a metric that characterizes the level of inflammation in the body. It is calculated using the following formula:

$$SII = \frac{N \times P}{L} \tag{B.4}$$

Here, N refers to the neutrophil count, P refers to the platelet (thrombocytes) count, and L refers to the lymphocyte count. All these values are measured in $10^9/L$. A SII value below 900 is considered normal or healthy according to medical experts from the Erasmus Medical Centre Rotterdam. The statistical results in Table B.57 indicate that the average SII values after the first chemotherapy treatment are higher than before, with both exceeding the healthy threshold of 900. However, closer examination of the table shows that while the median, mean, and third quartile increased, the minimum SII value decreased from 57.4 before treatment to 17.1 after, and the maximum value decreased from 7246.8 to 5563.2. This suggests an overall increase in SII values after the first chemotherapy treatment, but with some patients experiencing a decrease. The scatterplot, histogram, and boxplots in Figure B.166 and Figure B.167 demonstrate a left-skewed distribution with a long right tail due to high SII values for a couple of patients. Furthermore, the distribution appears to be more dispersed after treatment, which can be seen clearly in the scatterplot. However, no clear association between SII values and final response to chemotherapy can be deduced from the provided boxplots. The non-normality of SII values is confirmed by both QQplots and values in Table B.58.

Once all missing values have been removed, the SII dataset consists of 118 observations, with 62 male and 56 female patients. The final response is CR for 2 patients, PR for 21, PD for 24, SD for 63, and 8 were categorized as Unknown. The statistical analysis presented in Table B.59 shows that the central tendency and dispersion of SII values were similar to those observed in the full dataset. As depicted in Figure B.169 and Figure B.170, SII values tend to increase after the first chemotherapy treatment, although some patients displayed a decrease in SII values. The smallest difference between the SII values before and after treatment is observed in the PR group. Conversely, the largest differences are identified in the SD group of this dataset.

SII Value	*Before 1st Chemo*	*After 1st Chemo*
Min	57.4	17.1
1st Quartile	545.7	492.0
Median	789.7	852.8
Mean	1013.0	1121.3
3rd Quartile	1176.2	1481.2
Max	7246.8	5563.2
NA	75	106

Table B.57: Summary statistics values of the SII values of the entire dataset, n=247.

SII value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.645	0.833
SW-test pvalue	<2.2e-16	2.26e-11
KS test pvalue	<2.2e-16	<2.2e-16

Table B.58: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the SII values, n=247.

SII values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	57.4	17.1	-4357.6
1st Quartile	569.8	503.3	-469.6
Median	778.5	831.0	-46.6
Mean	968.9	1126.0	-157.1
3rd Quartile	1139.3	1349.5	392.3
Max	7246.8	5563.2	6178.6

Table B.59: Summary statistics of the SII values before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=118.

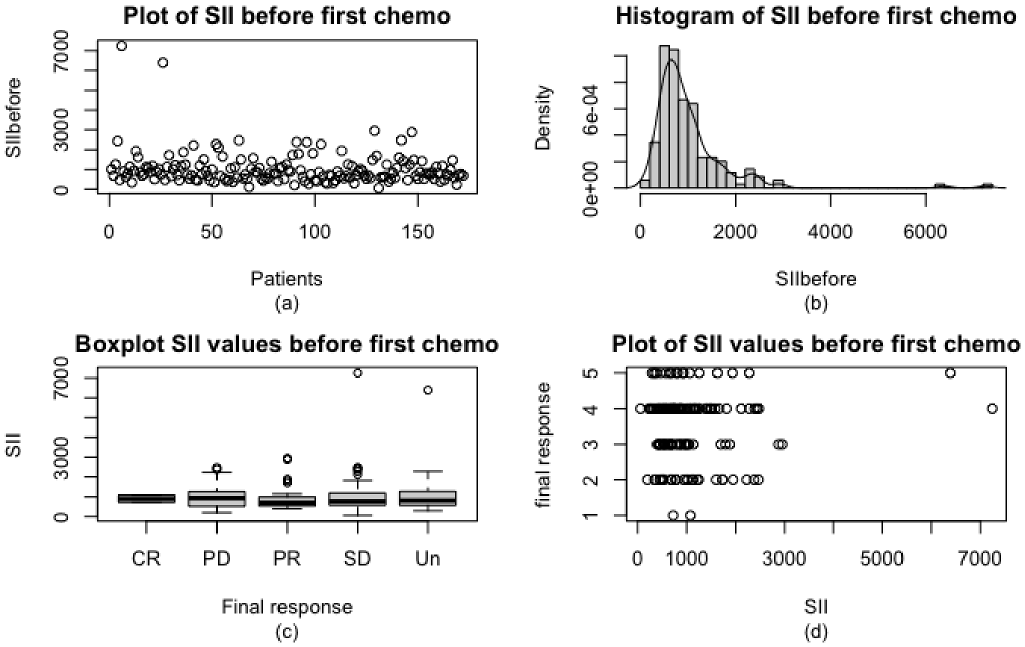


Figure B.166: SII distribution before the first chemotherapy cycle for the entire dataset: (a) scatterplot of SII values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of SII values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

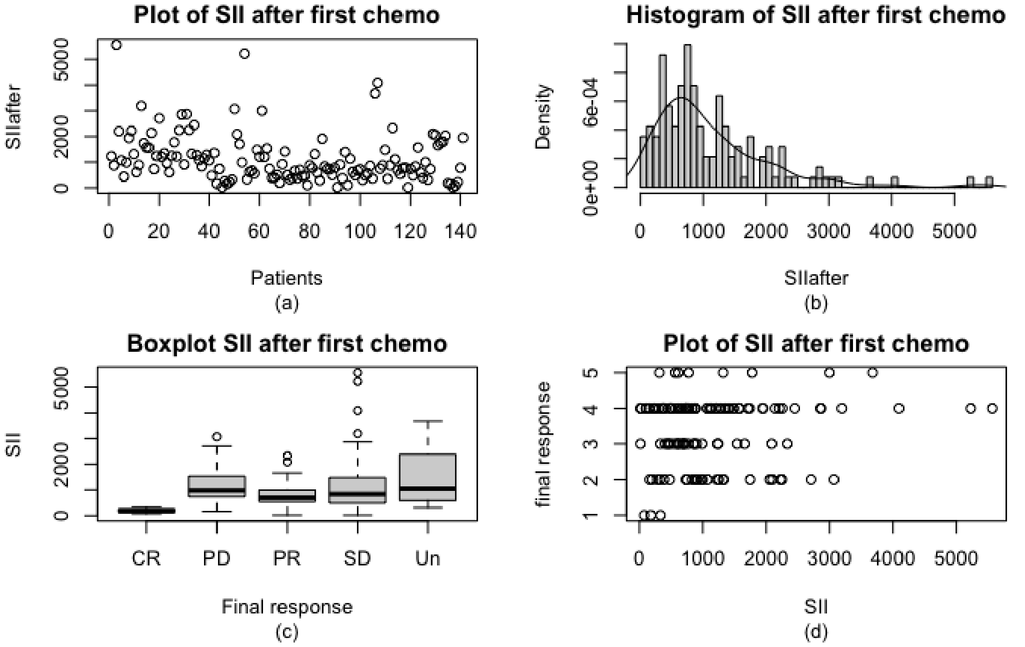


Figure B.167: SII distribution after the first chemotherapy cycle for the entire dataset: (a) scatterplot of SII values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of SII values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

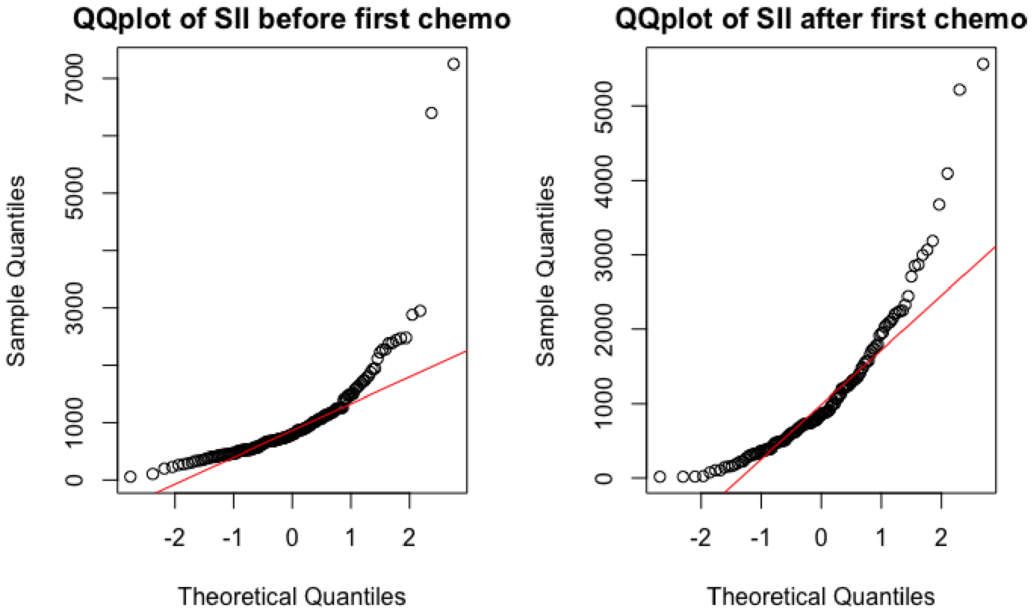


Figure B.168: QQplot of the SII values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

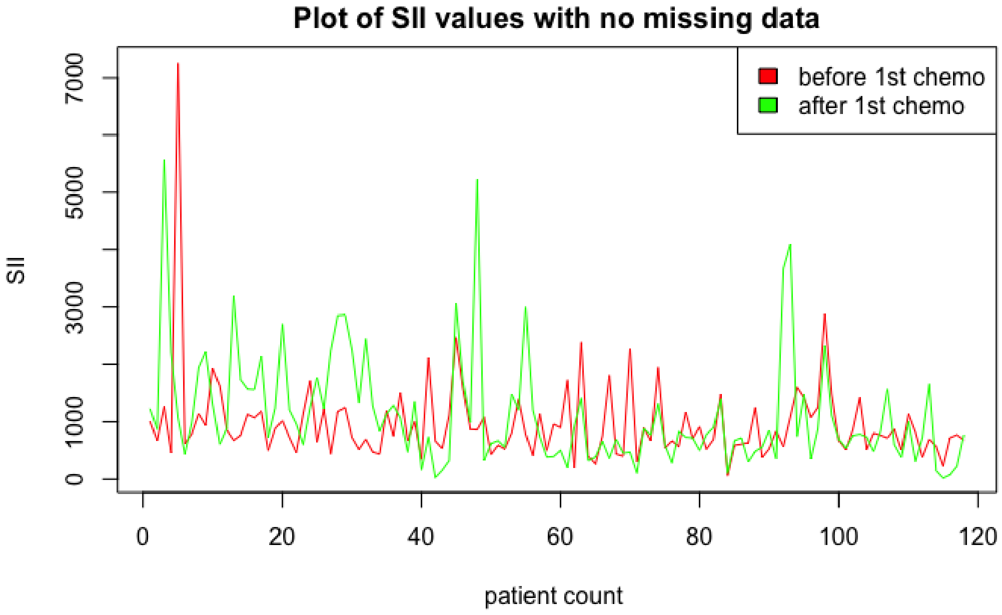


Figure B.169: SII values before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=118.

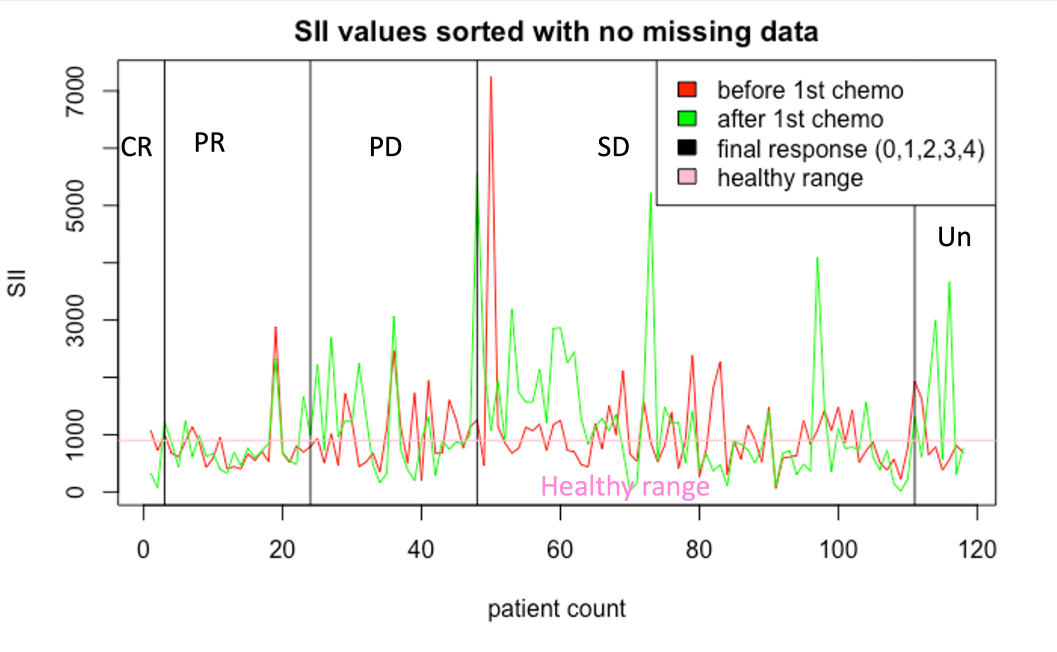


Figure B.170: SII values before and after the first chemotherapy cycle sorted by final response with no missing values, n=118. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=21), Progressive Disease (PD, n=24), Stable Disease (SD, n=63), Unknown (Un, n=8). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy range.

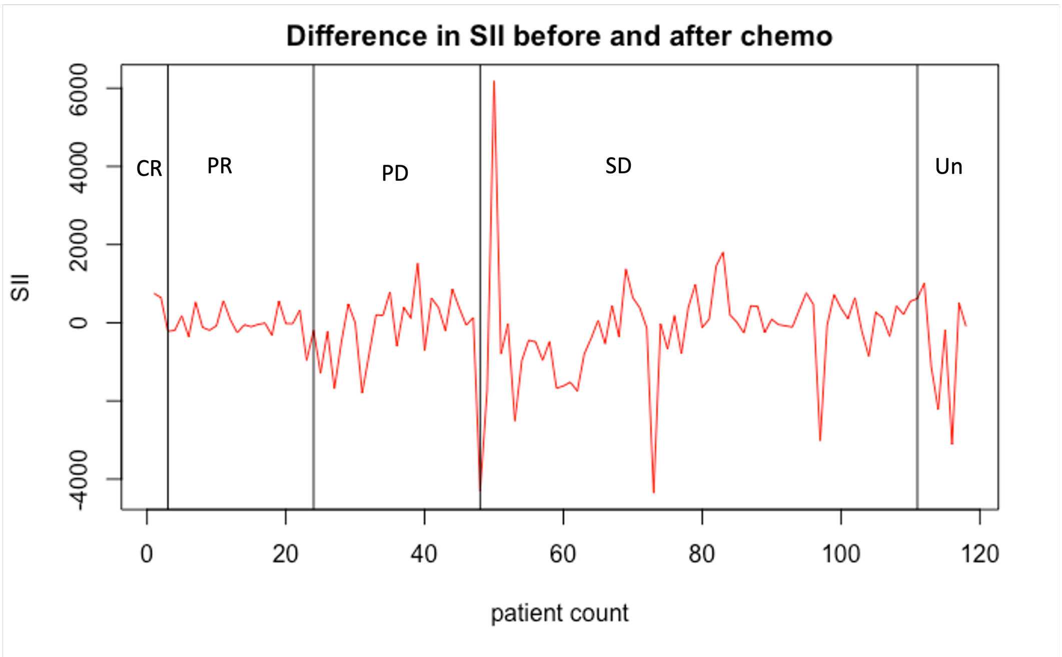


Figure B.171: Difference in SII values between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=118. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=21), Progressive Disease (PD, n=24), Stable Disease (SD, n=63), Unknown (Un, n=8).

B.1.3.19 Neutrophil to Lymphocyte Ratio (NLR)

The Neutrophil to Lymphocyte Ratio (NLR) and Platelet to Lymphocyte Ratio (PLR) are two ratios that have been investigated in relation to the response of PDAC patients to chemotherapy. The NLR can be calculated by dividing the neutrophil count by the lymphocyte count in the blood, as shown in Equation B.5.

$$NLR = \frac{N}{L} \tag{B.5}$$

where N is the neutrophil count and L the lymphocyte count in the blood. According to the study by Forget et al. [108] healthy adults have NLR values between 0.78 and 3.53. Based on the statistics presented in B.60, it is evident that the Neutrophil to Lymphocyte Ratio (NLR) increased on average for all metrics, except for the minimum value, which decreased from 0.43 to 0.14 after the first chemotherapy treatment. This increase was observed in the maximum value, which almost doubled from 19.8 to 36.8, suggesting a significant rise in neutrophil count relative to lymphocyte count in the blood. The scatter plots, histograms, and boxplots depicted in B.172 and B.173 reveal a left-skewed distribution of NLR values before and after the first chemotherapy treatment with a long right tail attributable to outliers. The boxplots do not exhibit any clear trend in the final response to chemotherapy concerning NLR values before or after the first chemotherapy treatment. The QQplots illustrated in B.174 and test results outlined in B.61 confirm non-normality, confirming the left-skewed distribution and long right tail.

Following the exclusion of missing values, the remaining dataset contains 119 observations, of which 63 are male and 56 are female. Of these, 2 patients had CR, 21 had PR, 24 had PD, 64 had SD, and 8 had an unknown final response. As previously observed, the NLR values continue to exhibit an increasing trend in all metrics, except for the minimum value, with the green graph consistently higher than the red graph. This pattern is evident in the plots and difference plot presented in Figure B.175, Figure B.177, and Figure B.181, respectively. In particular, the SD group displays the largest spikes in NLR values, while the smallest differences and NLR values are observed in the PR group. Nevertheless, almost all of NLR values exceed the healthy reference range, as shown in Figure B.178. Further analysis by gender indicates no discernible difference in NLR values between male and female patients, seen in Figure B.179 and Figure B.180.

NLR Value	*Before 1st Chemo*	*After 1st Chemo*
Min	0.43	0.14
1st Quartile	2.15	2.59
Median	3.05	4.41
Mean	3.60	5.51
3rd Quartile	4.42	6.77
Max	19.80	36.84
NA	75	105

Table B.60: Summary statistics values of the Neutrophil-to-Lymphocyte ratio values of the entire dataset, n=247.

NLR value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.761	0.695
SW-test pvalue	1.929e-15	8.269e-16
KS test pvalue	<2.2e-16	<2.2e-16

Table B.61: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the NLR values, n=247.

NLR values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	0.43	0.18	-32.96
1st Quartile	2.25	2.76	-3.06
Median	2.96	4.34	-1.21
Mean	3.49	5.62	-2.13
3rd Quartile	4.15	6.73	0.59
Max	19.80	36.84	12.48

Table B.62: Summary statistics of the Neutrophil-to-Lymphocyte ratio before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=119.

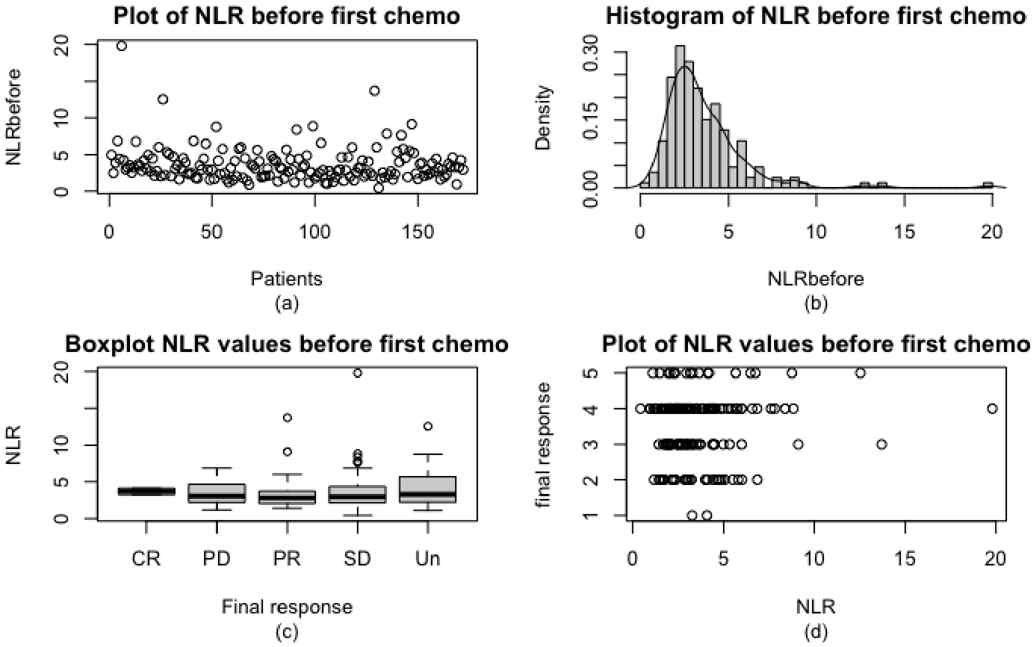


Figure B.172: Neutrophil-to-Lymphocyte Ratio distribution before the first chemotherapy cycle for the entire dataset: (a) scatterplot of NLR values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of NLR values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

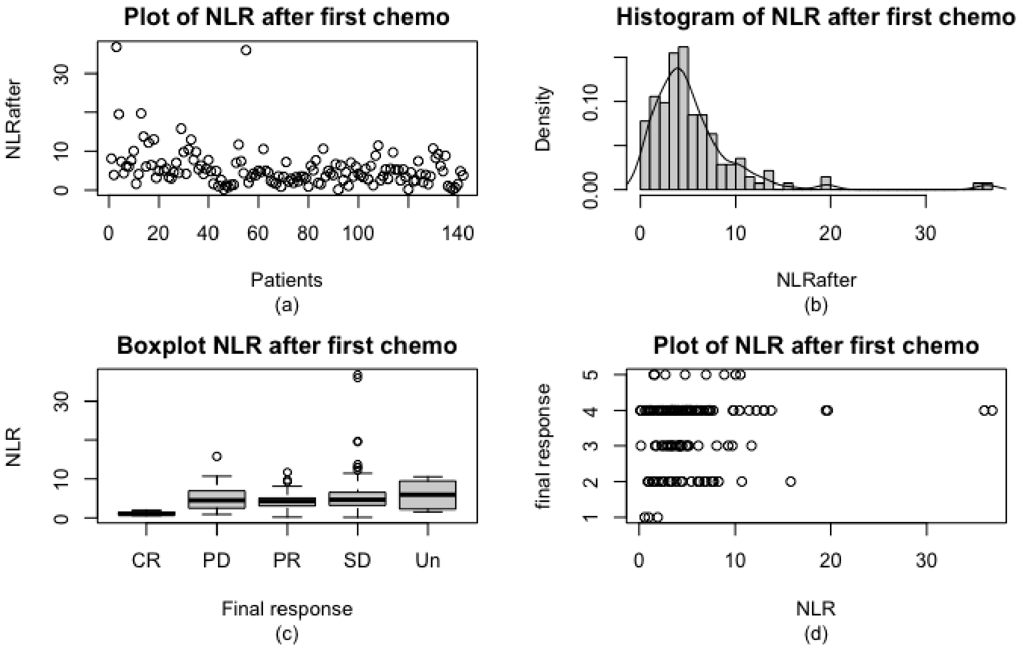


Figure B.173: Neutrophil-to-Lymphocyte distribution after the first chemotherapy cycle for the entire dataset: (a) scatterplot of NLR values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of NLR values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

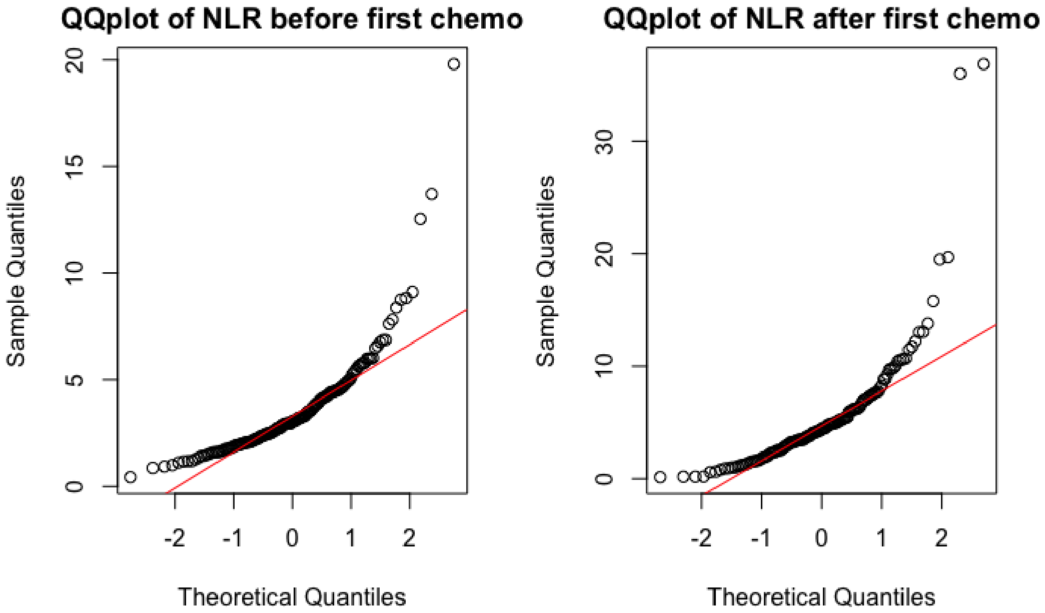


Figure B.174: QQplot of the NLR values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

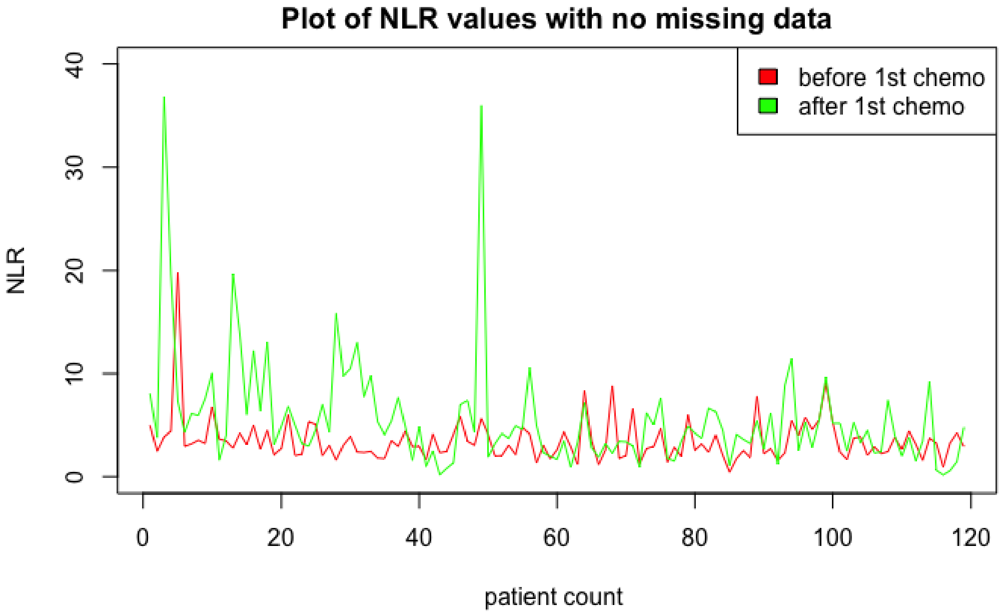


Figure B.175: Neutrophil-to-Lymphocyte ratio before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=119.

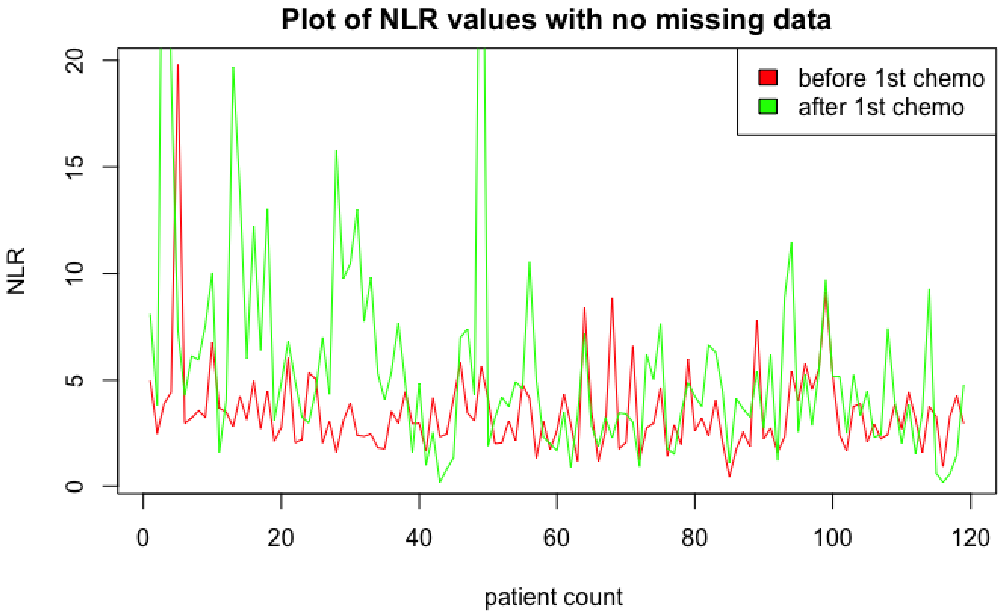


Figure B.176: Same plot as Figure B.175 but with the y-axis restricted from 0-20.

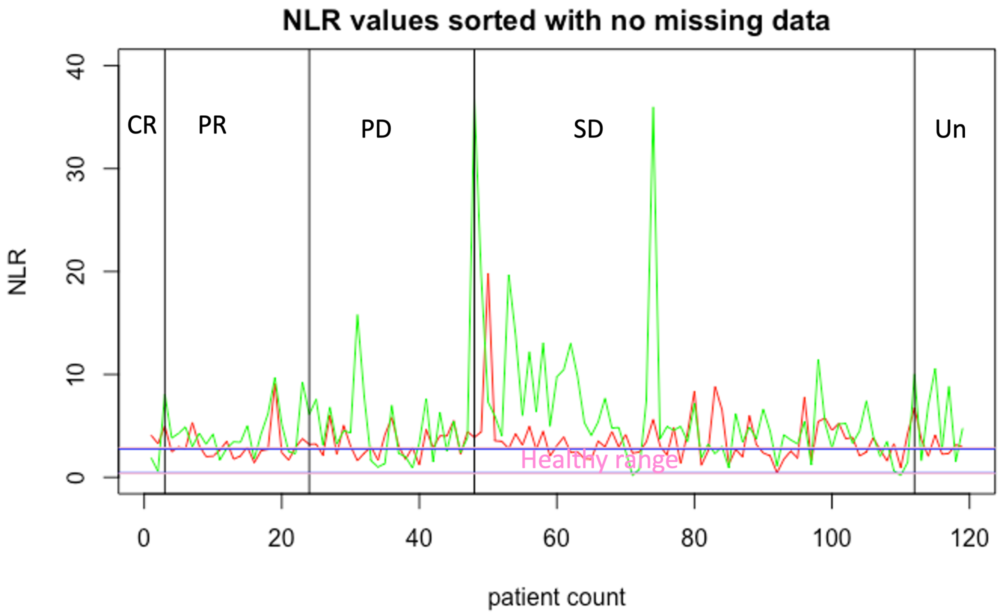


Figure B.177: Neutrophil-to-Lymphocyte ratio before and after the first chemotherapy cycle sorted by final response with no missing values, n=119. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=21), Progressive Disease (PD, n=24), Stable Disease (SD, n=64), Unknown (Un, n=8). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

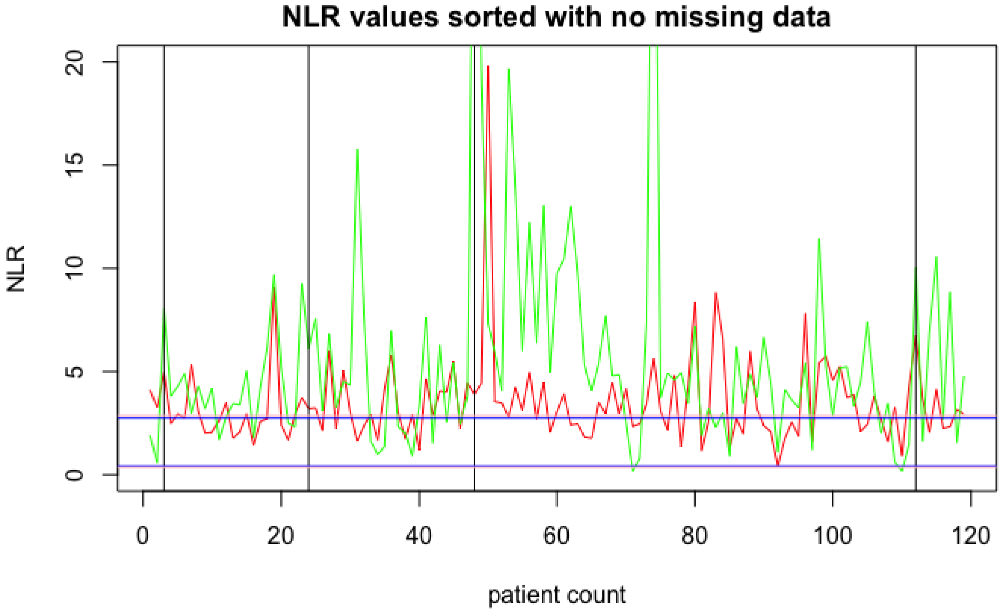


Figure B.178: Same plot as Figure B.177 but with the y-values restricted between 0 and 20.

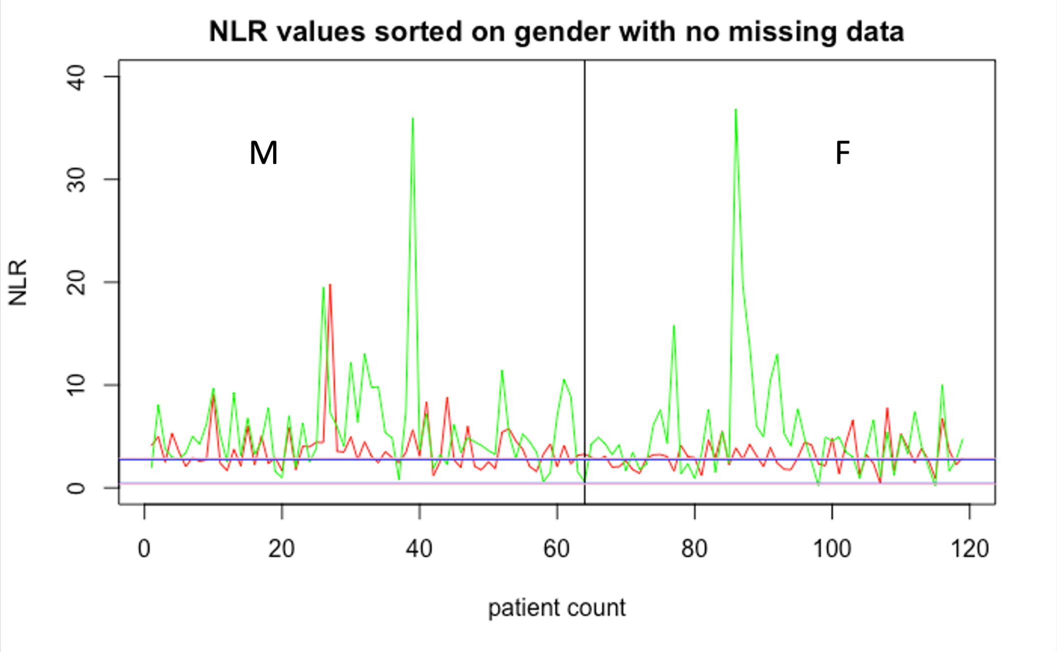


Figure B.179: Neutrophil-to-Lymphocyte ratio before and after the first chemotherapy cycle sorted by gender with no missing values, n=119 (63 male, 56 female). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

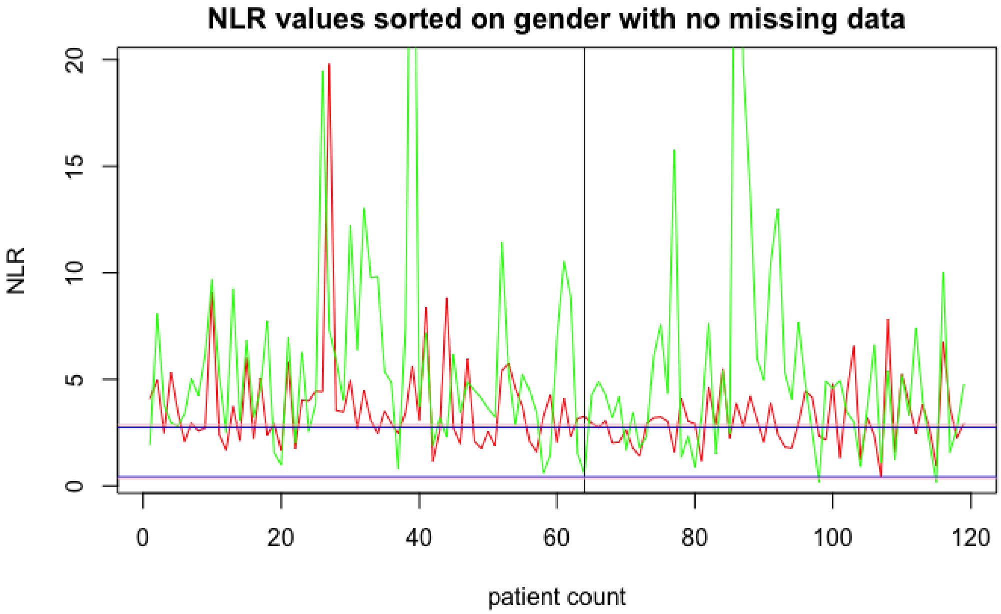


Figure B.180: Same plot as Figure B.179 but with the y-values restricted between 0-20.

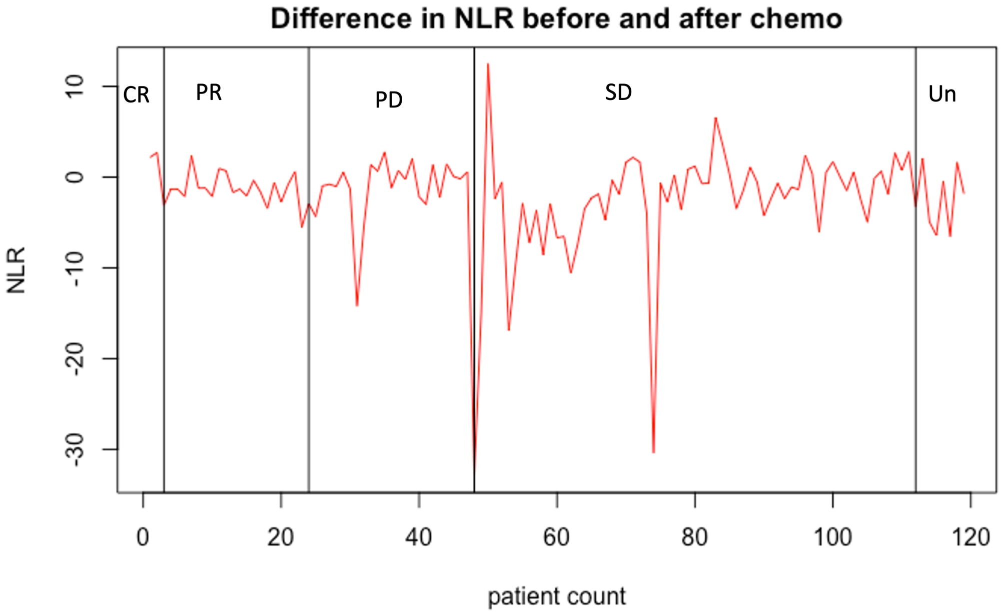


Figure B.181: Difference in Neutrophil-to-Lymphocyte ratios between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=119. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=21), Progressive Disease (PD, n=24), Stable Disease (SD, n=64), Unknown (Un, n=8).

B.1.3.20 Platelet to Lymphocyte Ratio (PLR)

The platelet to lymphocyte ratio (PLR) is a commonly studied metric, defined as the ratio between the platelet count and lymphocyte count in the blood, denoted by P and L respectively, and expressed mathematically as in Equation B.6. Medical experts have identified a healthy range for PLR in males to be within the interval of 36.63–149.13, and in females to be within 43.36–172.68.

$$PLR = \frac{P}{L} \tag{B.6}$$

where P is the platelet (thrombocyte) count in the blood and L the lymphocyte count. A healthy range for males is between 36.63–149.13 and for females between 43.36–172.68 according to experts. To proceed, the data presented in Table B.63 indicate that the PLR values have generally decreased, in contrast to the NLR values. The observed decrease in PLR values after the first chemotherapy treatment can be attributed to either a decrease in platelet count or an increase in lymphocyte count, or a combination of both, resulting in an overall lower PLR value. The distribution of PLR values can be visualized through scatter plots, histograms, and box plots shown in Figure B.182 and Figure B.183. The majority of PLR values are concentrated between 0 and 400 before the first chemotherapy treatment and between 0 and 200 after the first chemotherapy treatment, with a few outlier values indicating high PLR values in some patients. The histograms also indicate a left-skewed distribution with a long right tail due to the high outlier values. However, no clear trend is evident in the final response groups except for the observation that the PD group has the smallest outlier values. The normality of the PLR values is further confirmed by the QQplots presented in Figure B.184and the values provided in Table B.64, indicating clear deviations from normal distributions in the tail, while the majority of values are relatively normally distributed.

Having removed any missing data points, the dataset consists of 121 observations, with 63 being male and 58 being female. The final response distribution is as follows: 2 patients showed complete response (CR), 22 patients showed partial response (PR), 24 patients showed progressive disease (PD), 65 patients showed stable disease (SD), and 8 patients had unknown final response. Consistent with the previous analysis, the PLR values exhibit a clear decrease after the first chemotherapy treatment, as evident from Table B.65. Further visualization through scatter plots and line graphs in Figure B.185 and Figure B.186 show that the green line (representing PLR values after chemotherapy) was generally below the red line (representing PLR values before chemotherapy), indicating a decrease in PLR values. However, a few exceptions were observed, where some patients had higher PLR values after chemotherapy. The majority of PLR values fall above the healthy range, as evident from Figure B.187. Nonetheless, after the first chemotherapy treatment, many values fall within the healthy range represented by the blue and pink lines. Further analysis by gender, as presented in Figure B.188 and Figure B.189, reveal that males tend to have higher PLR values compared to females. Lastly, Figure B.190 confirms that no clear differences were observed between the final response groups.

PLR Value	*Before 1st Chemo*	*After 1st Chemo*
Min	14.4	8.7
1st Quartile	117.0	73.0
Median	163.2	110.4
Mean	185.4	133.1
3rd Quartile	220.3	153.0
Max	915.0	725.0
NA	75	101

Table B.63: Summary statistics values of the Platelet-to-Lymphocyte ratio values of the entire dataset, n=247.

PLR value	*Before 1st Chemo*	*After 1st Chemo*
SW-test W	0.768	0.766
SW-test pvalue	3.144e-15	5.425e-14
KS test pvalue	<2.2e-16	<2.2e-16

Table B.64: Results of the Shapiro-Wilk test and the Kolmogorov-Smirnov test for assessing the normality of the latelet-to-lymphocyte ratio values, n=247.

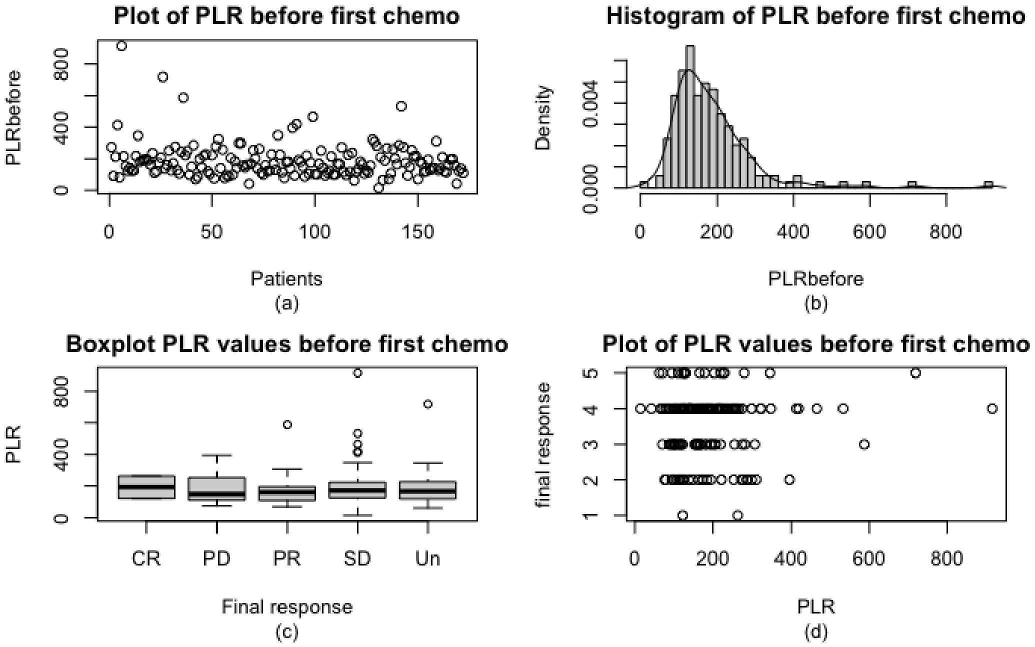


Figure B.182: Platelet-to-Lymphocyte ratio distribution before the first chemotherapy cycle for the entire dataset: (a) scatterplot of PLR values before the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of PLR values before the first chemotherapy cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

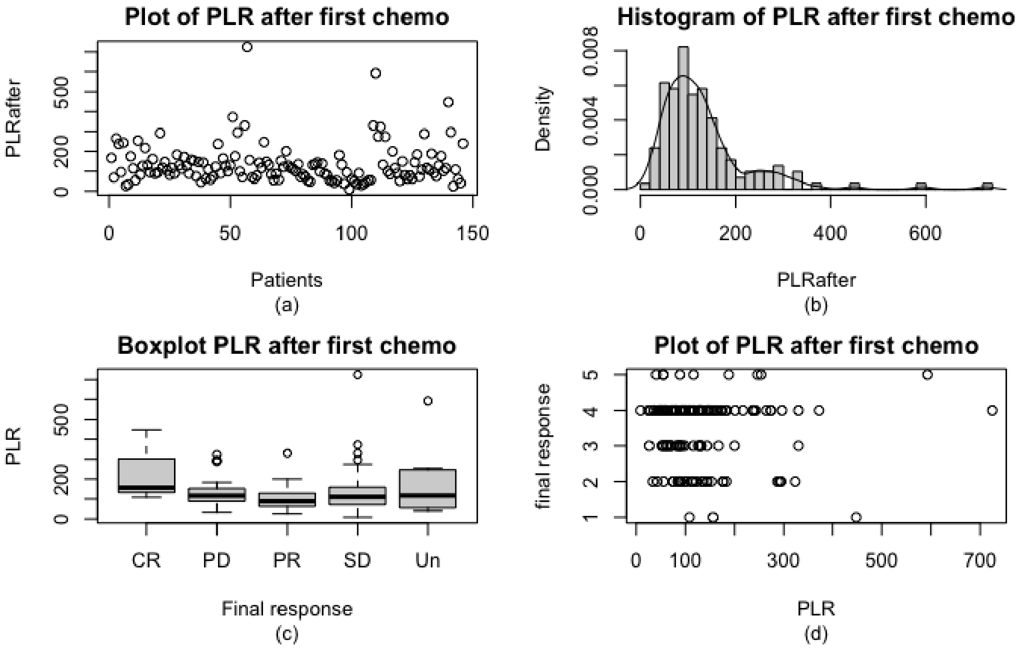


Figure B.183: Platelet-to-Lymphocyte ratio distribution after the first chemotherapy cycle for the entire dataset: (a) scatterplot of PLR values after the first cycle, (b) histogram with a fitted density curve, (c) boxplot showing the distribution of PLR values after the first cycle across different final response categories, including Complete response (CR, n=3), Partial response (PR, n=48), Progressive disease (PD, n=43), Stable disease (SD, n=122), and Unknown (Un, n=31), (d) the same information as (c) using a different graphical approach (1=CR, 2=PR, 3=PD, 4=SD, 5=Un), total n=247.

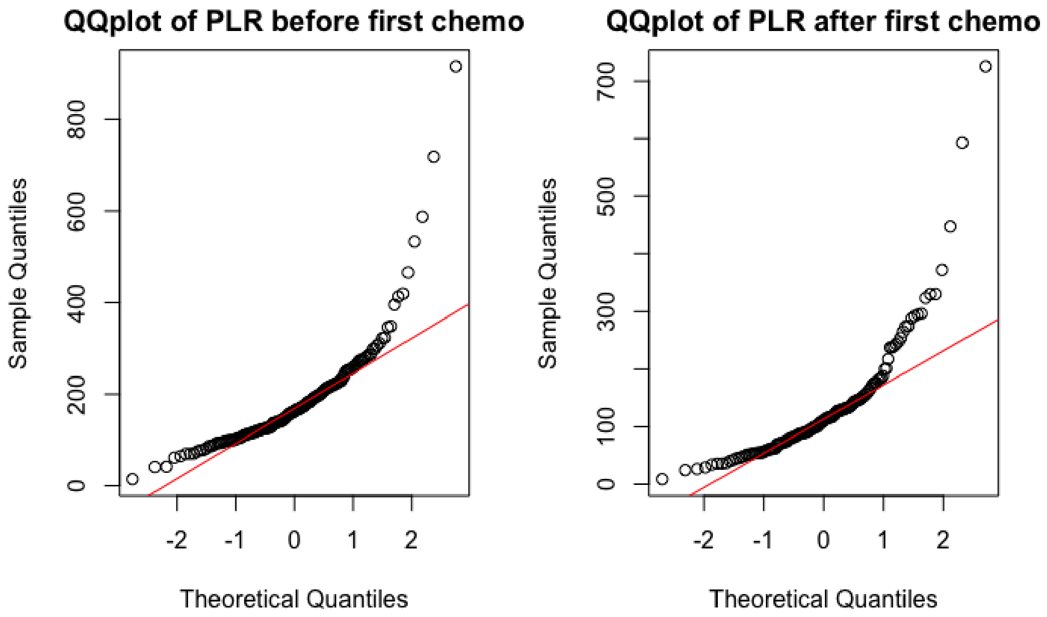


Figure B.184: QQplot of the PLR values before and after the first chemotherapy cycle, with a normal distribution fitted in red, n=247.

PLR values	*Before 1st Chemo*	*After 1st Chemo*	*Difference*
Min	14.4	8.7	-585.9
1st Quartile	117.0	72.6	5.6
Median	159.1	100.3	37.7
Mean	185.0	132.6	52.3
3rd Quartile	223.0	153.1	94.6
Max	915.0	725.0	671.7

Table B.65: Summary statistics of the Platelet-to-Lymphocyte ratio values before and after the first chemotherapy cycle as well as the difference for a cohort of patients with all missing values removed, n=121.

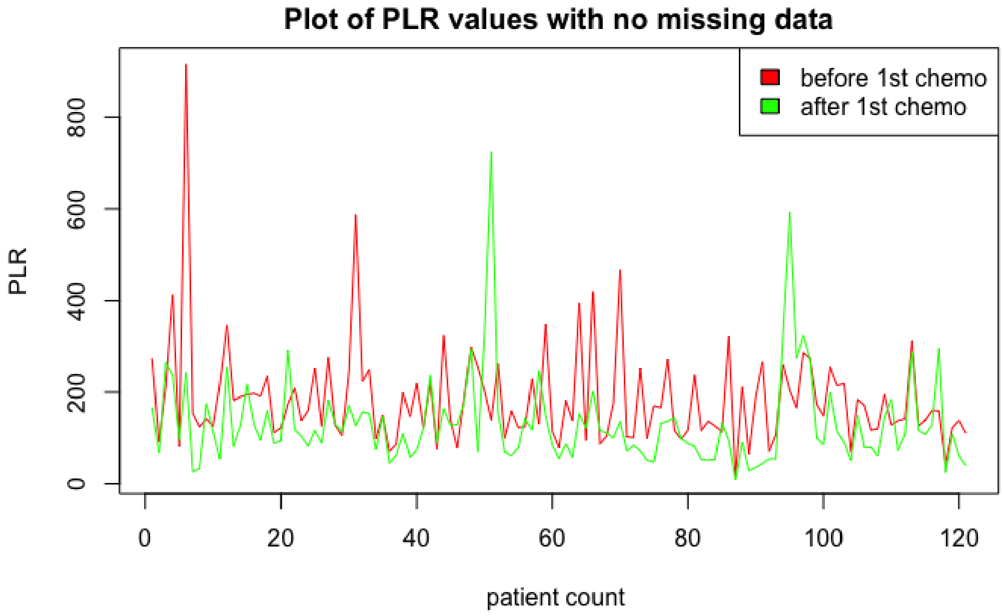


Figure B.185: Platelet-to-Lymphocyte ratio before and after the first chemotherapy cycle. Red = values before the first cycle, Green = values after the first chemotherapy cycle. The plot exclusively contains data with no missing values, n=121.

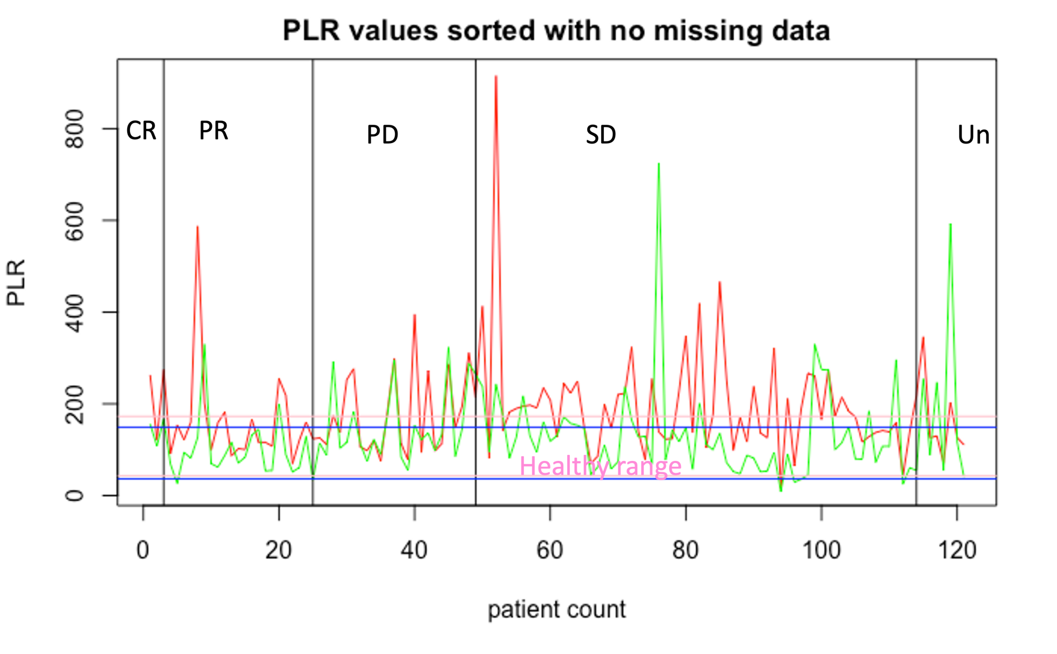


Figure B.186: Platelet-to-Lymphocyte ratio before and after the first chemotherapy cycle sorted by final response with no missing values, n=121. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=22), Progressive Disease (PD, n=24), Stable Disease (SD, n=65), Unknown (Un, n=8). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

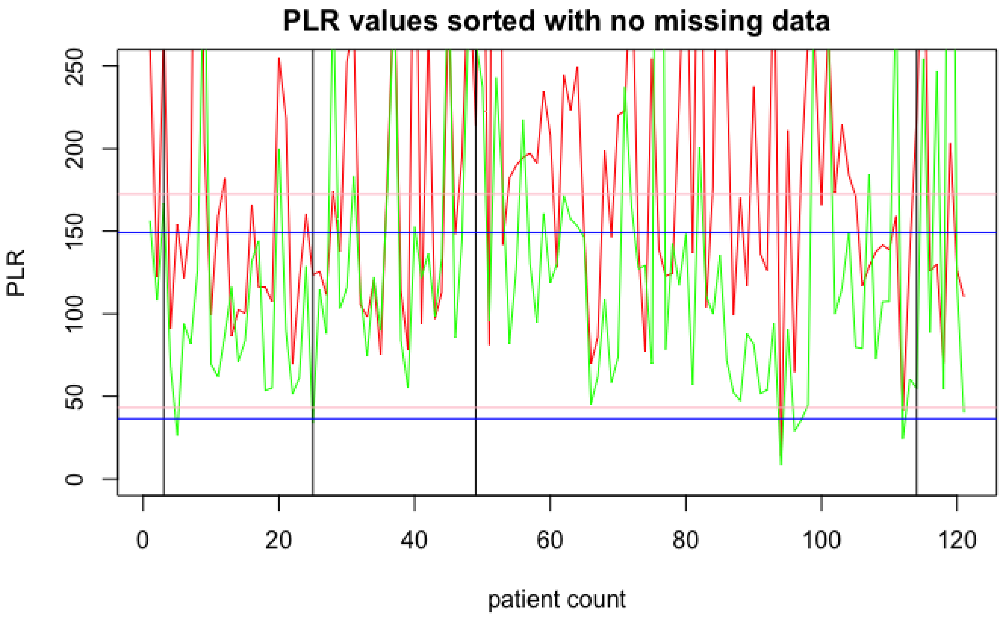


Figure B.187: Same plot as Figure B.186, but with the y-values restricted between 0 and 250.

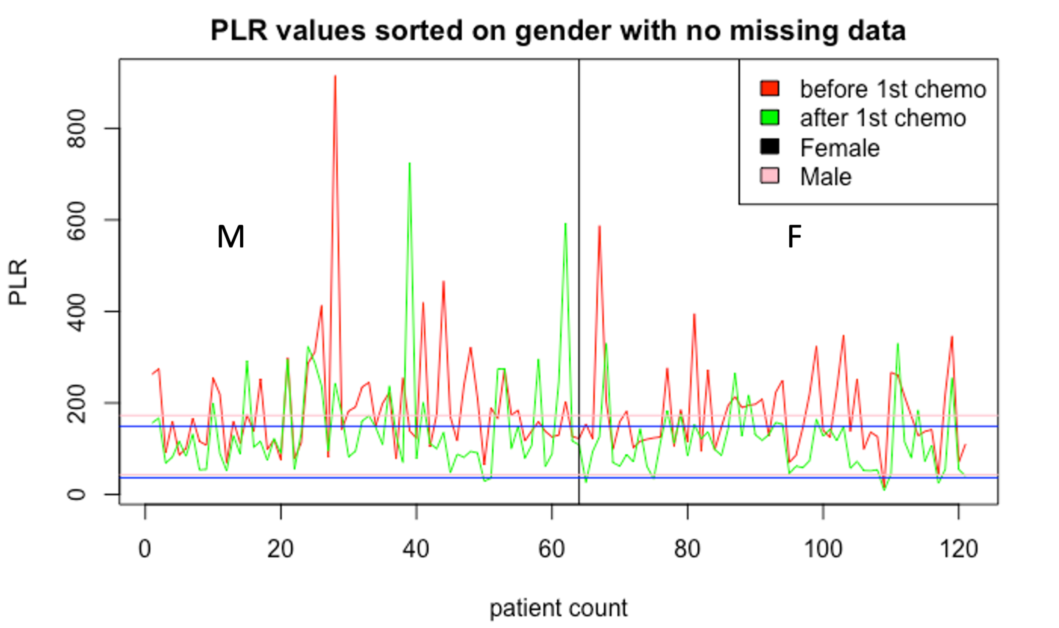


Figure B.188: Platelet-to-Lymphocyte ratio before and after the first chemotherapy cycle sorted by gender with no missing values, n=121 (63 male, 58 female). Red = values before the first cycle, Green = values after the first chemotherapy cycle, Pink = healthy female range, Blue = healthy male range.

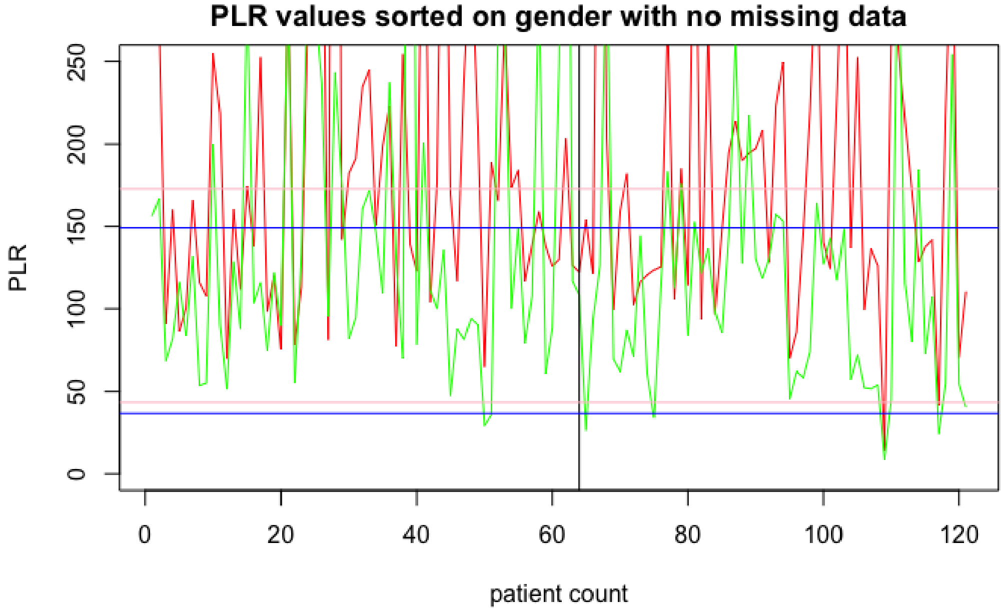


Figure B.189: Same plot as Figure B.188 but with the y-values restricted between 0-250.

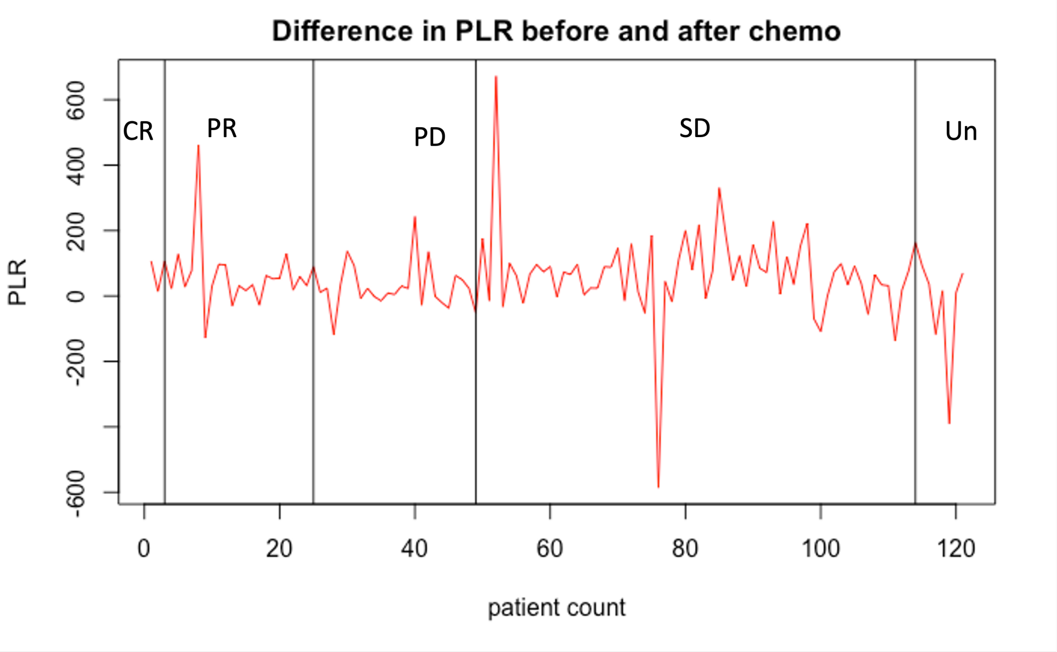


Figure B.190: Difference in Platelet-to-Lymphocyte ratios between the values before and after the first chemotherapy cycle for all the patients with no missing data, sorted by their final response, n=121. The final response is classified as Complete Response (CR, n=2), Partial response (PR, n=22), Progressive Disease (PD, n=24), Stable Disease (SD, n=65), Unknown (Un, n=8).

B.2. Outlier Analysis (full)

The section contains the full outlier analysis performed on the provided dataset.

B.2.1 Tumor Markers: CA19-9 and CEA Outlier analysis

This subsection contains the outlier analysis of the measured tumor markers CA19-9 and CEA at diagnosis, before and after the first chemotherapy cycle.

B.2.1.1 CA19-9

CA19-9 at diagnosis		**CA19-9 before**		**CA19-9 after**	
Patient ID	*Value (kU/L)*	*Patient ID*	*Value (kU/L)*	*Patient ID*	*Value (kU/L)*
001PANC0002	6430	001PANC0002	10448	001PANC0002	9746
001PANC0035	5815	001PANC0035	26459	001PANC0035	35717
001PANC0009	4803	001PANC0009	6416	001PANC0009	5694
001PANC0025	15121	001PANC0025	20427	001PANC0025	15594
001PANC0037	5146	001PANC0037	11050	001PANC0037	5190
001PANC0004	5456	001PANC0004	10316	001PANC0004	6646
001PANC0019	84341	001PANC0019	83313	001PANC0019	85225
001PP20022	4146	001PANC0011	5265	001PP20022	12401
151PP20011	2790	001PP20022	8332	151PP20011	4643
165PP20010	3800	151PP20011	4562	165PP20010	95696
		165PP20010	45305		

Table B.66: Overview of the outlier values determined using the IQR-method for the CA19-9 (kU/L) tumor marker measured at diagnosis, before and after the first chemotherapy cycle.

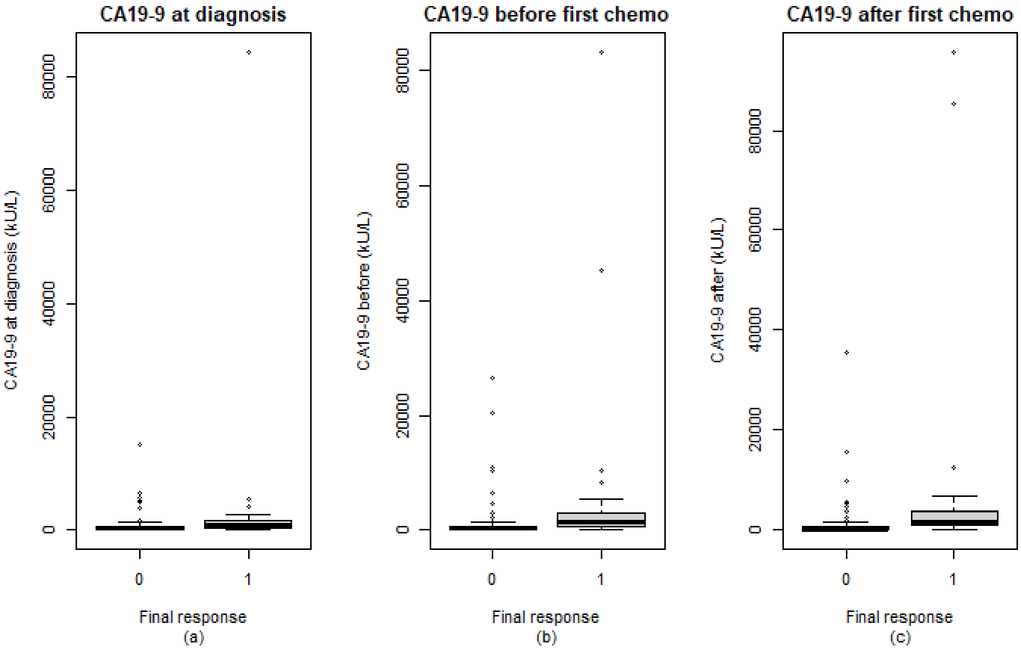


Figure B.191: Boxplot of the CA19-9 (kU/L) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=62), 1 = Progressive disease (n=21), total n=83)

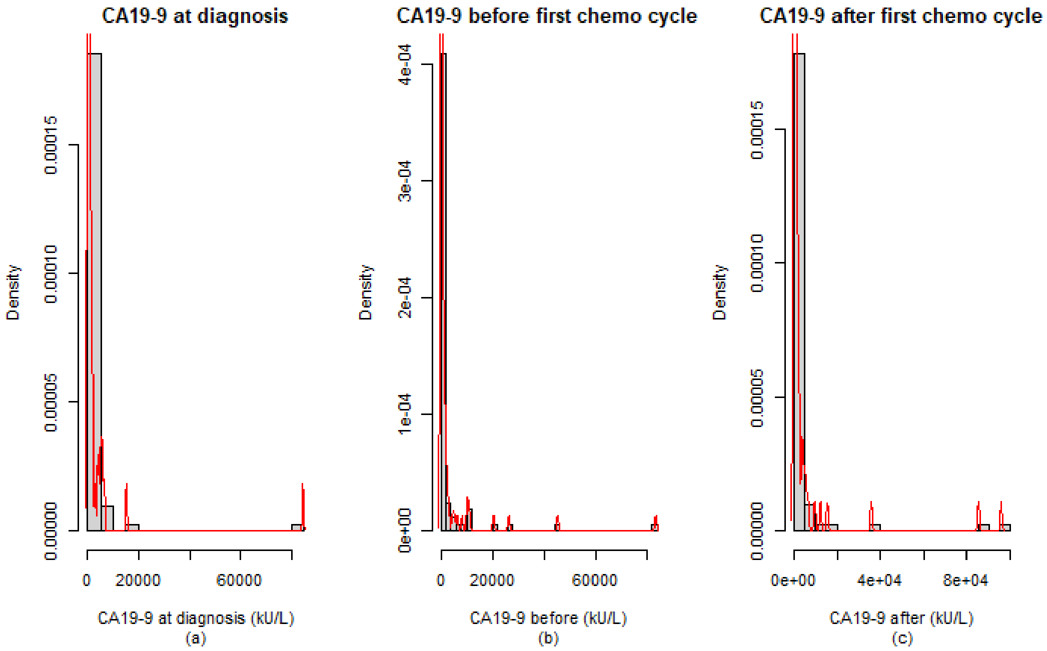


Figure B.192: Histogram of the CA19-9 (kU/L) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=62), 1 = Progressive disease (n=21), total n=83)

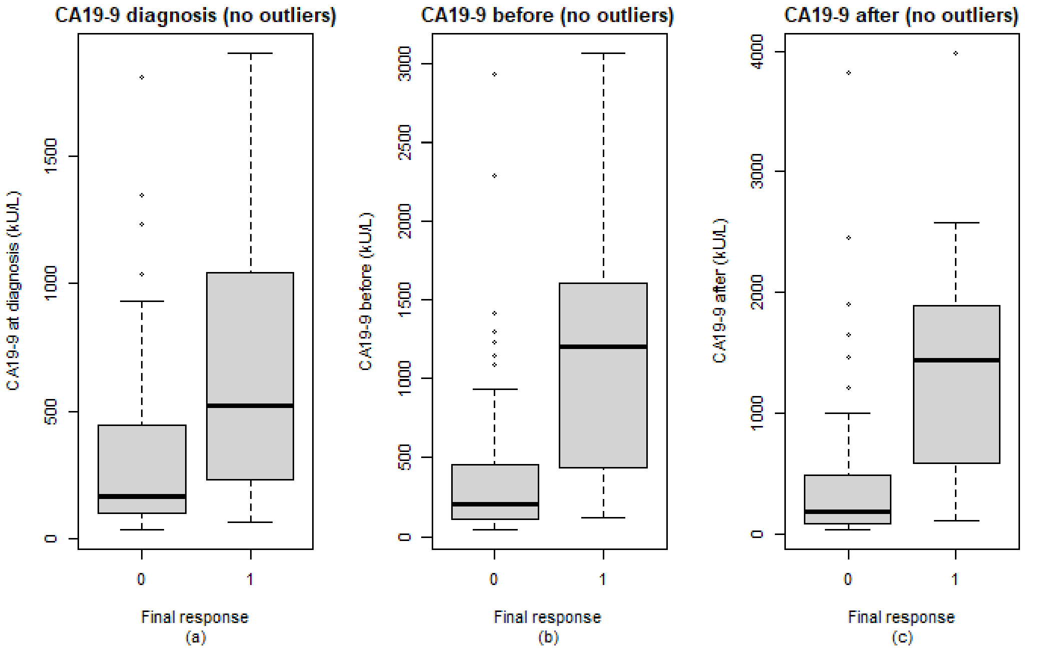


Figure B.193: Boxplot of the CA19-9 (kU/L) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle. Plots are categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=56), 1 = Progressive Disease (n=15), total n=71)

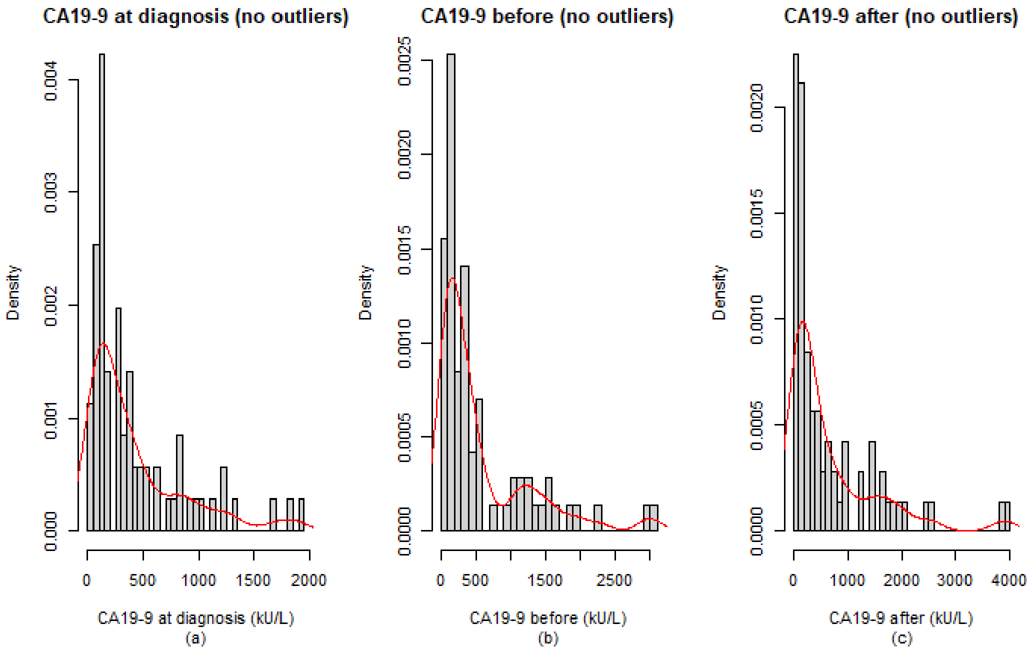


Figure B.194: Histogram of the CA19-9 (kU/L) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=56), 1 = Progressive Disease (n=15), total n=71).

B.2.1.2 CEA

CEA at diagnosis		**CEA before**		**CEA after**	
Patient ID	*Value (µg/L)*	*Patient ID*	*Value (µg/L)*	*Patient ID*	*Value (µg/L)*
001PANC0002	31.0	001PANC0002	41.5	001PANC0002	40.1
001PANC0035	29.0	001PANC0035	93.2	001PANC0035	98.8
001PANC0016	23.1	001PANC0016	20.3	001PANC0016	21.4
001PANC0019	255.0	001PANC0032	79.6	001PANC0032	67.0
001PP20022	35.0	001PANC0004	22.9	001PANC0004	20.1
001PP20017	16.9	001PANC0033	45.8	001PANC0033	36.4
078PANC0001	112.7	001PANC0019	226.0	001PANC0019	241.0
078PP20028	17.3	001PP20022	46.4	001PP20022	56.2
001PP20045	21.0	001PP20017	20.4	001PP20017	21.6
		078PANC0001	104.8	001PP20045	139.0
		001PP20045	119.0		

Table B.67: Overview of the outlier values determined using the IQR-method for the CEA ($\mu g/L$) tumor marker measured at diagnosis, before and after the first chemotherapy cycle.

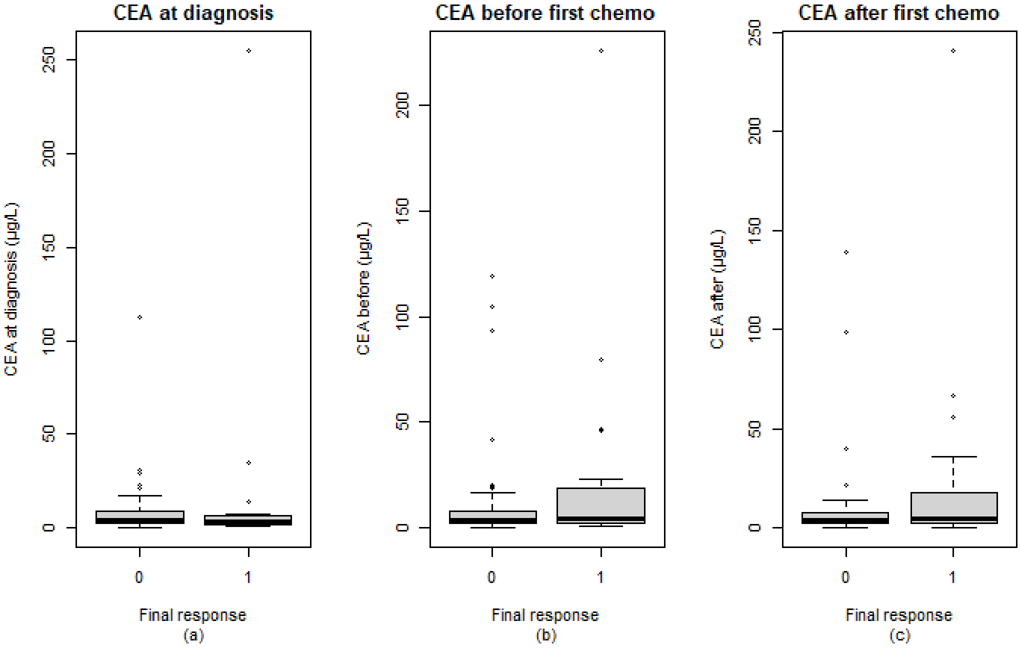


Figure B.195: Boxplot of the CEA ($\mu g/L$) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=76), 1 = Progressive Disease (n=20), total n=96).

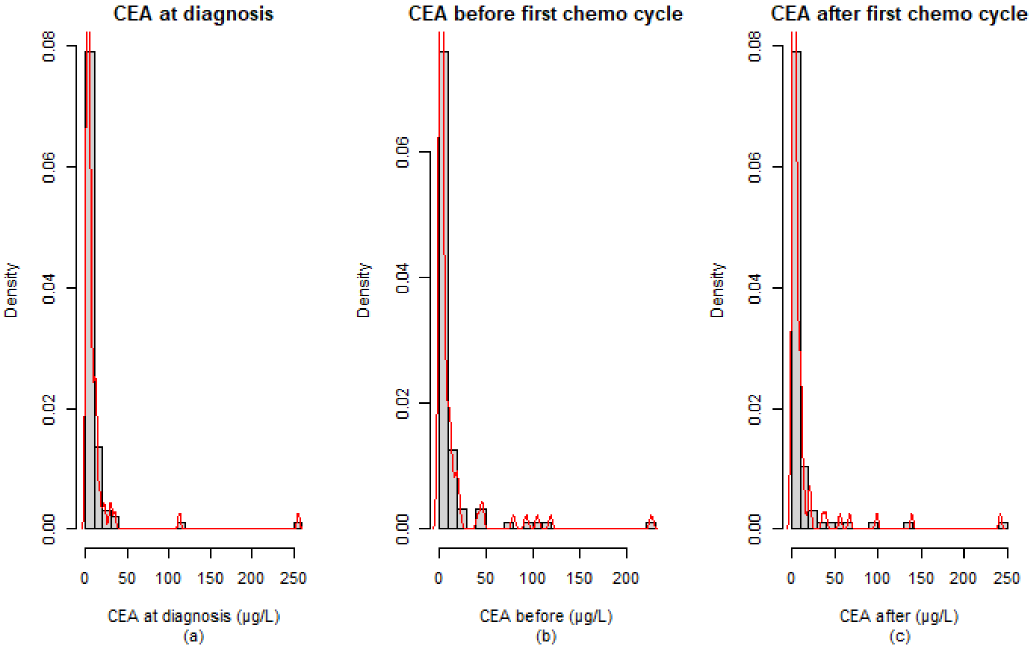


Figure B.196: Histogram of the CEA ($\mu g/L$) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=76), 1 = Progressive Disease (n=20), total n=96)

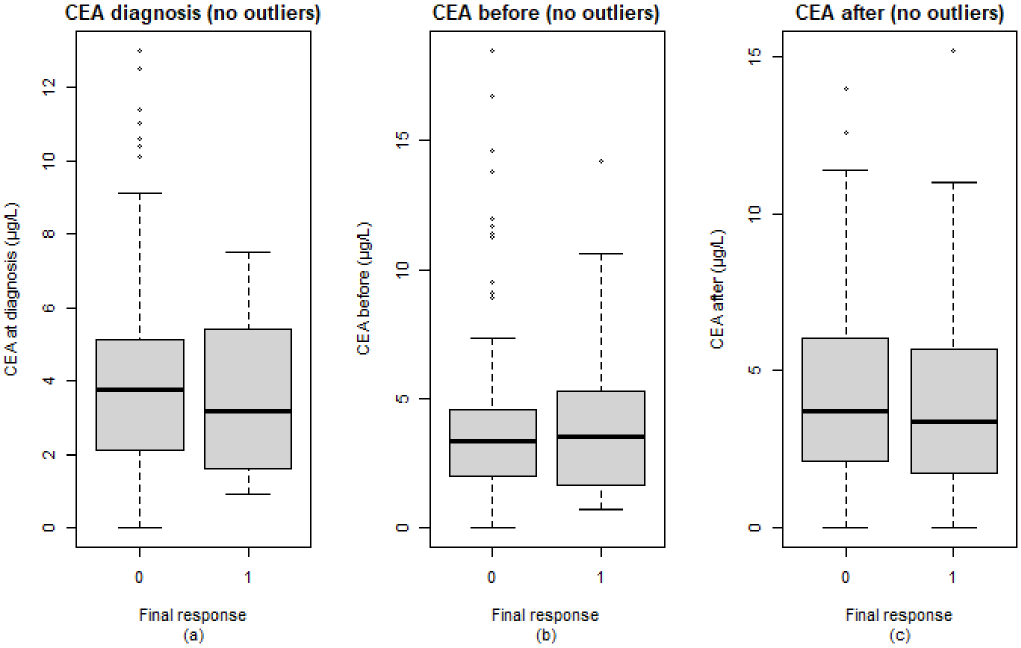


Figure B.197: Boxplot of the CEA ($\mu g/L$) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=69), 1 = Progressive Disease (n=15), total n=84).

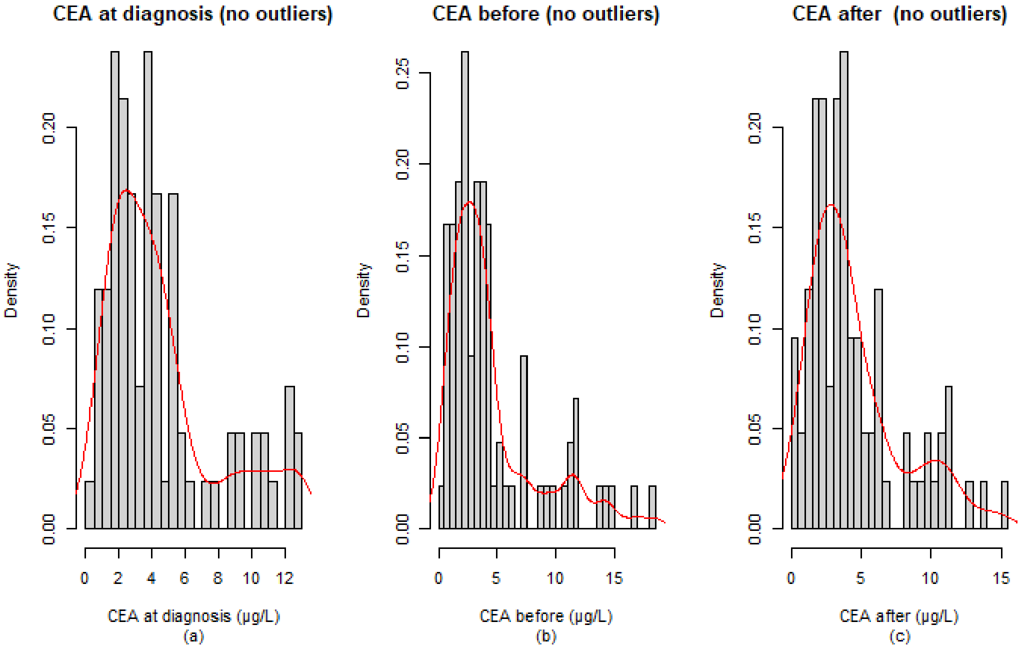


Figure B.198: Histogram of the CEA ($\mu g/L$) data (a) at diagnosis, (b) before the first chemotherapy cycle and (c) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=69), 1 = Progressive Disease (n=15), total n=84).

B.2.2 Blood Markers

This sub-section contains the outlier analysis of the measured blood markers before and after the first chemotherapy cycle as well as the added inflammation indices: Systemic Inflammation Index (SII), Neutrophils-to-Lymphocyte Ratio (NLR) and Platelet-to-Lymphocyte Ratio (PLR). Note that the International Normalized Ratio (INR) is not taken into consideration due to the lack of data.

B.2.2.1 Hemoglobin

HB before		**HB after**	
Patient ID	*Value (mmol/L)*	*Patient ID*	*Value (mmol/L)*
078PP20007	10.7	078PP20028	9.8

Table B.68: Overview of the outlier values determined using the IQR-method for the Hemoglobin values ($mmol/L$) measured before and after the first chemotherapy cycle.

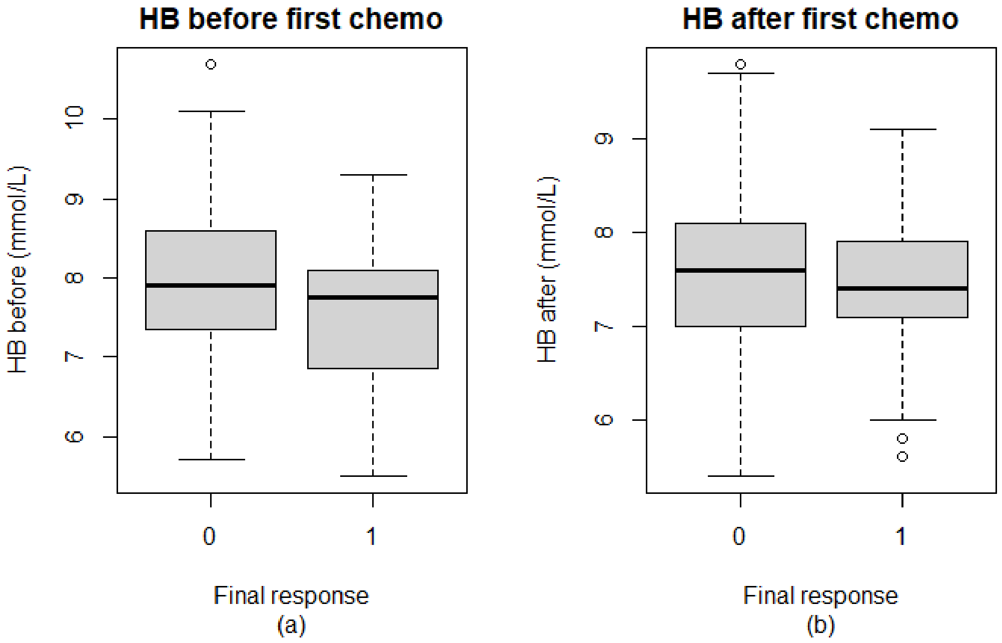


Figure B.199: Boxplot of the Hemoglobin (HB) ($mmol/L$) data (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=167), 1 = Progressive Disease (n=40), total n=207).

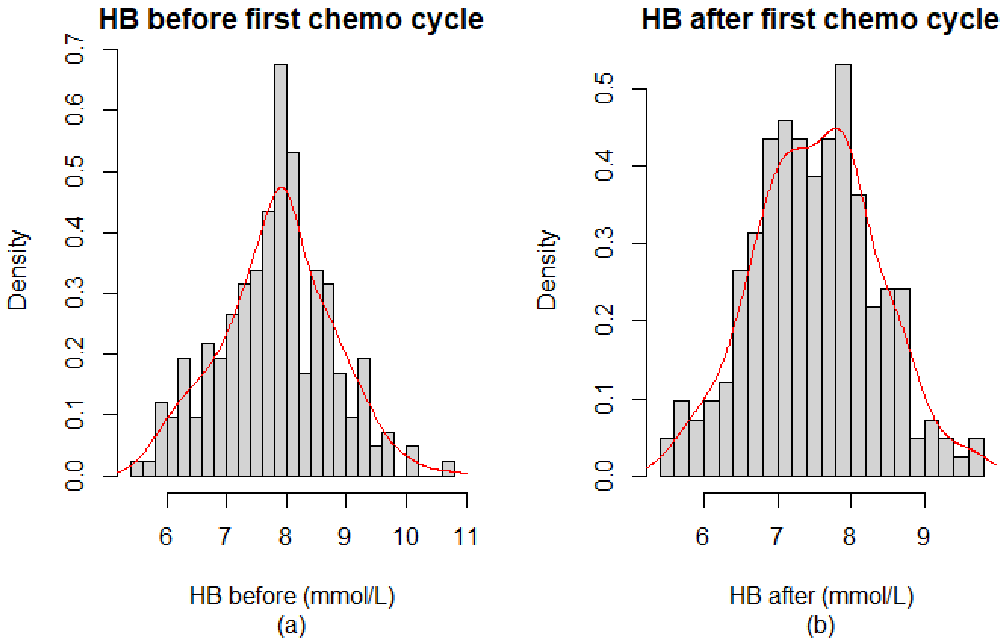


Figure B.200: Histogram of the Hemoglobin (HB) ($mmol/L$) data (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density in red. (0 = Disease control (n=167), 1 = Progressive Disease (n=40), total n=207)

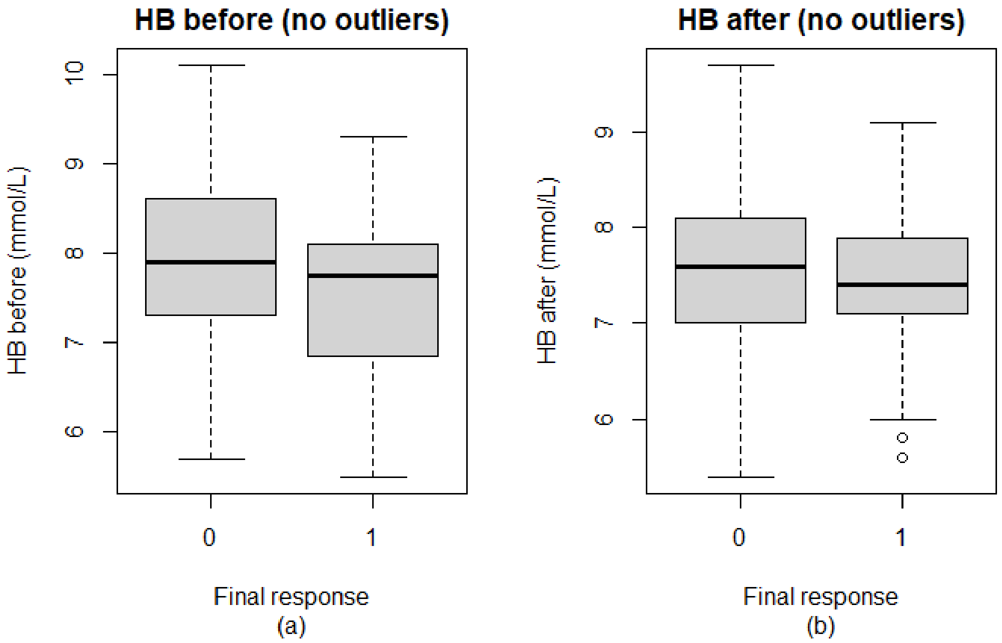


Figure B.201: Boxplot of the Hemoglobin (HB) ($mmol/L$) data (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=165), 1 = Progressive Disease (n=40), total n=205).

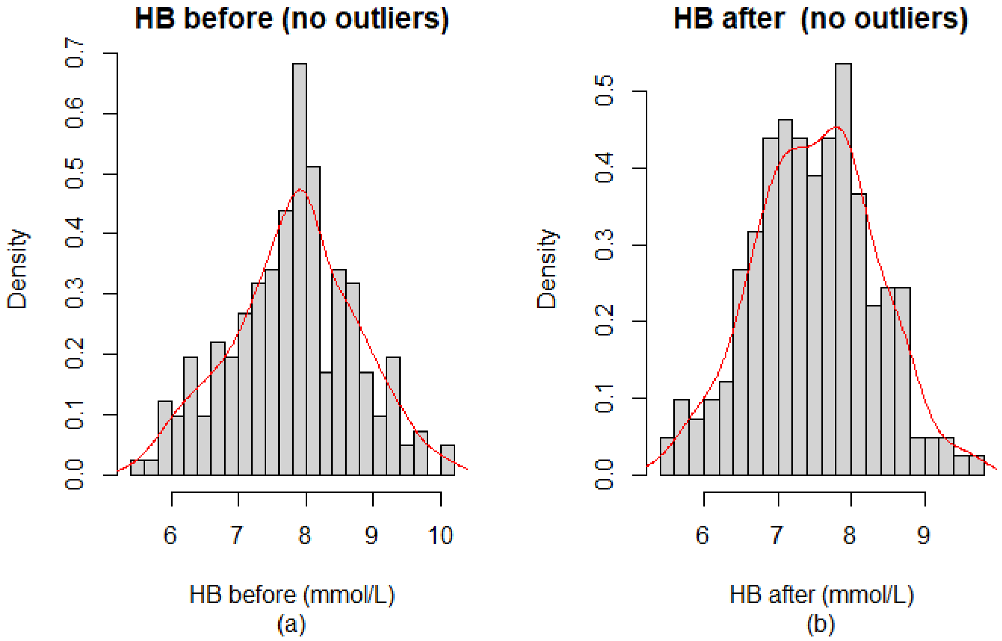


Figure B.202: Histogram of the Hemoglobin (HB) ($mmol/L$) data (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=165), 1 = Progressive Disease (n=40) , total n=205)

B.2.2.2 Thrombocytes

TB before		**TB after**	
Patient ID	*Value* $(10^9/L)$	*Patient ID*	*Value* $(10^9/L)$
002PP20011	508	001PANC0004	397
002PP20005	511	002PP20008	630
065PP20005	43	018PP20004	455
001PANC0051	593	059PP20001	480
001PP20036	488	078PANC0002	439
133PP20007	709	001PP20036	482
151PP20017	521	151PP20018	417
		165PP20006	629
		165PP20014	394
		165PP20018	527

Table B.69: Overview of the outlier values determined using the IQR-method for the Thrombocyte count $(10^9/L)$ measured before and after the first chemotherapy cycle.

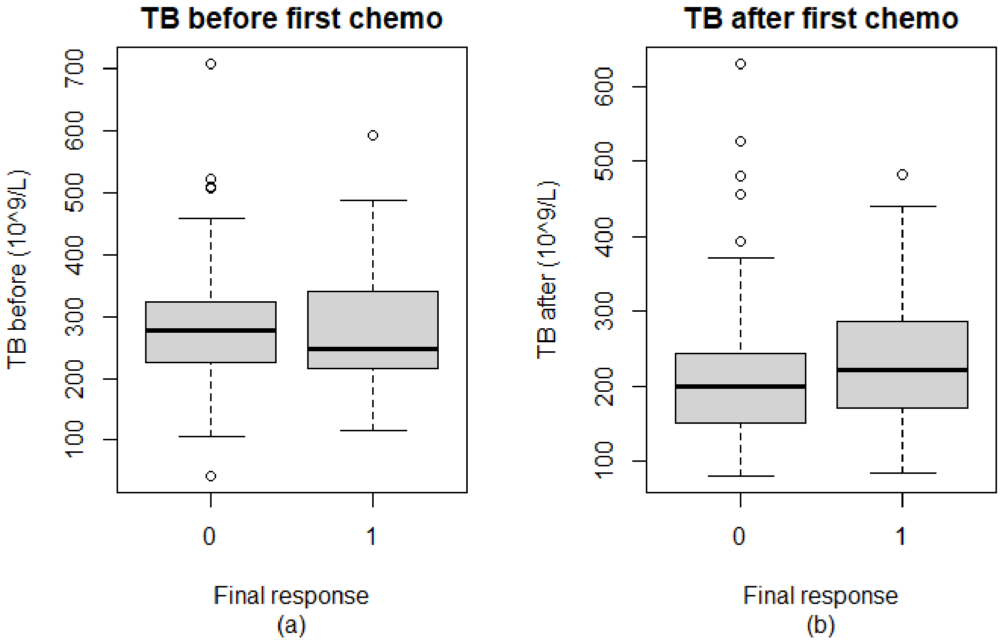


Figure B.203: Boxplot of the Thrombocytes (TB) $(10^9/L)$ count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=168), 1 = Progressive Disease (n=40), total n=208).

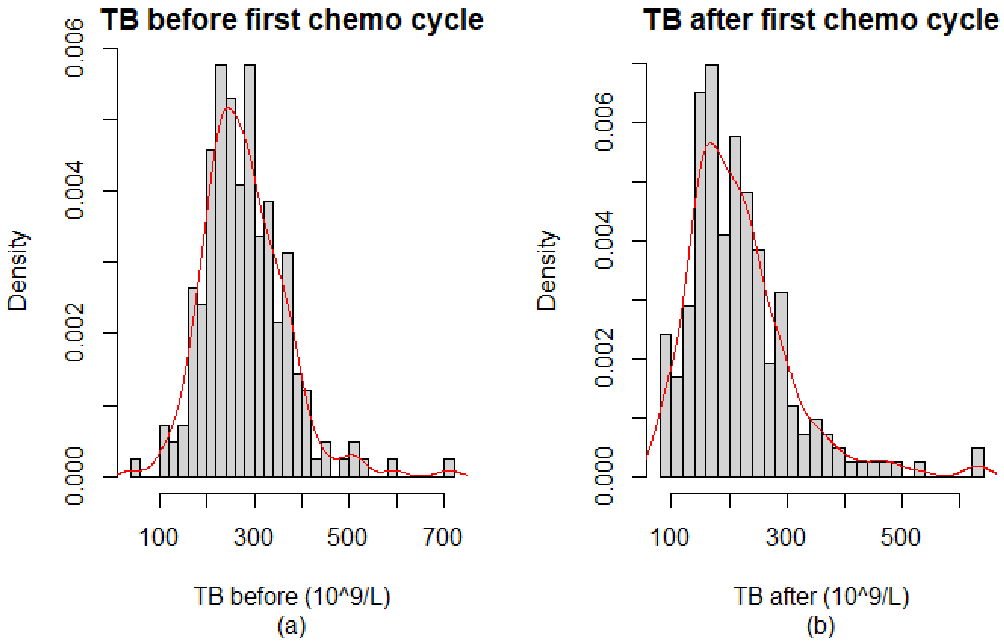


Figure B.204: Histogram of the Thrombocytes (TB) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=168), 1 = Progressive Disease (n=40), total n=208)

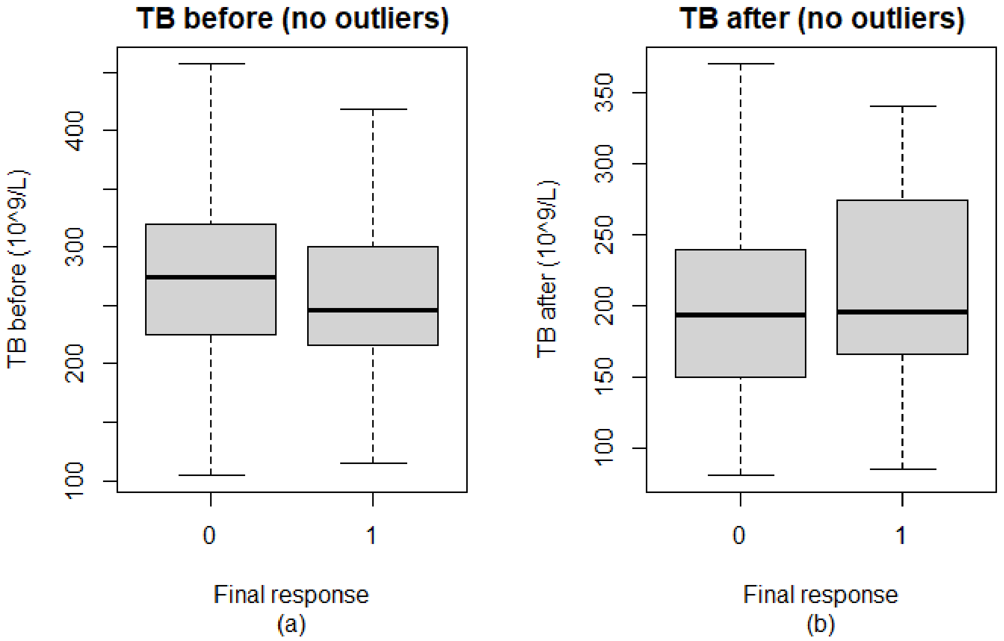


Figure B.205: Boxplot of the Thrombocytes (TB) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=157), 1 = Progressive Disease (n=35), total n=192)

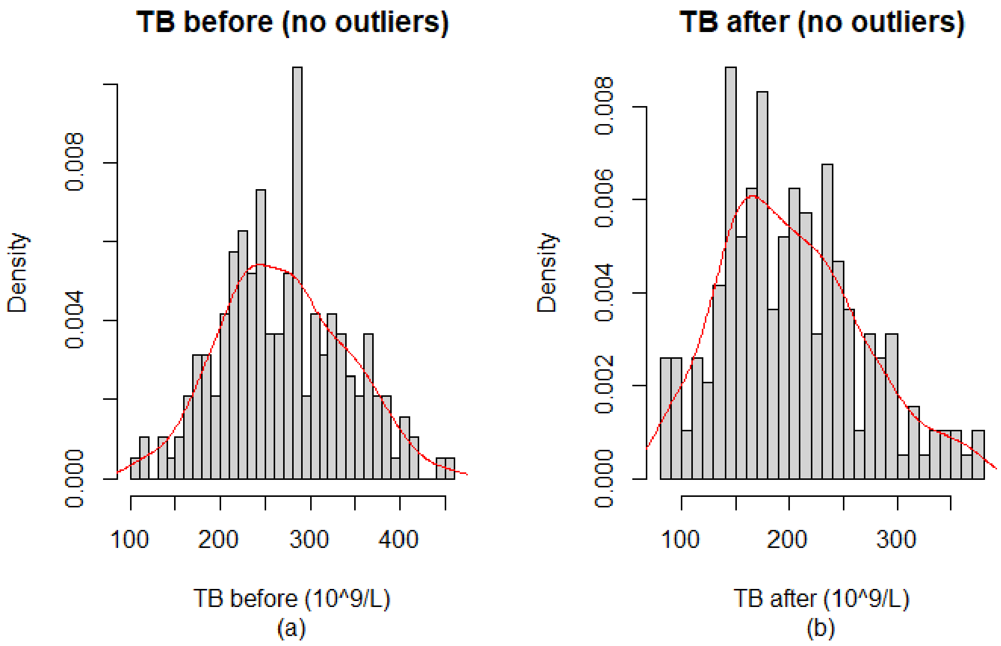


Figure B.206: Histogram of the Thrombocytes (TB) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density in red. (0 = Disease control (n=157), 1 = Progressive Disease (n=35), total n=192)

B.2.2.3 Leukocytes

LK before		**LK after**	
Patient ID	*Value* ($10^9/L$)	*Patient ID*	*Value* ($10^9/L$)
001PANC0018	16.60	001PANC0014	41.00
078PP20007	14.22	001PP20017	41.00
001PANC0052	2.50	078PP20004	39.16
001PANC0039	14.90		
001PP20036	14.30		
148PANC0003	13.80		
148PP20013	13.20		
148PP20019	14.00		
001PANC0061	14.40		

Table B.70: Overview of the outlier values determined using the IQR-method for the Leukocyte count ($10^9/L$) measured before and after the first chemotherapy cycle.

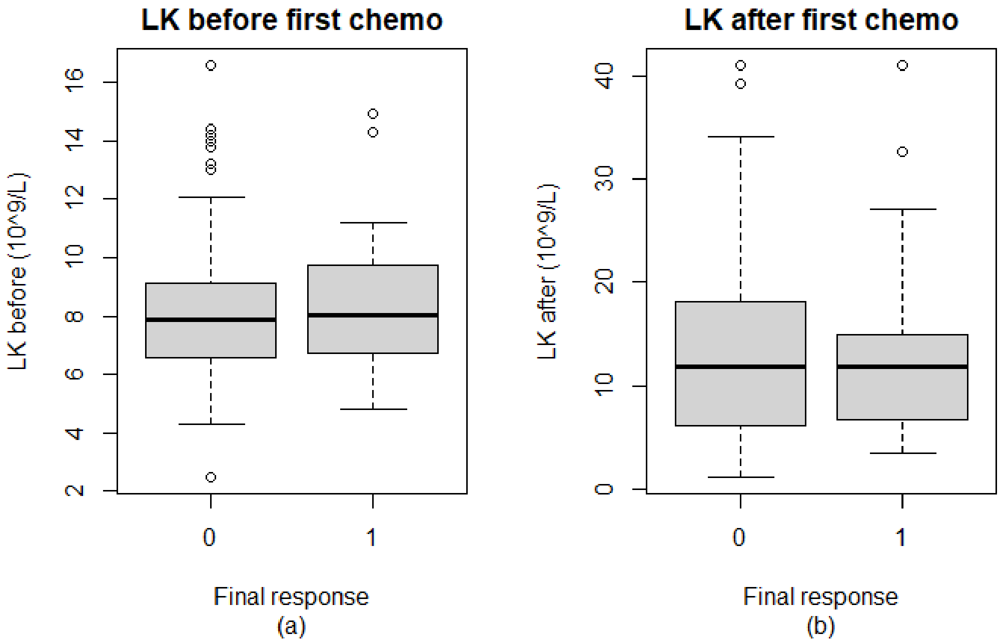


Figure B.207: Boxplot of the Leukocytes (LK) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=170), 1 = Progressive Disease (n=40), total n=210)

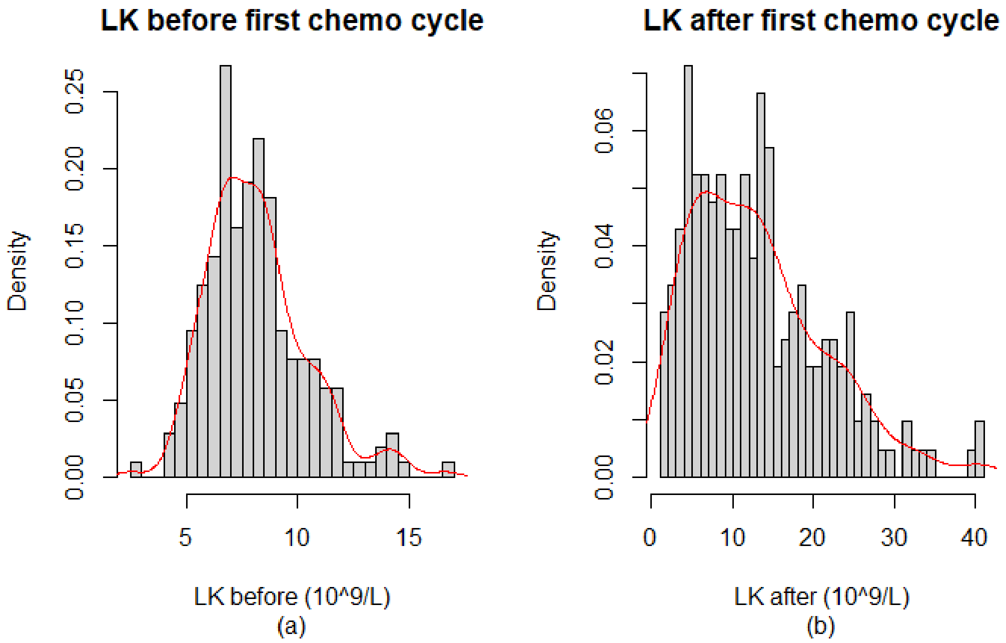


Figure B.208: Histogram of the Leukocytes (LK) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=170), 1 = Progressive Disease (n=40), total n=210)

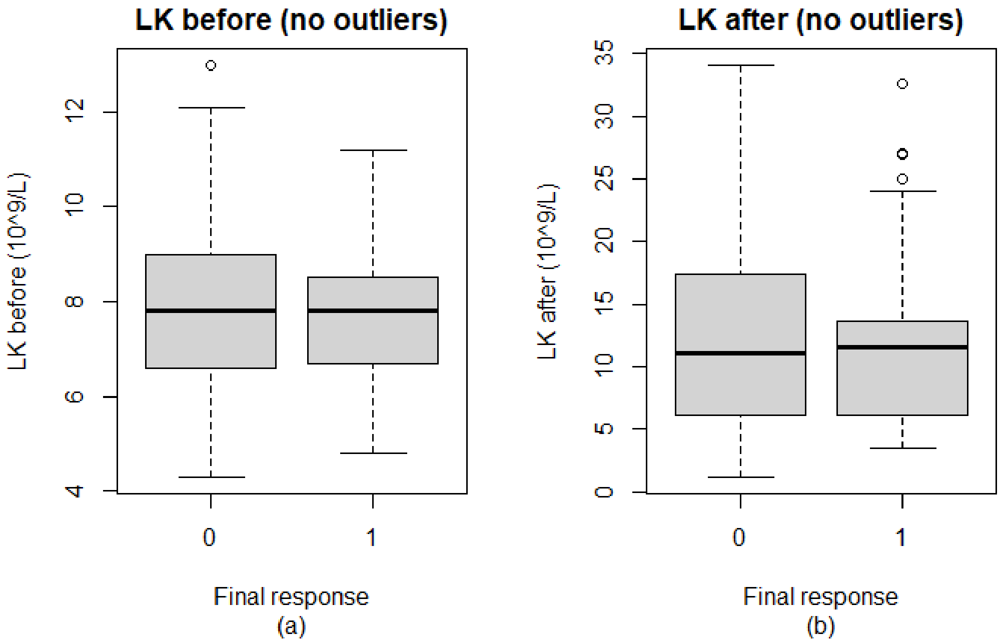


Figure B.209: Boxplot of the Leukocytes (LK) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=161), 1 = Progressive Disease (n=37), total n=198)

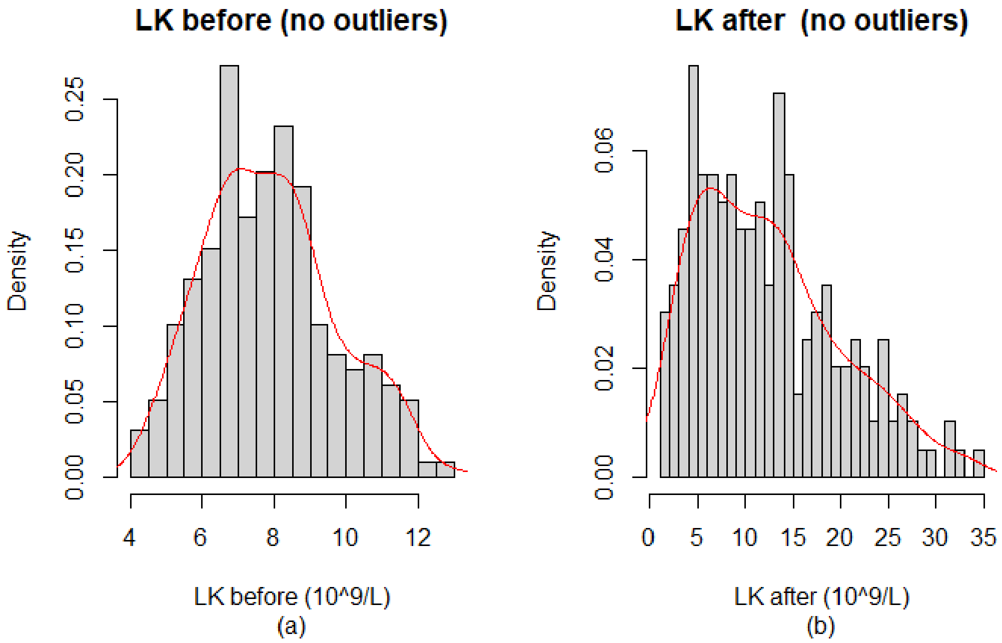


Figure B.210: Histogram of the Leukocytes (LK) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=161), 1 = Progressive Disease (n=37), total n=198)

B.2.2.4 Neutrophils

NP before		**NP after**	
Patient ID	*Value* $(10^9/L)$	*Patient ID*	*Value* $(10^9/L)$
001PANC0018	13.20	001PANC0014	29.00
001PANC0039	9.92	001PP20017	28.00
001PP20036	10.78	078PP20004	36.14
133PP20005	9.60	148PANC0004	30.61
148PP20019	9.42	078PP20039	27.66
001PANC0061	11.30		

Table B.71: Overview of the outlier values determined using the IQR-method for the Neutrophil count $(10^9/L)$ measured before and after the first chemotherapy cycle.

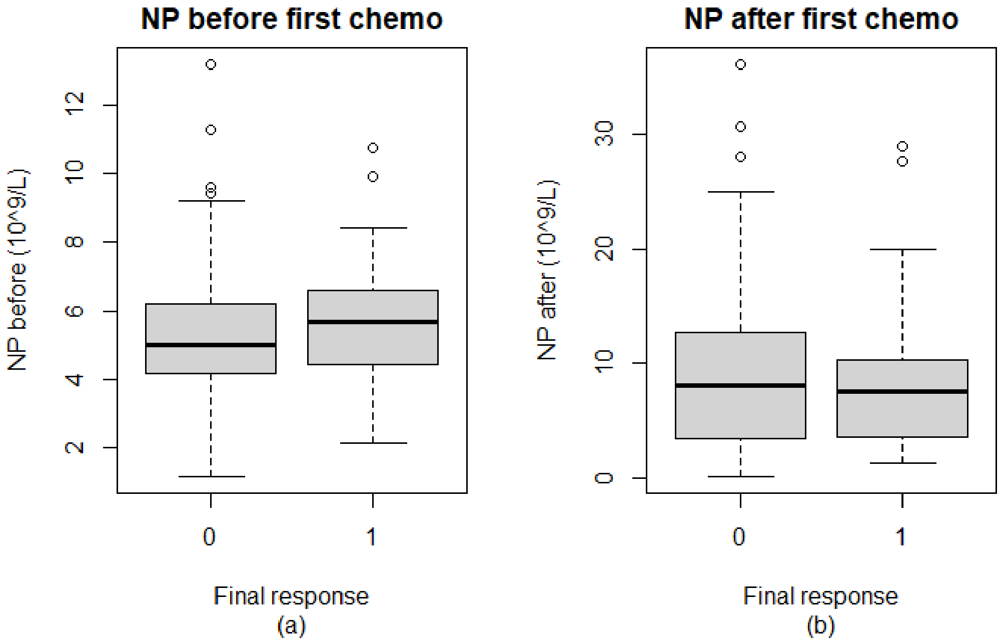


Figure B.211: Boxplot of the Neutrophils (NP) $(10^9/L)$ count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=153), 1 = Progressive Disease (n=35), total n=188)

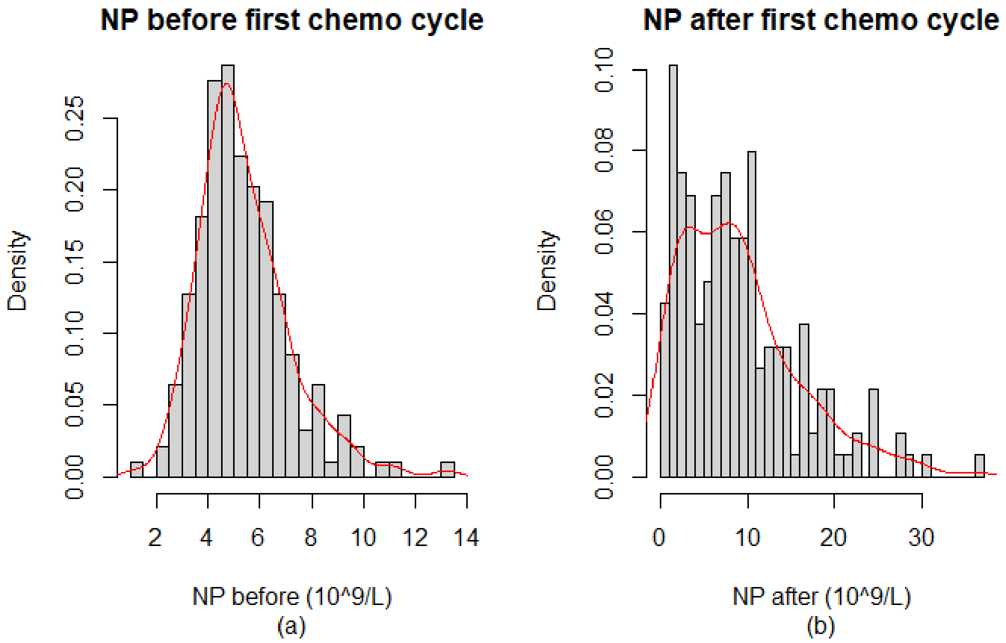


Figure B.212: Histogram of the Neutrophils (NP) $(10^9/L)$ count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=153), 1 = Progressive Disease (n=35), total n=188)

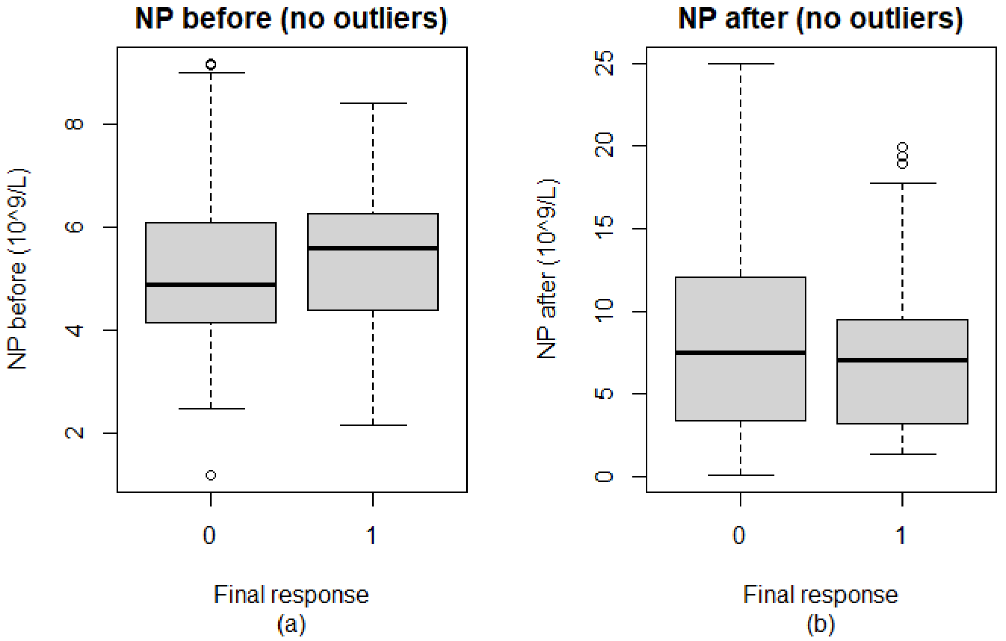


Figure B.213: Boxplot of the Neutrophils (NP) $(10^9/L)$ count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=146), 1 = Progressive Disease (n=31), total n=177)

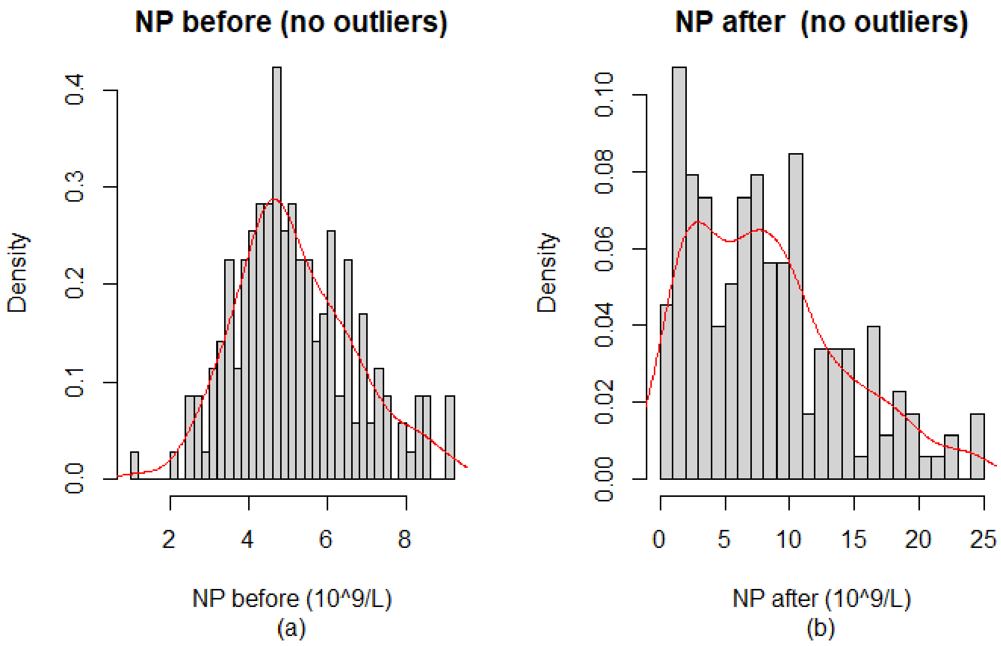


Figure B.214: Histogram of the Neutrophils (NP) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=146), 1 = Progressive Disease (n=31), total n=177)

B.2.2.5 Lymphocytes

LC before		**LC after**	
Patient ID	*Value* ($10^9/L$)	*Patient ID*	*Value* ($10^9/L$)
001PP20017	3.73	001PANC0014	4.74
001PANC0047	3.68	001PP20017	5.27
148PANC0003	9.20	148PANC0003	11.00
148PP20019	3.70	148PP20019	6.80
001PP20045	4.43		
151PP20034	5.90		

Table B.72: Overview of the outlier values determined using the IQR-method for the Lymphocyte count ($10^9/L$) measured before and after the first chemotherapy cycle.

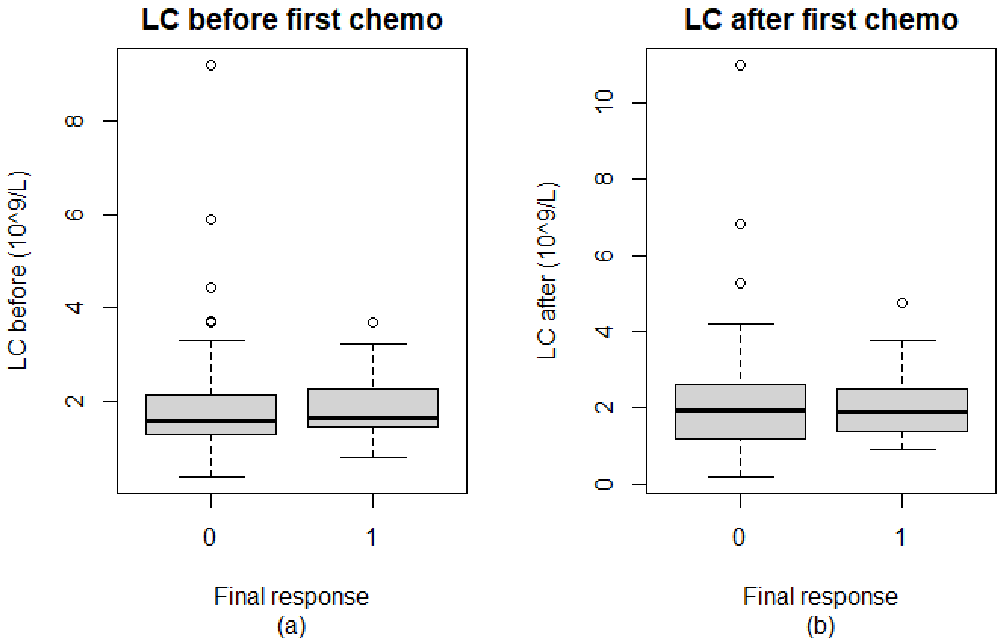


Figure B.215: Boxplot of the Lymphocytes (LC) $(10^9/L)$ count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=90), 1 = Progressive Disease (n=24), total n=114)

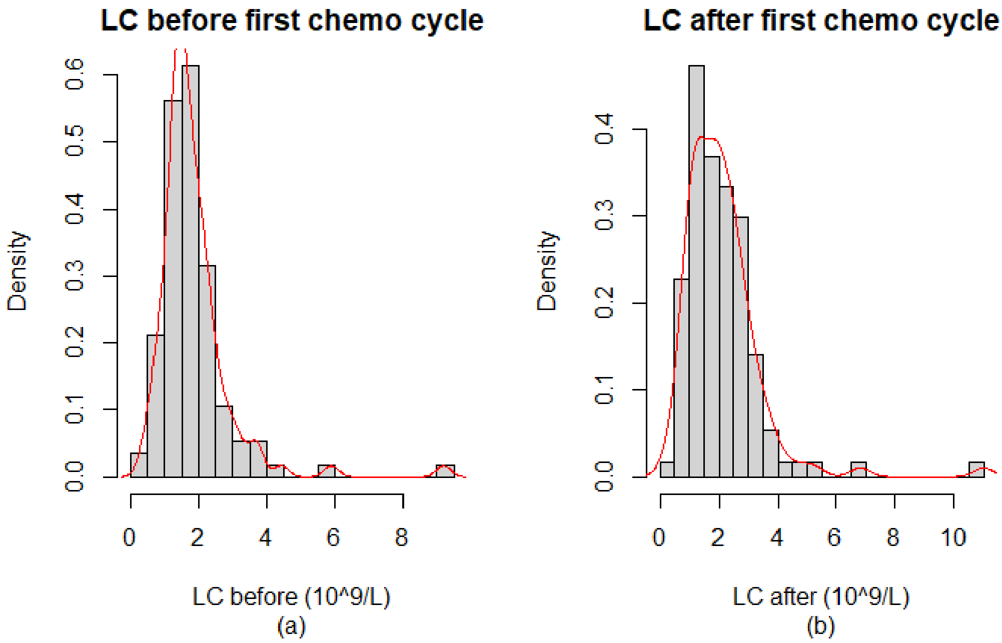


Figure B.216: Histogram of the Lymphocytes (LC) $(10^9/L)$ count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=90), 1 = Progressive Disease (n=24), total n=114)

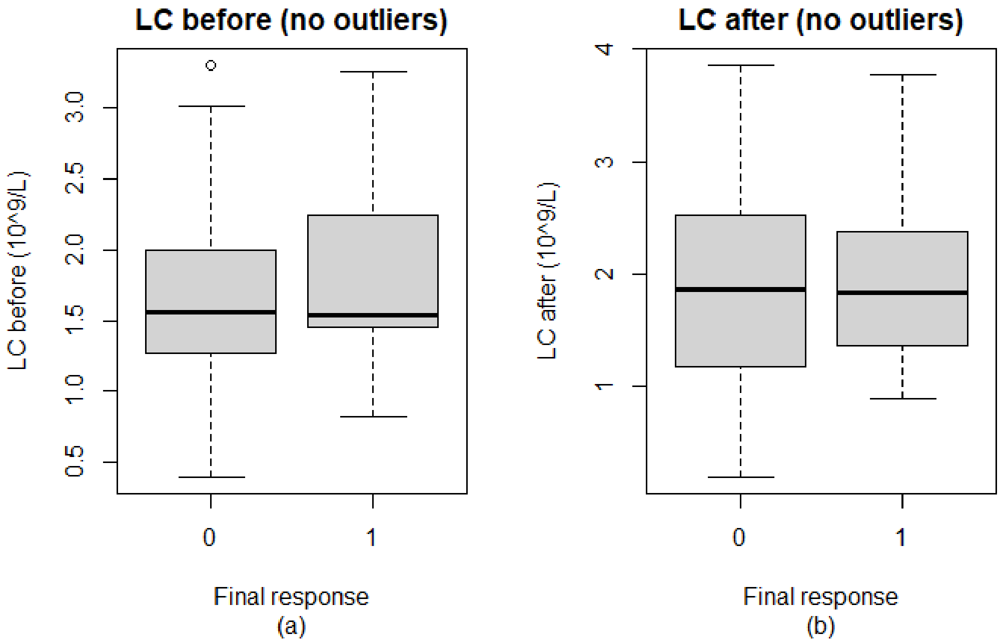


Figure B.217: Boxplot of the Lymphocytes (LC) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=85), 1 = Progressive Disease (n=22), total n=107)

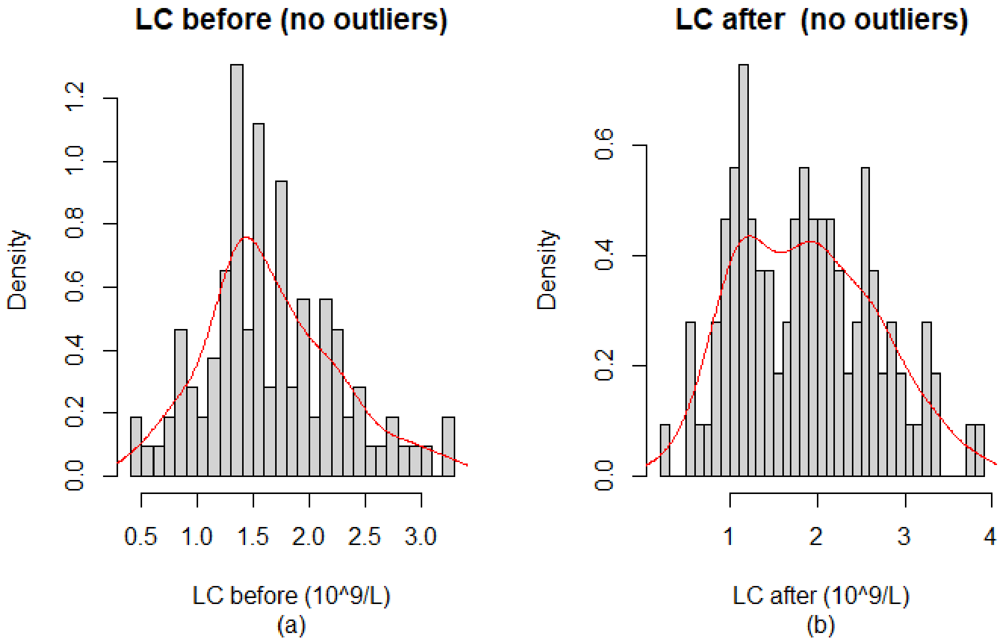


Figure B.218: Histogram of the Lymphocytes (LC) ($10^9/L$) count (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=85), 1 = Progressive Disease (n=22), total n=107)

B.2.2.6 Creatinin

CR before		**CR after**	
Patient ID	*Value ($\mu mol/L$)*	*Patient ID*	*Value ($\mu mol/L$)*
002PANC0007	119	001PP20008	172
001PANC0058	149	002PANC0007	127
151PP20013	124	001PANC0058	131
002PP20069	117	078PP20026	178
		001PP20040	952

Table B.73: Overview of the outlier values determined using the IQR-method for the Creatinin values ($\mu mol/L$) measured before and after the first chemotherapy cycle.

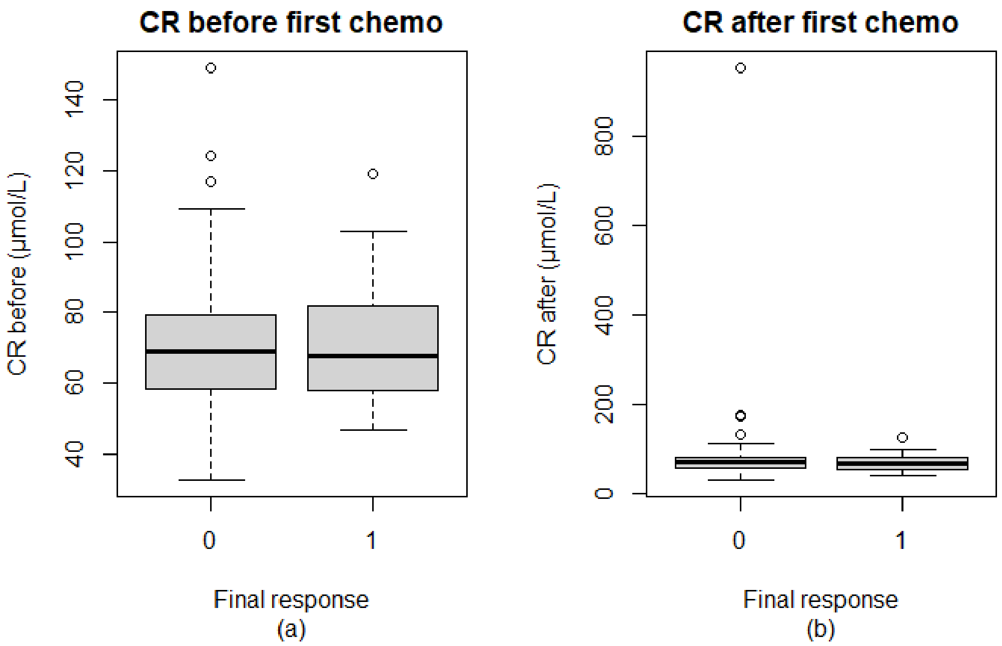


Figure B.219: Boxplot of the Creatinin (CR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=168), 1 = Progressive Disease (n=42), total n=210)

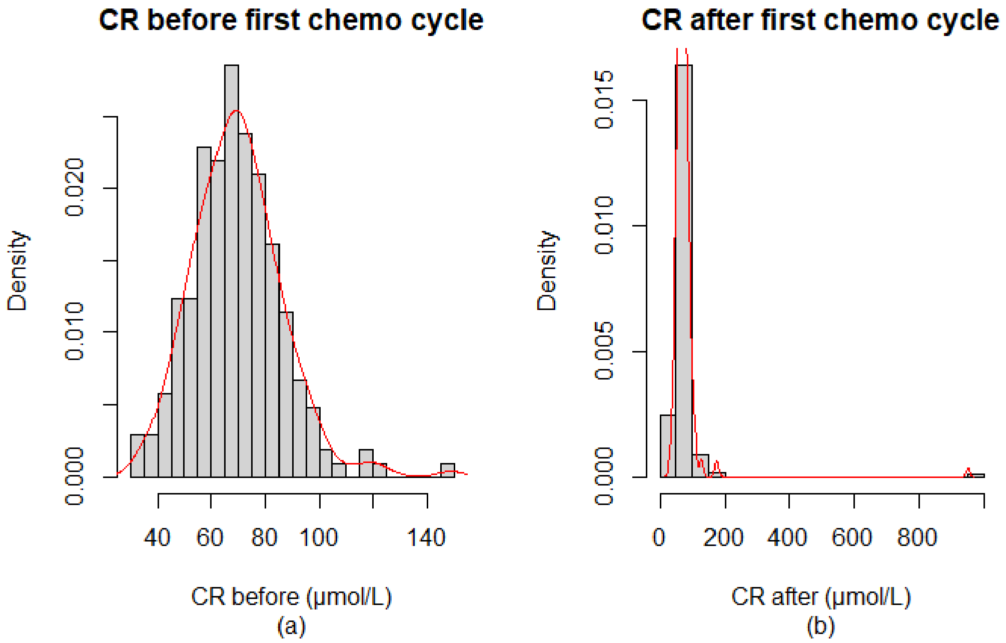


Figure B.220: Histogram of the Creatinin (CR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=168), 1 = Progressive Disease (n=42), total n=210)

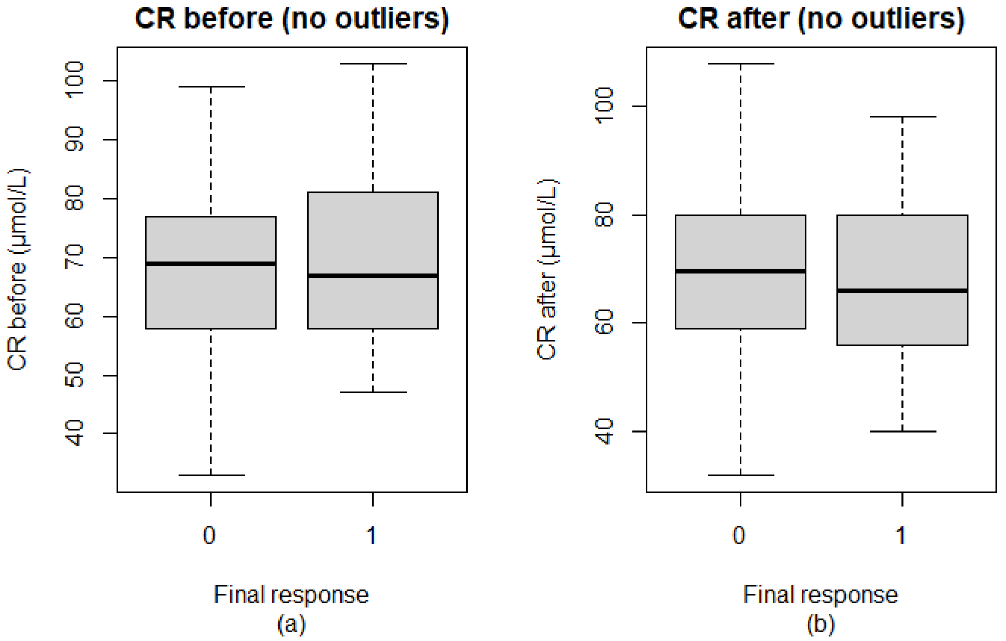


Figure B.221: Boxplot of the Creatinin (CR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=162), 1 = Progressive Disease (n=41), total n=203)

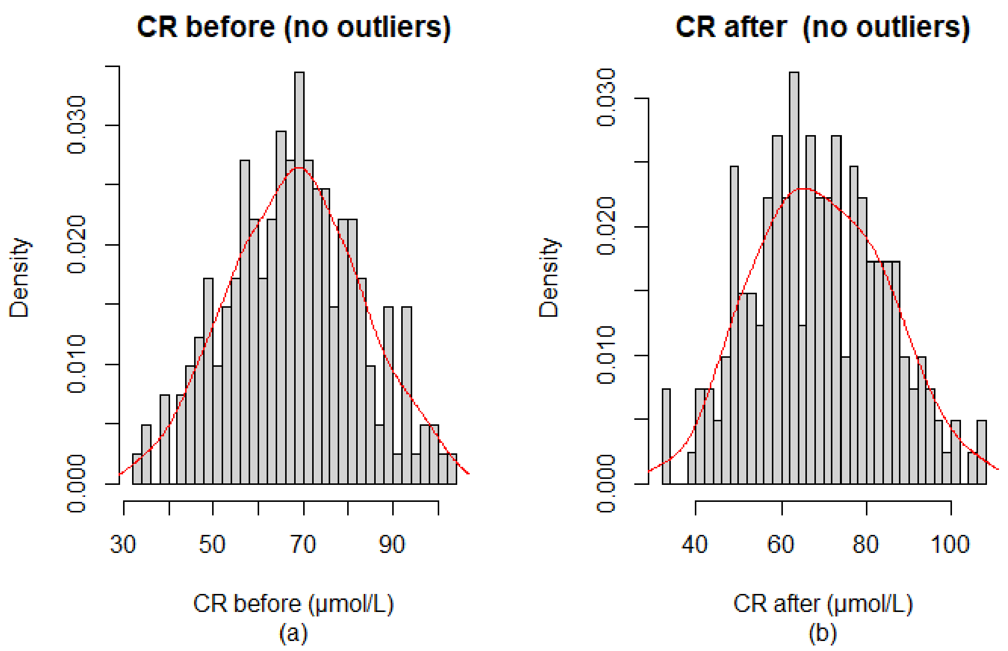


Figure B.222: Histogram of the Creatinin (CR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=162), 1 = Progressive Disease (n=41), total n=203)

B.2.2.7 Glomular Filtration Rate

GFR before		**GFR after**	
Patient ID	*Value (mL/min)*	*Patient ID*	*Value (mL/min)*
001PANC0002	106	001PANC0015	62
001PANC0023	62	001PANC0023	64
001PANC0010	115	001PANC0010	106
001PANC0013	105	001PANC0003	112
001PANC0003	111	001PANC0011	116
001PANC0011	55	001PP20022	124
001PP20013	105	001PP20008	35
001PP20022	126	002PANC0007	50
002PANC0007	54	065PP20003	58
002PP20007	65	001PANC0058	45
065PP20003	52	001PANC0042	58
151PP20009	109	001PANC0052	125
001PANC0058	39	001PANC0051	106
001PANC0042	54	001PANC0043	114
001PANC0052	124	001PP20029	57
001PANC0047	66	002PP20048	59
001PANC0051	106	002PP20063	60
001PANC0038	65	078PP20026	32
001PANC0043	122	078PP20028	58
001PP20029	53	133PP20006	60
078PP20029	66	151PP20013	64
151PP20013	54	001PP20046	45
151PP20021	58	002PP20069	57
001PP20046	51	005PP20021	61
001PP20040	58	151PP20034	50
002PP20069	54		

Table B.74: Overview of the outlier values determined using the IQR-method for the Glomular Filtration Rate values (mL/min) measured before and after the first chemotherapy cycle.

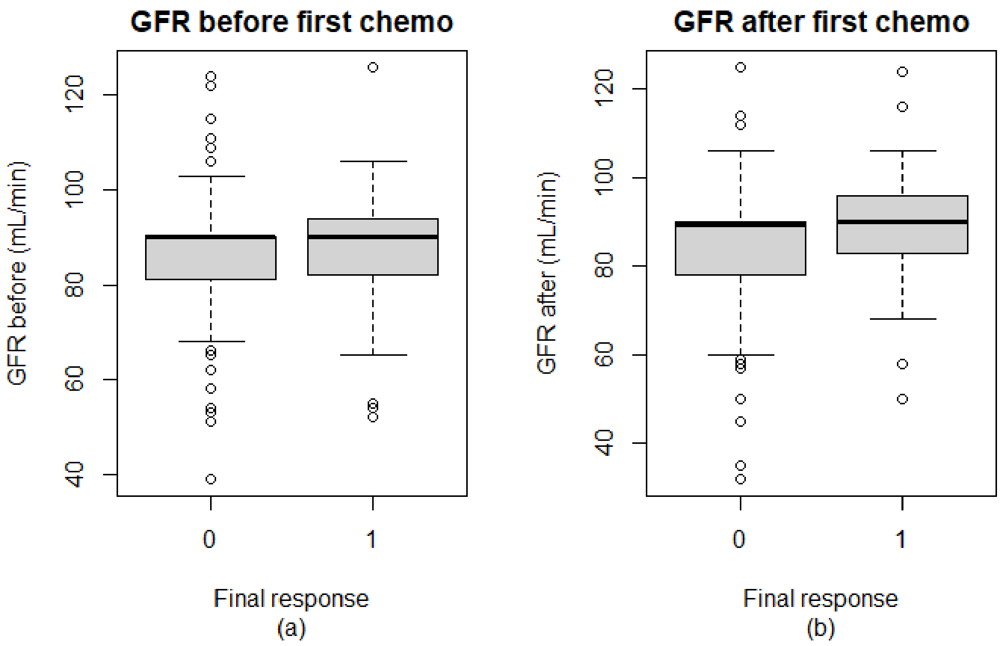


Figure B.223: Boxplot of the Glomular Filtration Rate (GFR) (mL/min) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=168), 1 = Progressive Disease (n=41), total n=209)

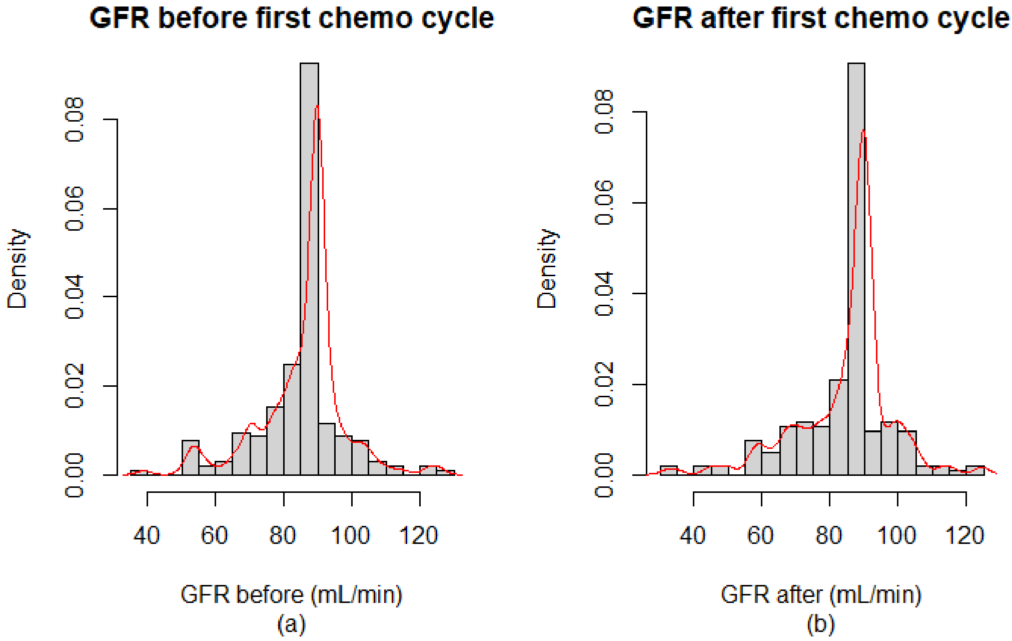


Figure B.224: Histogram of the Glomular Filtration Rate (GFR) (mL/min) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=168), 1 = Progressive Disease (n=41), total n=209)

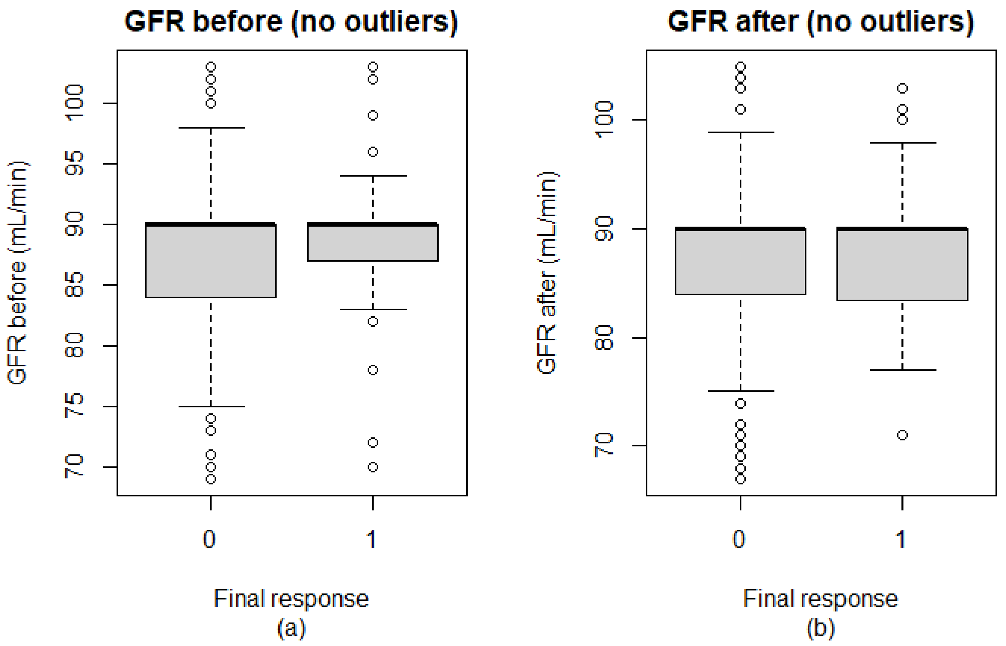


Figure B.225: Boxplot of the Glomular Filtration Rate (GFR) (mL/min) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=142), 1 = Progressive Disease (n=32), total n=174)

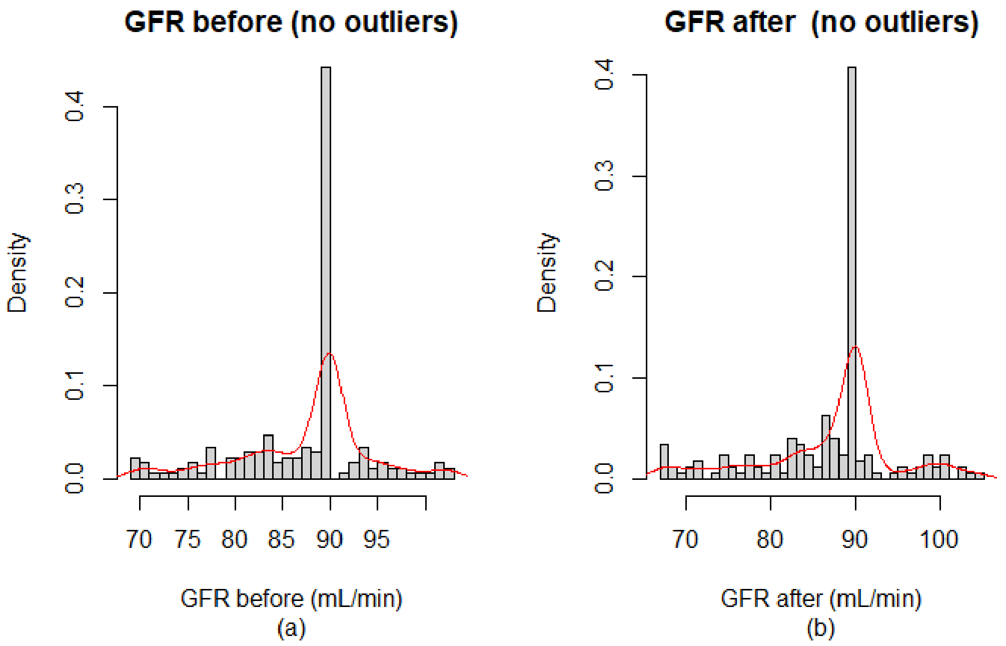


Figure B.226: Histogram of the Glomular Filtration Rate (GFR) (mL/min) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=142), 1 = Progressive Disease (n=32), total n=174)

B.2.2.8 Sodium

Na before		**Na after**	
Patient ID	*Value (mmol/L)*	*Patient ID*	*Value (mmol/L)*
001PANC0002	129	001PANC0029	146
		001PANC0011	131
		001PANC0047	133
		001PANC0050	132
		020PP20006	132
		165PP20006	131
		165PP20018	131

Table B.75: Overview of the outlier values determined using the IQR-method for the Sodium ($mmol/L$) measured before and after the first chemotherapy cycle.

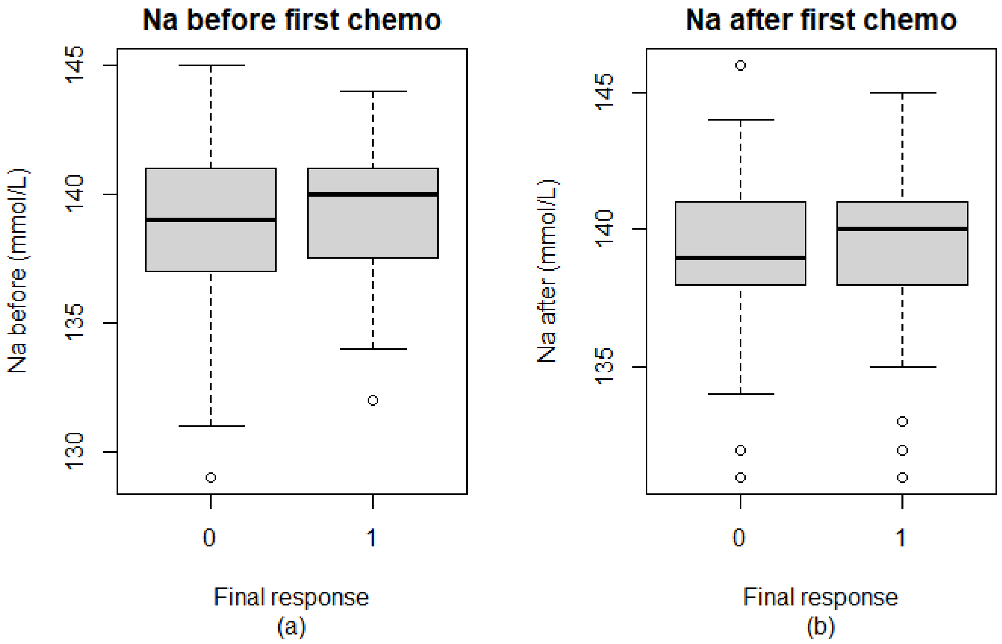


Figure B.227: Boxplot of the Sodium (Na) ($mmol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=115), 1 = Progressive Disease (n=32), total n=147)

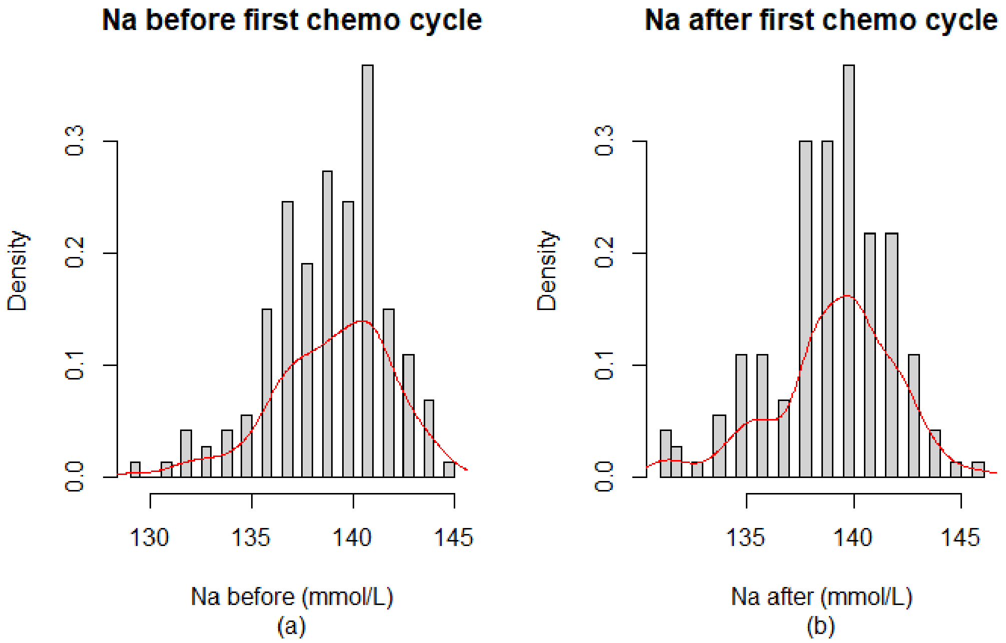


Figure B.228: Histogram of the Sodium (Na) $(mmol/L)$ values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=115), 1 = Progressive Disease (n=32), total n=147)

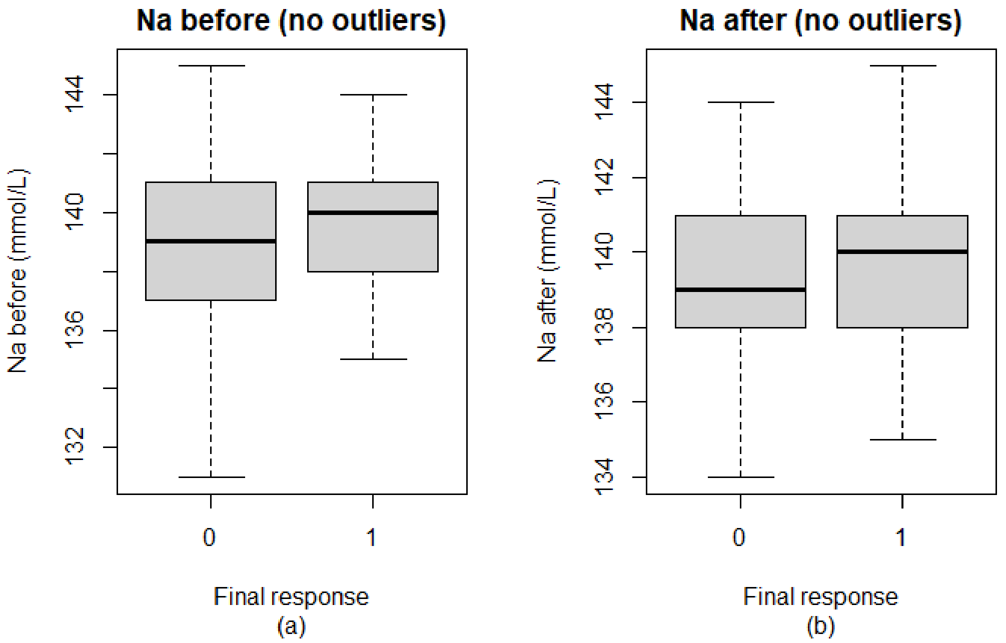


Figure B.229: Boxplot of the Sodium (Na) $(mmol/L)$ values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=110), 1 = Progressive Disease (n=29), total n=139)

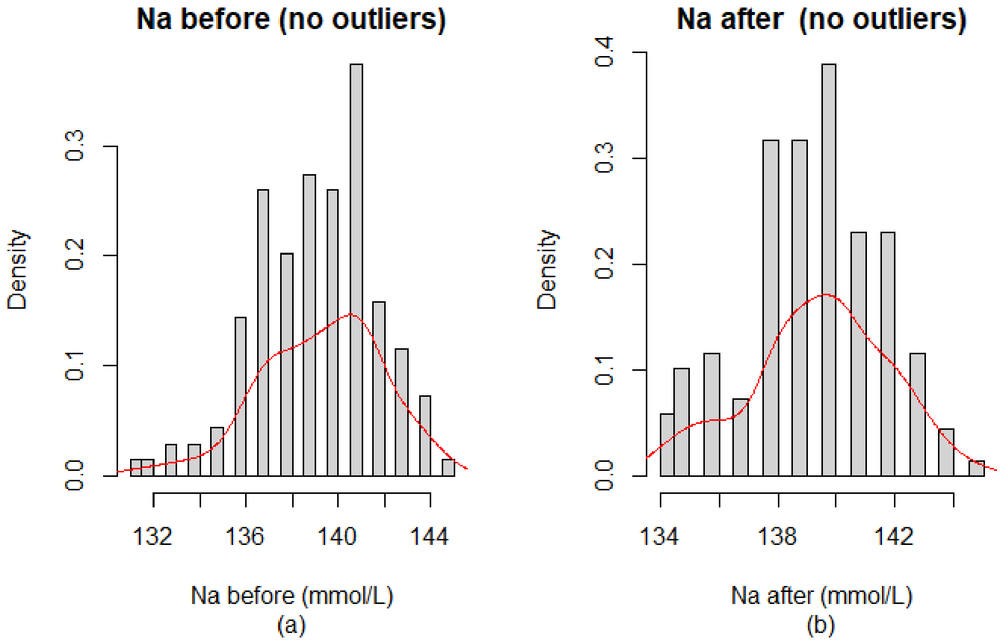


Figure B.230: Histogram of the Sodium (Na) ($mmol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=110), 1 = Progressive Disease (n=29), total n=139)

B.2.2.9 Potassium

K before		**K after**	
Patient ID	*Value (mmol/L)*	*Patient ID*	*Value (mmol/L)*
		001PANC0015	2.4

Table B.76: Overview of the outlier values determined using the IQR-method for the Potassium ($mmol/L$) measured before and after the first chemotherapy cycle.

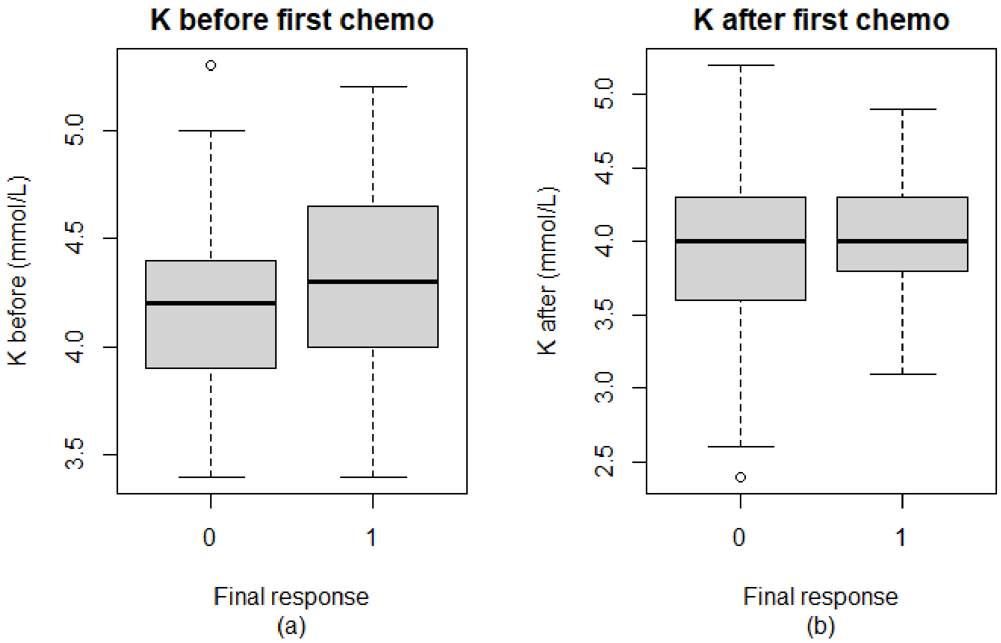


Figure B.231: Boxplot of the Potassium (K) ($mmol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=114), 1 = Progressive Disease (n=31), total n=145)

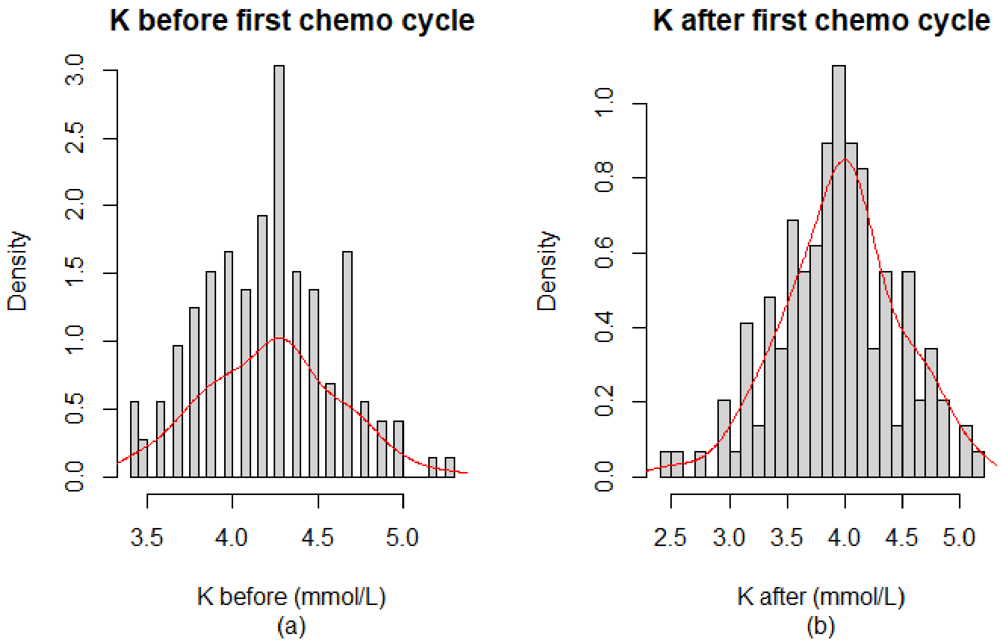


Figure B.232: Histogram of the Potassium (K) ($mmol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=114), 1 = Progressive Disease (n=31), total n=145)

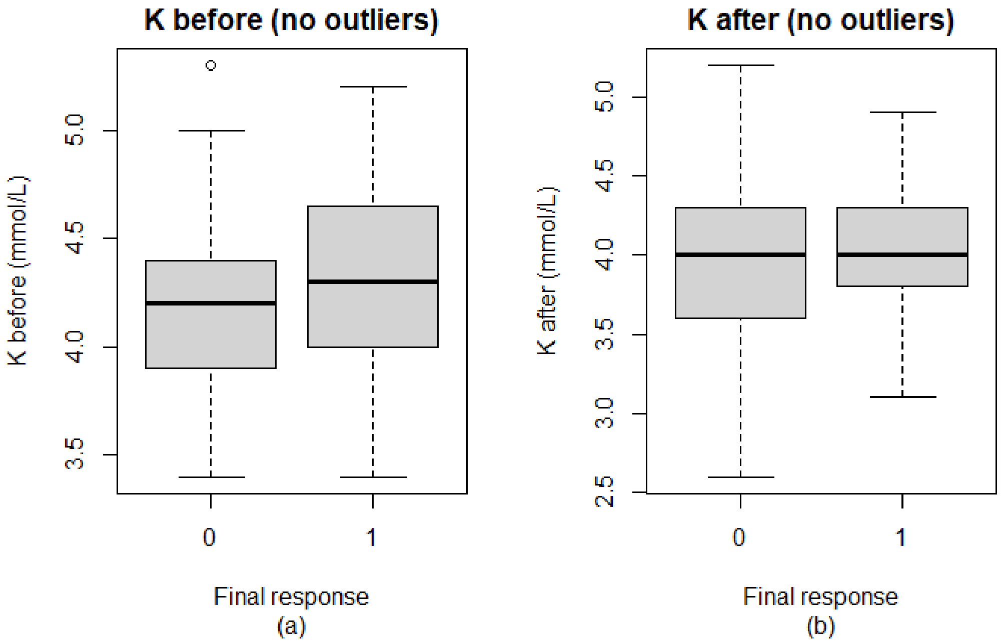


Figure B.233: Boxplot of the Potassium (K) $(mmol/L)$ values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=113), 1 = Progressive Disease (n=31), total n=144)

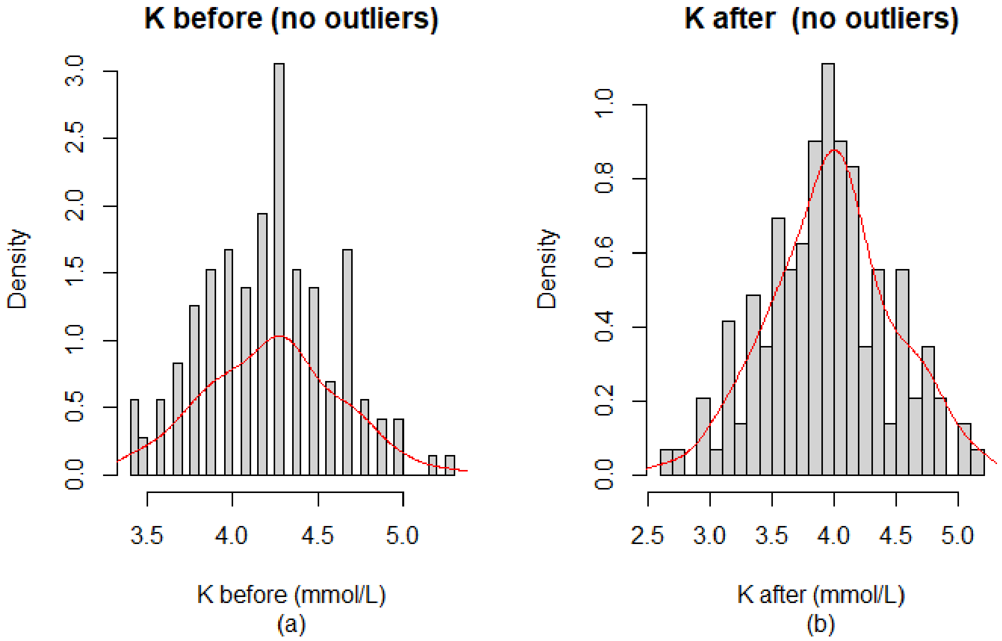


Figure B.234: Histogram of the Potassium (K) $(mmol/L)$ values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=113), 1 = Progressive Disease (n=31), total n=144)

B.2.2.10 ASAT

ASAT before		**ASAT after**	
Patient ID	*Value (U/L)*	*Patient ID*	*Value (U/L)*
001PANC0001	104	001PANC0029	68
001PANC0025	178	001PANC0032	222
001PANC0013	285	001PANC0004	196
151PP20009	87	001PANC0019	67
001PANC0052	1100	001PP20007	69
001PP20032	207	018PP20003	153
002PANC0013	135	148PP20006	58
002PP20051	231	001PP20031	69
078PP20016	77	002PANC0013	56
133PP20001	300	002PP20063	56
133PP20006	66	133PP20001	57
133PP20005	305	148PANC0008	75
148PANC0004	73	148PANC0003	54
148PP20013	69	148PP20019	61
148PP20027	77	005PP20016	69
151PP20017	106	078PP20040	69
005PP20016	79		
148PANC0009	150		
151PP20028	98		

Table B.77: Overview of the outlier values determined using the IQR-method for the ASAT (U/L) measured before and after the first chemotherapy cycle.

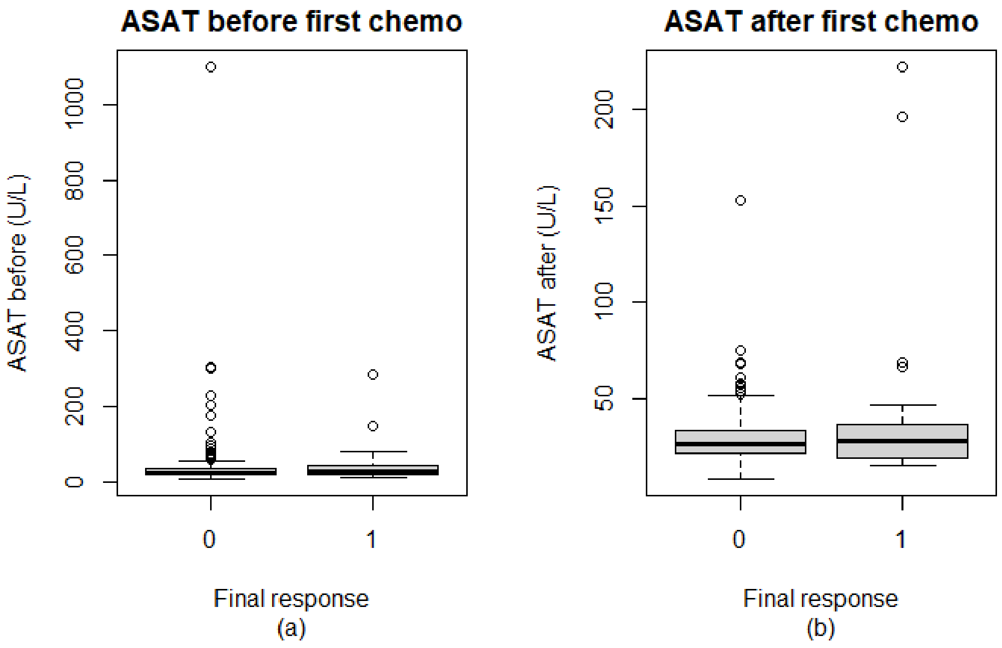


Figure B.235: Boxplot of the ASAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=146), 1 = Progressive Disease (n=34), total n=180)

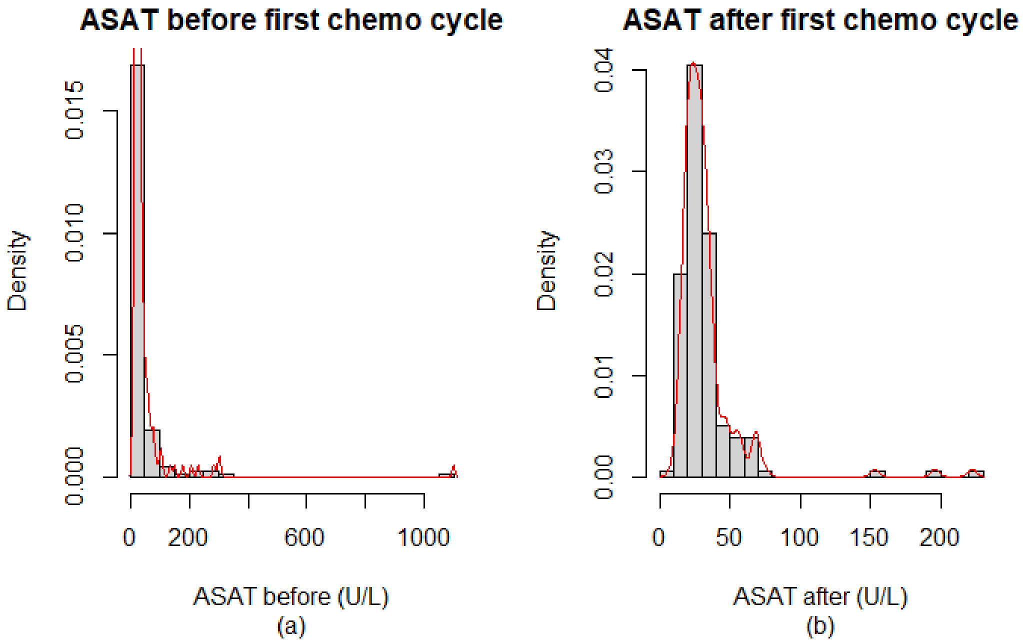


Figure B.236: Histogram of the ASAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=146), 1 = Progressive Disease (n=34), total n=180)

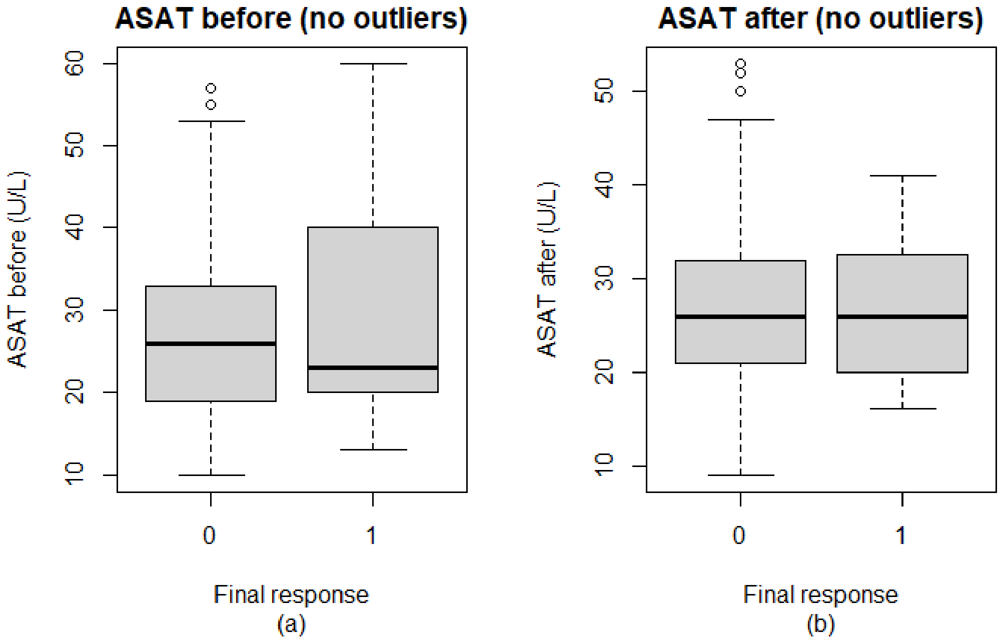


Figure B.237: Boxplot of the ASAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=121), 1 = Progressive Disease (n=27), total n=148)

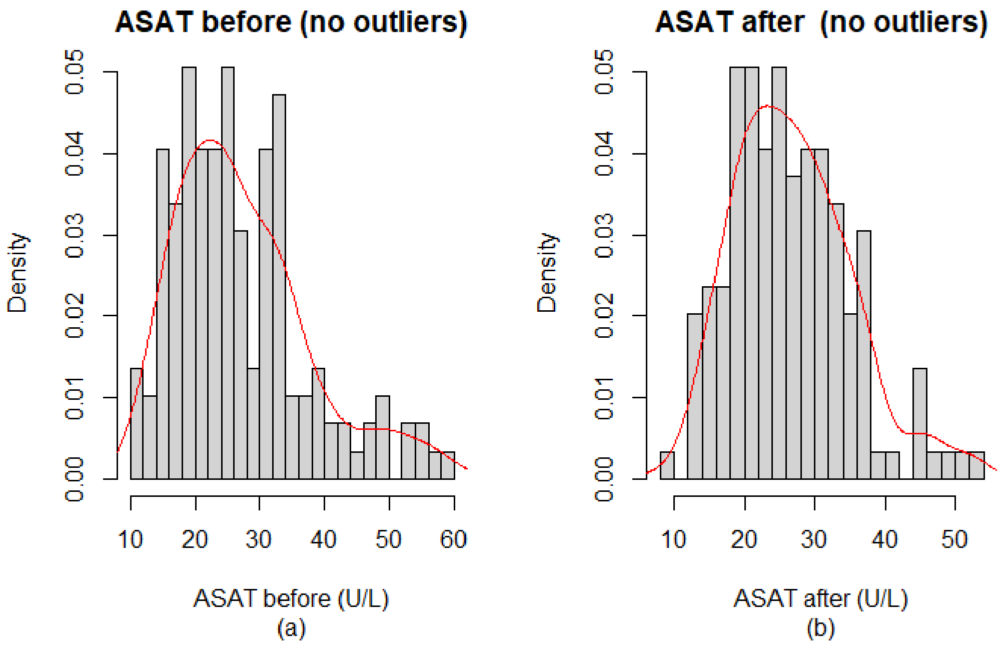


Figure B.238: Histogram of the ASAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=121), 1 = Progressive Disease (n=27), total n=148)

B.2.2.11 ALAT

ALAT before		**ALAT after**	
Patient ID	*Value (U/L)*	*Patient ID*	*Value (U/L)*
001PANC0001	128	001PANC0029	155
001PANC0025	400	001PANC0032	264
001PANC0013	636	001PANC0004	175
001PP20002	125	001PP20007	96
002PANC0005	122	148PP20002	176
002PP20024	143	001PP20031	154
002PP20005	114	002PANC0013	151
148PP20009	126	002PP20055	116
151PP20009	153	002PP20063	90
001PANC0052	1241	078PP20029	114
001PP20032	200	133PP20007	183
002PANC0013	279	133PP20005	124
002PP20047	107	148PANC0004	150
002PP20063	105	148PP20019	106
078PP20016	143	165PP20006	103
133PP20001	133	078PP20040	90
133PP20007	172		
133PP20005	389		
148PP20013	122		
148PP20027	125		
151PP20017	194		
148PANC0009	347		
151PP20028	178		

Table B.78: Overview of the outlier values determined using the IQR-method for the ALAT (U/L) measured before and after the first chemotherapy cycle.

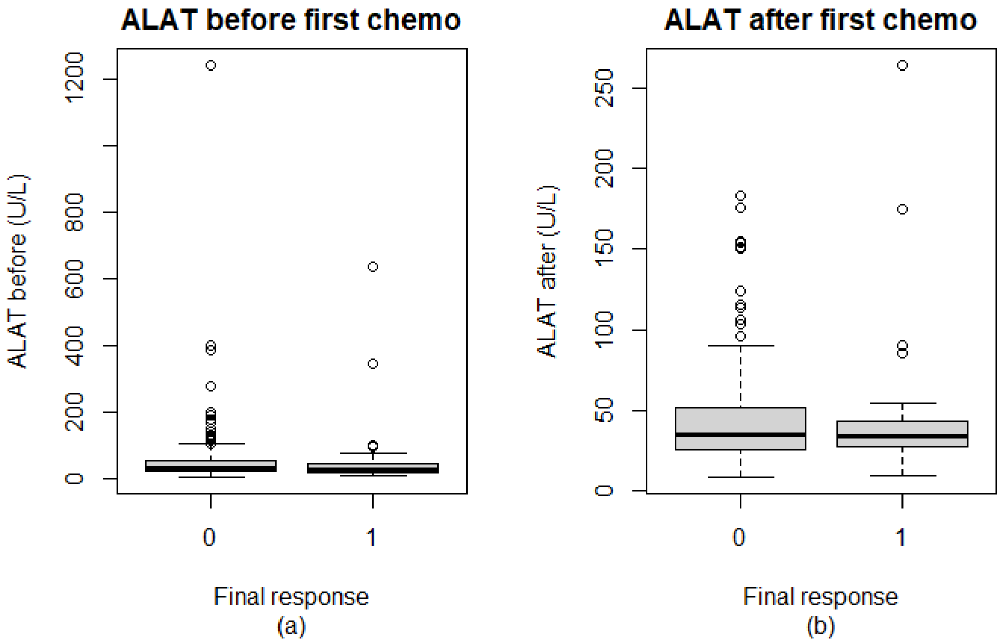


Figure B.239: Boxplot of the ALAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=164), 1 = Progressive Disease (n=39), total n=203)

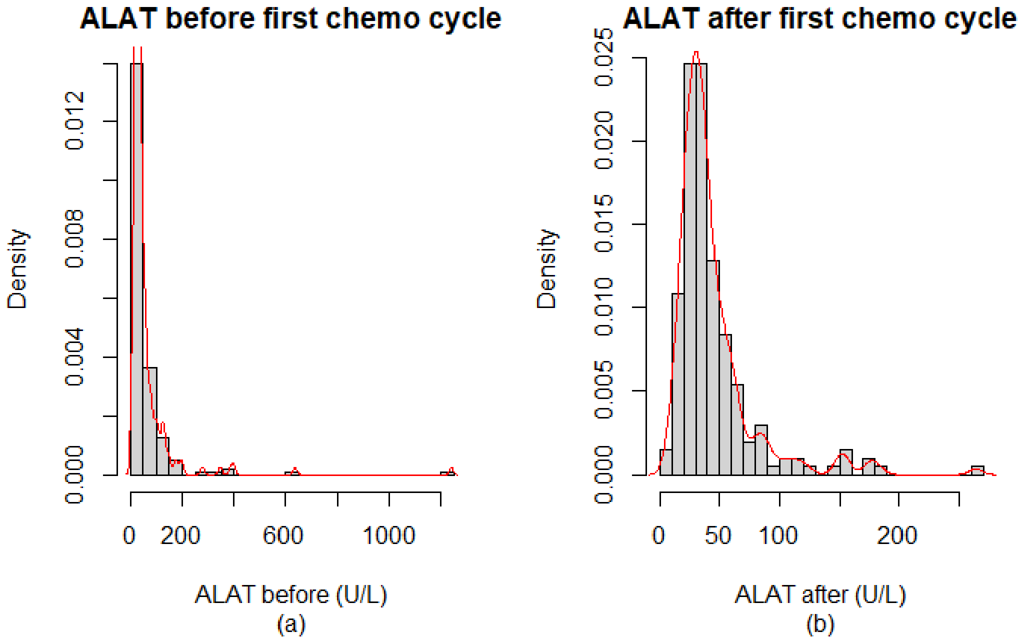


Figure B.240: Histogram of the ALAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=164), 1 = Progressive Disease (n=39), total n=203)

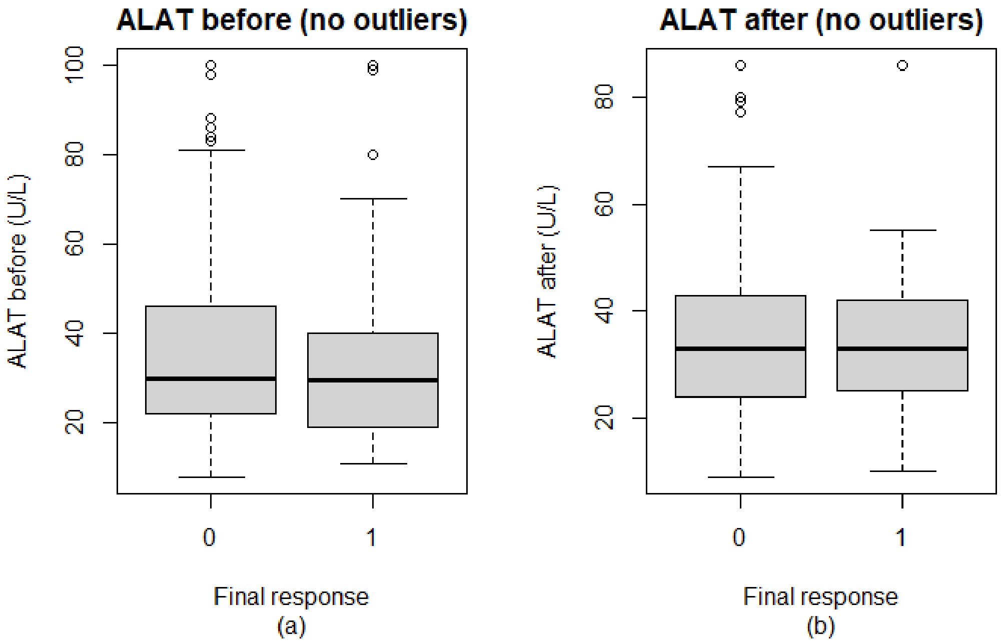


Figure B.241: Boxplot of the ALAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=134), 1 = Progressive Disease (n=34), total n=168)

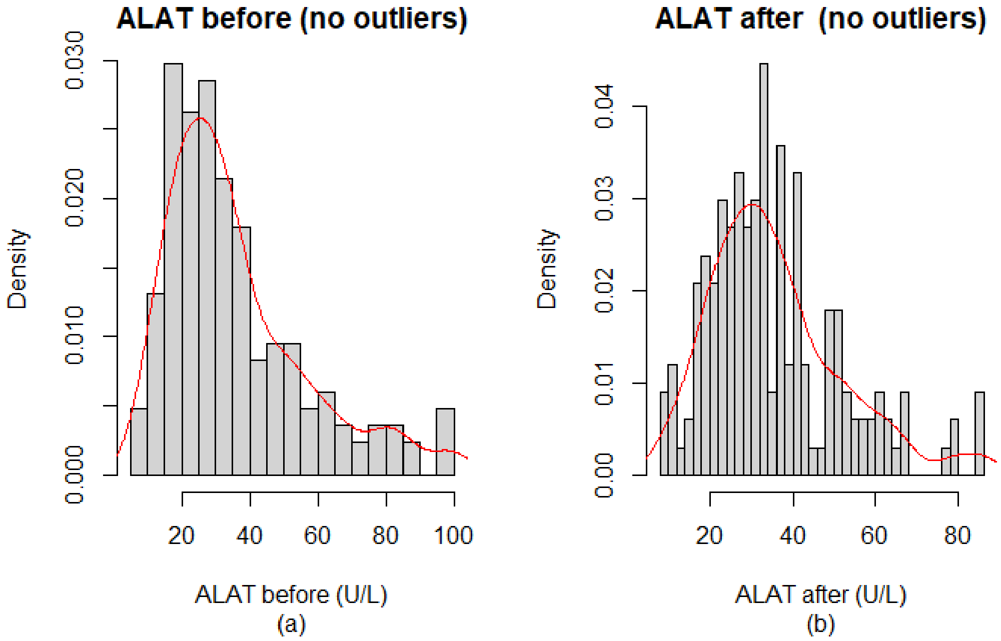


Figure B.242: Histogram of the ALAT (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=134), 1 = Progressive Disease (n=34), total n=168)

B.2.2.12 Alkaline Phosphatase

AF before		**AF after**	
Patient ID	*Value (U/L)*	*Patient ID*	*Value (U/L)*
001PANC0028	596	001PANC0028	454
001PANC0016	402	001PANC0014	345
001PANC0018	385	001PANC0016	456
001PANC0033	892	001PANC0032	336
001PANC0011	562	001PANC0004	588
148PP20006	706	001PANC0033	735
151PP20004	431	001PANC0011	493
002PP20047	367	148PP20006	533
078PP20016	509	078PP20016	334
133PP20007	717	133PP20007	418
133PP20005	750	133PP20005	515
148PP20013	397	078PP20040	787
078PP20040	36		
148PANC0009	564		

Table B.79: Overview of the outlier values determined using the IQR-method for the Alkaline Phosphatase (U/L) measured before and after the first chemotherapy cycle.

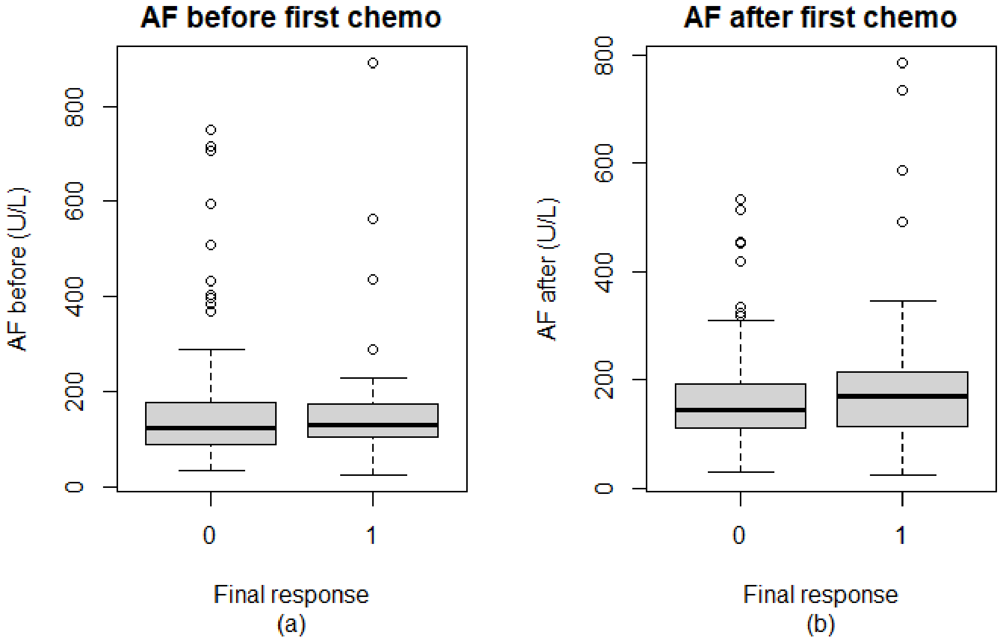


Figure B.243: Boxplot of the Alkaline Phosphatase (AF) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=161), 1 = Progressive Disease (n=39), total n=200)

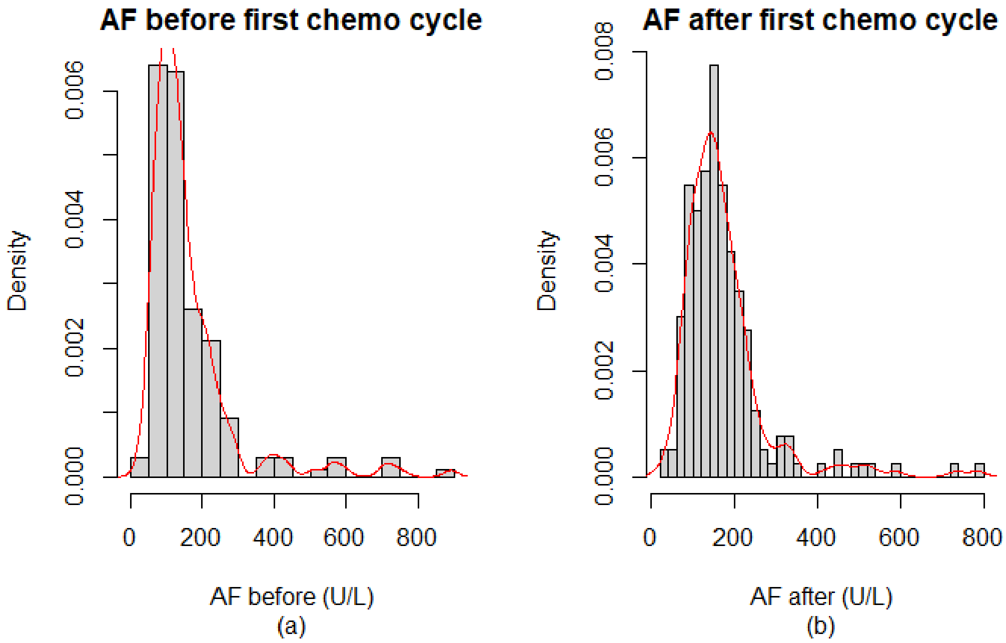


Figure B.244: Histogram of the Alkaline Phosphatase (AF) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=161), 1 = Progressive Disease (n=39), total n=200)

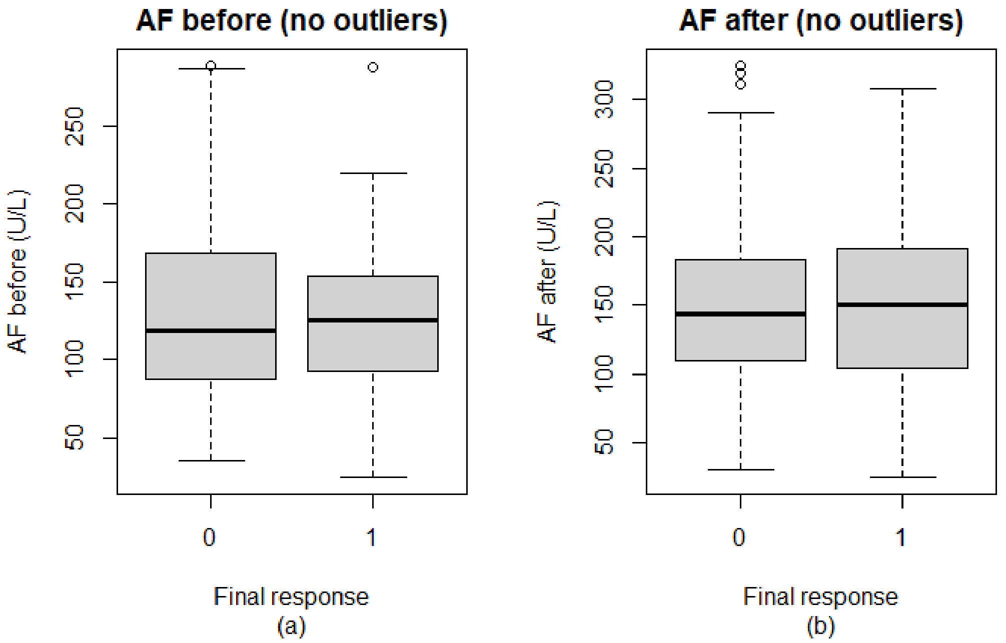


Figure B.245: Boxplot of the Alkaline Phosphatase (AF) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=151), 1 = Progressive Disease (n=32), total n=183)

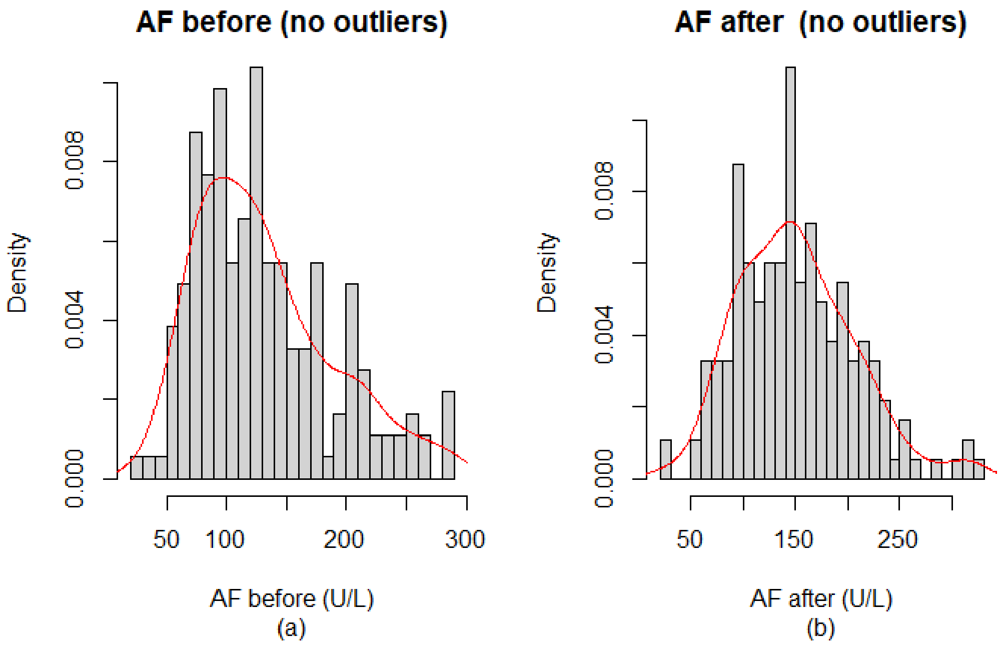


Figure B.246: Histogram of the Alkaline Phosphatase (AF) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=151), 1 = Progressive Disease (n=32), total n=183)

B.2.2.13 γ-Glutamyl Transferase

GGT before		**GGT after**	
Patient ID	*Value (U/L)*	*Patient ID*	*Value (U/L)*
001PANC0001	509	001PANC0001	395
001PANC0028	790	001PANC0028	777
001PANC0025	924	001PANC0009	402
001PANC0010	405	001PANC0032	634
001PANC0018	551	001PANC0004	488
001PANC0013	515	001PANC0010	348
001PANC0033	895	001PANC0018	377
148PP20006	792	001PANC0033	718
151PP20004	1110	001PP20013	306
001PANC0052	693	148PP20006	552
001PP20032	719	001PANC0039	379
002PP20047	494	001PP20031	314
005PP20004	405	005PP20004	432
020PP20006	443	133PP20005	827
078PP20016	417	148PP20019	606
133PP20005	1576	001PANC0061	451
001PANC0061	398	078PP20040	942
078PP20040	676		

Table B.80: Overview of the outlier values determined using the IQR-method for the γ-Glutamyl Transferase (U/L) measured before and after the first chemotherapy cycle.

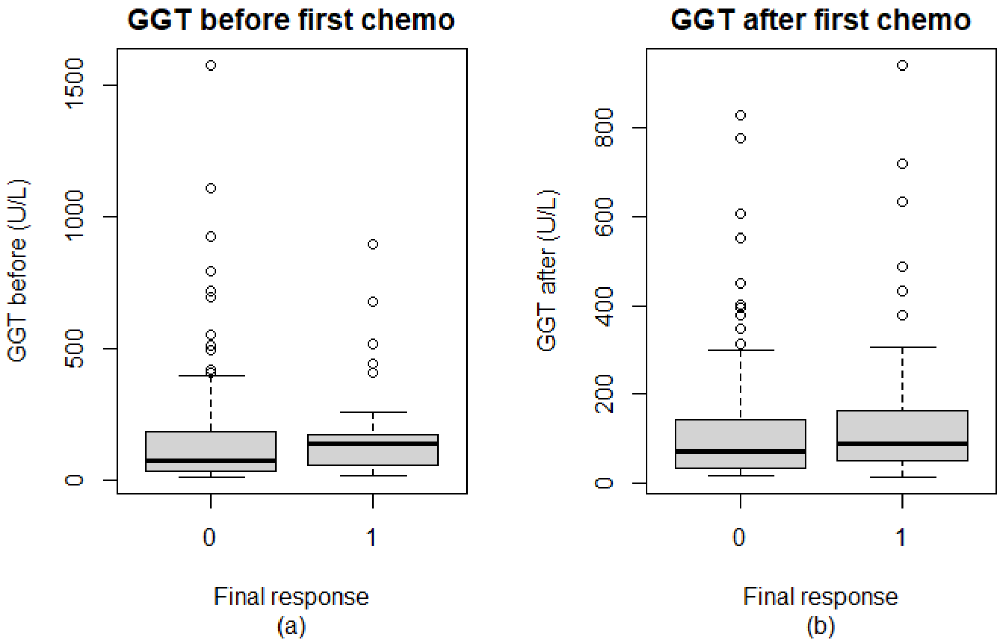


Figure B.247: Boxplot of the γ-Glutamyl Transferase (GGT) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=146), 1 = Progressive Disease (n=34), total n=180)

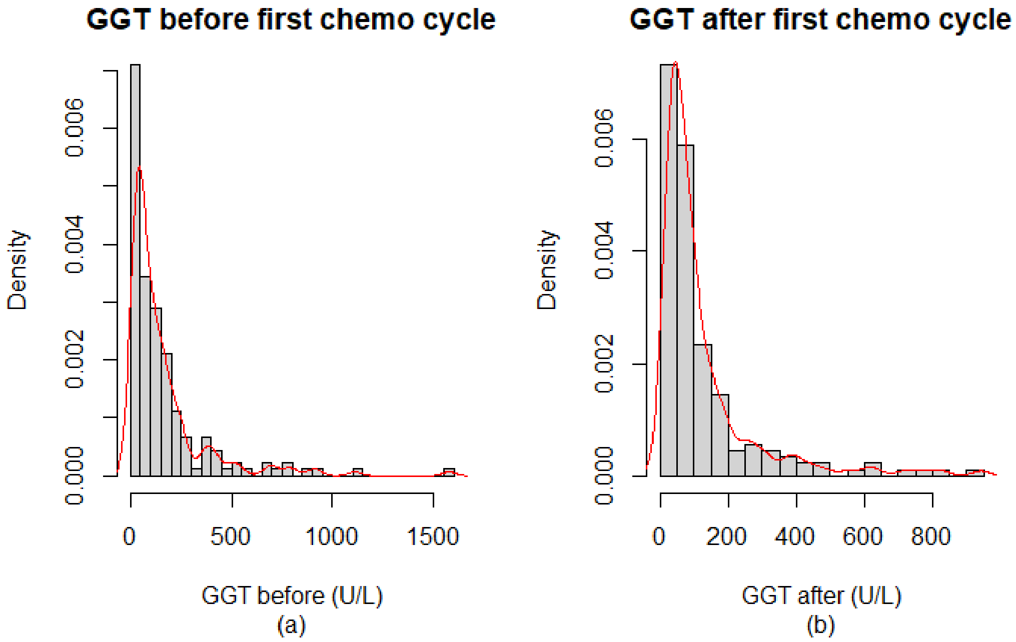


Figure B.248: Histogram of the γ-Glutamyl Transferase (GGT) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=146), 1 = Progressive Disease (n=34), total n=180)

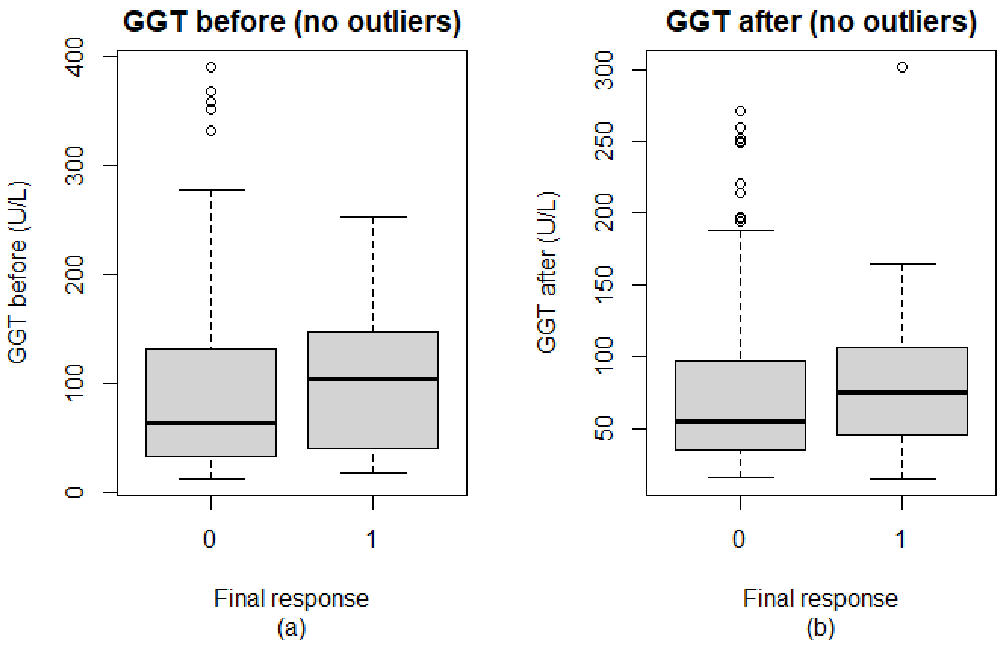


Figure B.249: Boxplot of the γ-Glutamyl Transferase (GGT) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=130), 1 = Progressive Disease (n=25), total n=155)

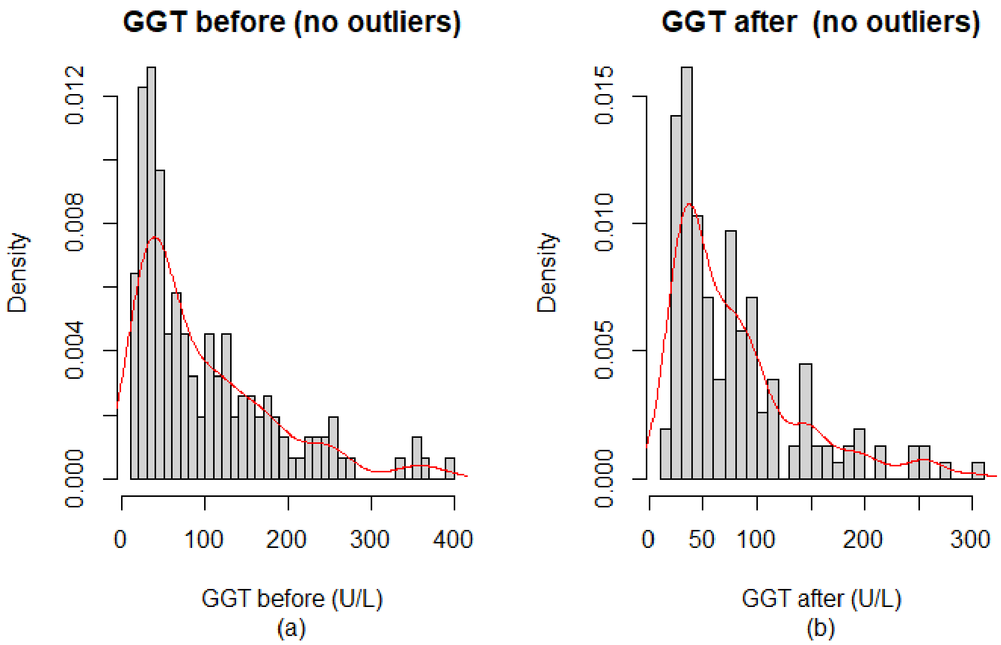


Figure B.250: Histogram of the γ-Glutamyl Transferase (GGT) (U/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red.(0 = Disease control (n=130), 1 = Progressive Disease (n=25), total n=155)

B.2.2.14 Bilirubin

BR before		**BR after**	
Patient ID	*Value ($\mu mol/L$)*	*Patient ID*	*Value ($\mu mol/L$)*
001PANC0025	132	001PANC0011	24
001PANC0013	83	002PP20002	23
133PP20001	55	078PP20026	27
148PP20013	84	165PP20012	30
151PP20028	60	133PP20009	27

Table B.81: Overview of the outlier values determined using the IQR-method for the Bilirubin values ($\mu mol/L$) measured before and after the first chemotherapy cycle.

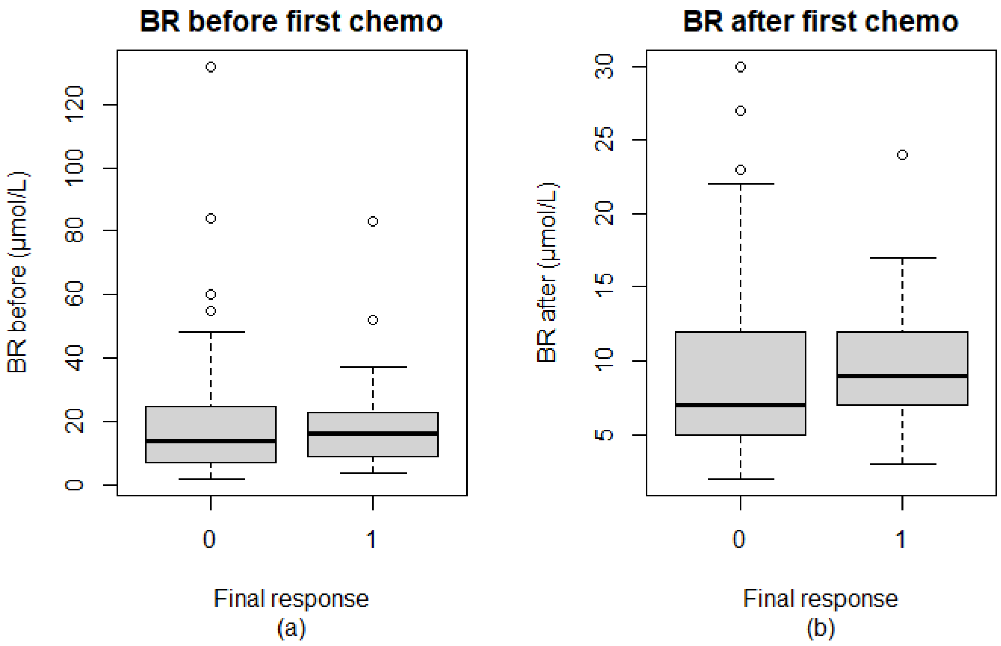


Figure B.251: Boxplot of the Bilirubin (BR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=163), 1 = Progressive Disease (n=41), total n=204)

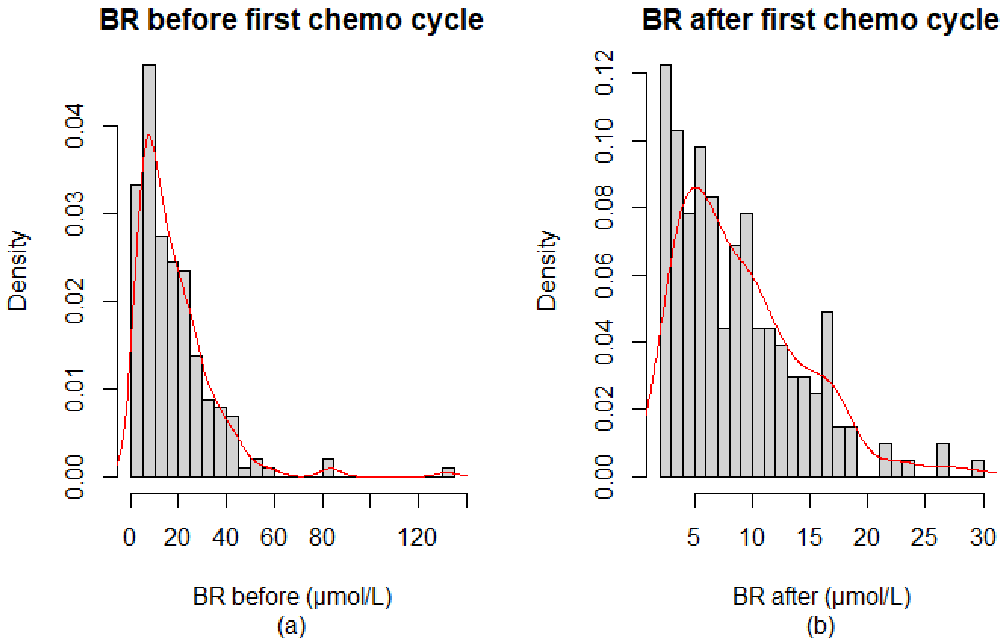


Figure B.252: Histogram of the Bilirubin (BR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=163), 1 = Progressive Disease (n=41), total n=204)

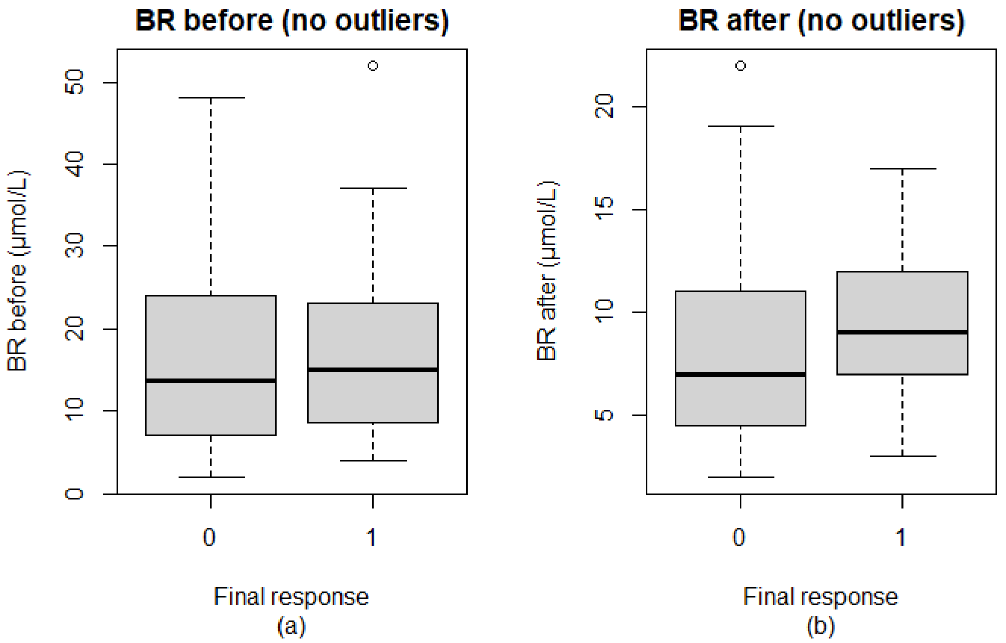


Figure B.253: Boxplot of the Bilirubin (BR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=155), 1 = Progressive Disease (n=39), total n=194)

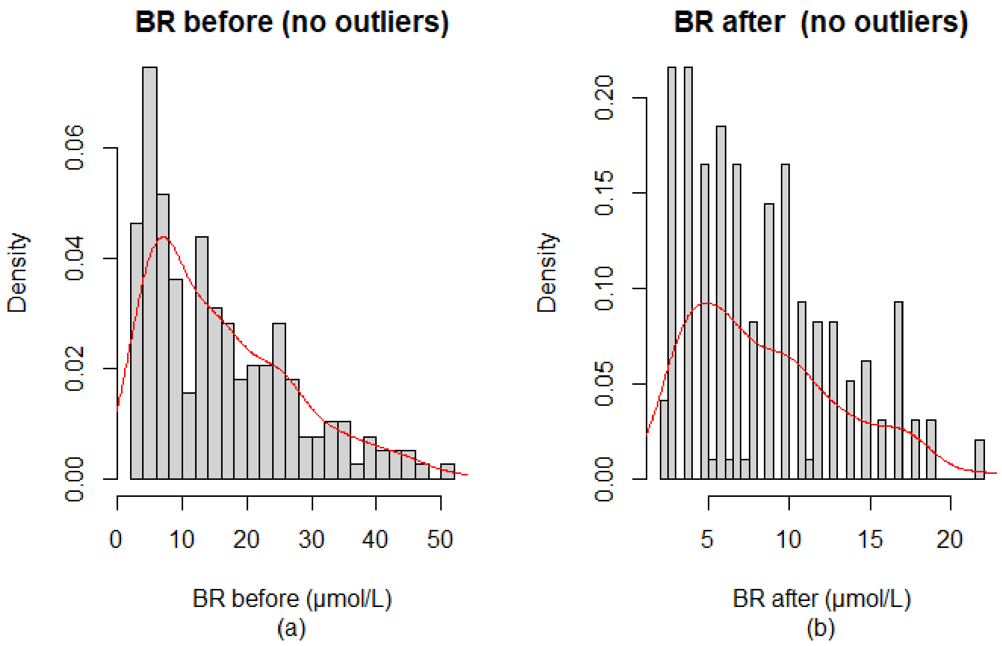


Figure B.254: Histogram of the Bilirubin (BR) ($\mu mol/L$) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=155), 1 = Progressive Disease (n=39), total n=194)

B.2.2.15 Albumin

Alb before		**Alb after**	
Patient ID	*Value (g/L)*	*Patient ID*	*Value (g/L)*
001PANC0043	23	001PANC0039	22
166PP20006	25	002PP20037	64
		165PP20012	8

Table B.82: Overview of the outlier values determined using the IQR-method for the Albumin values (g/L) measured before and after the first chemotherapy cycle.

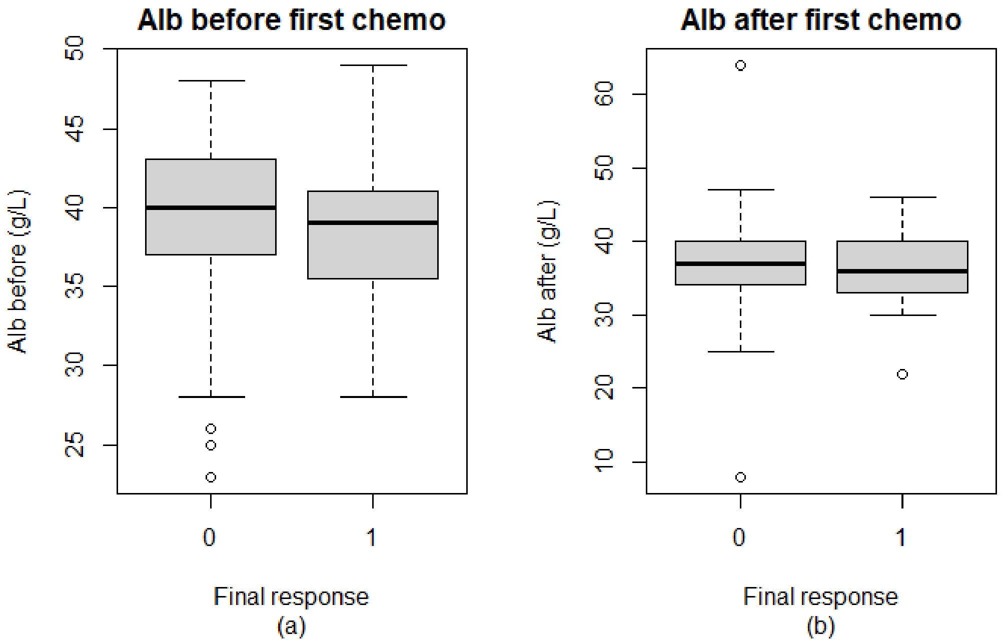


Figure B.255: Boxplot of the Albumin (Alb) (g/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=122), 1 = Progressive Disease (n=28), total n=150)

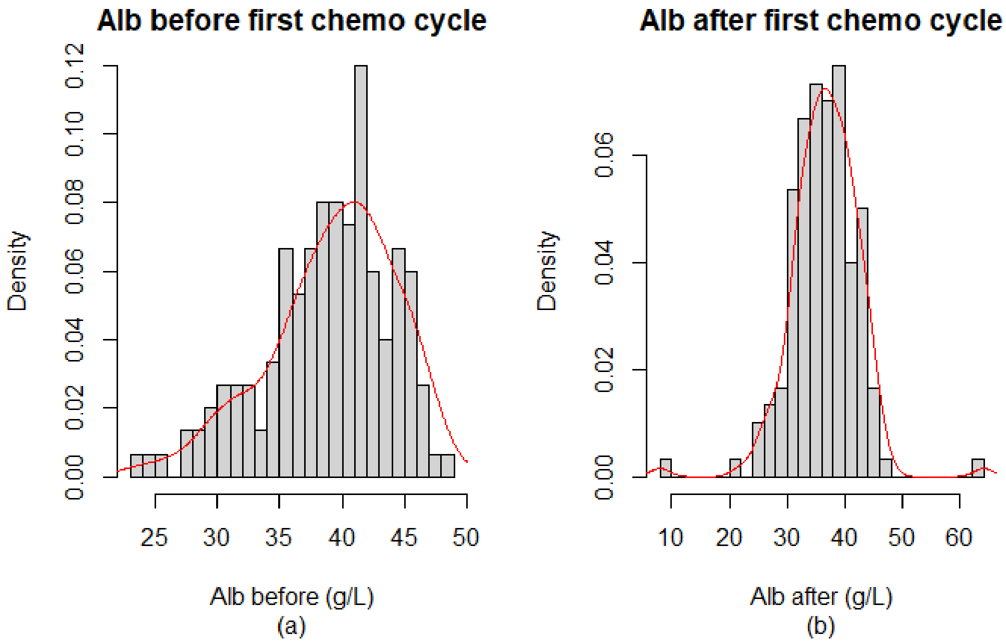


Figure B.256: Histogram of the Albumin (Alb) (g/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red.(0 = Disease control (n=122), 1 = Progressive Disease (n=28), total n=150)

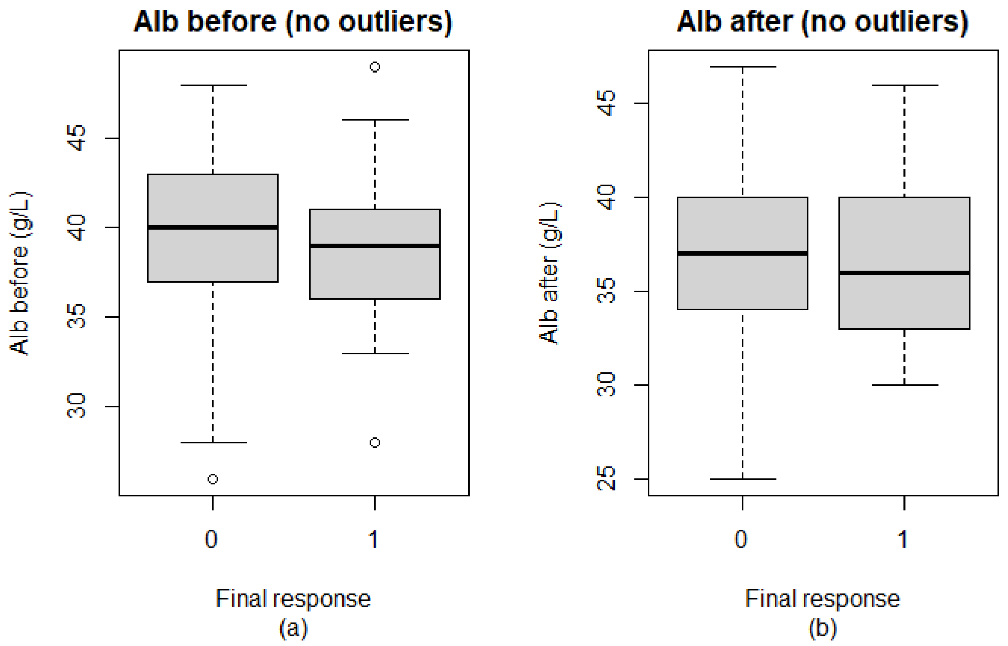


Figure B.257: Boxplot of the Albumin (Alb) (g/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=118), 1 = Progressive Disease (n=27), total n=145)

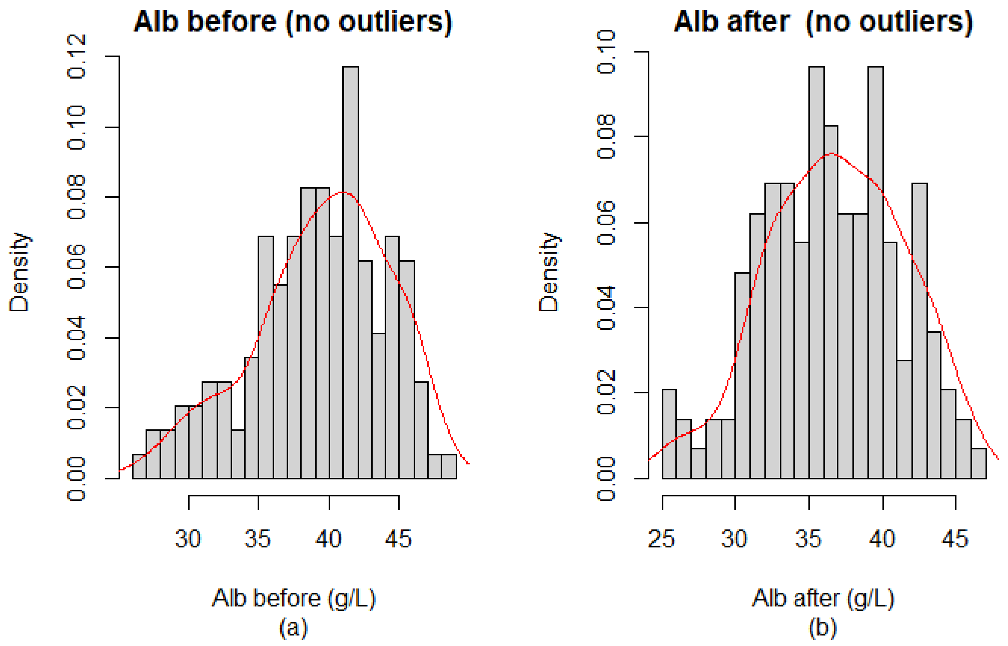


Figure B.258: Histogram of the Albumin (Alb) (g/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=118), 1 = Progressive Disease (n=27), total n=145)

B.2.2.16 C-Reactive Protein

CRP before		**CRP after**	
Patient ID	*Value (mg/L)*	*Patient ID*	*Value (mg/L)*
001PANC0028	44.0	001PANC0036	96.0
001PANC0002	56.0	001PANC0016	81.0
001PANC0036	91.0	001PANC0011	87.0
001PANC0016	91.0	065PP20005	77.0
001PANC0004	38.0	078PANC0002	86.9
001PANC0033	47.0	001PANC0050	70.0
001PANC0011	27.0	001PP20032	70.0
001PP20013	38.0	078PP20026	270.3
001PP20015	26.0	148PP20013	78.0
001PP20012	57.0	148PP20015	90.0
151PP20004	150.0	001PANC0061	104.0
001PANC0039	88.0	001PP20048	75.0
001PP20032	49.0		
148PP20015	27.0		
001PANC0061	73.0		
001PP20048	66.0		
005PP20020	71.0		

Table B.83: Overview of the outlier values determined using the IQR-method for the C-Reactive Protein values (mg/L) measured before and after the first chemotherapy cycle.

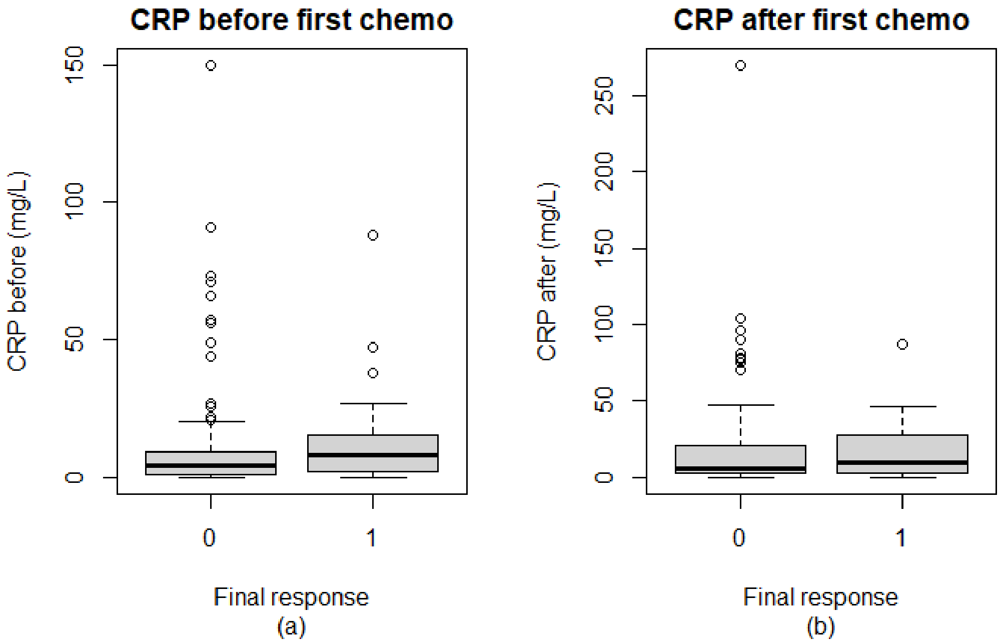


Figure B.259: Boxplot of the C-Reactive Protein (CRP) (mg/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=96), 1 = Progressive Disease (n=24), total n=120)

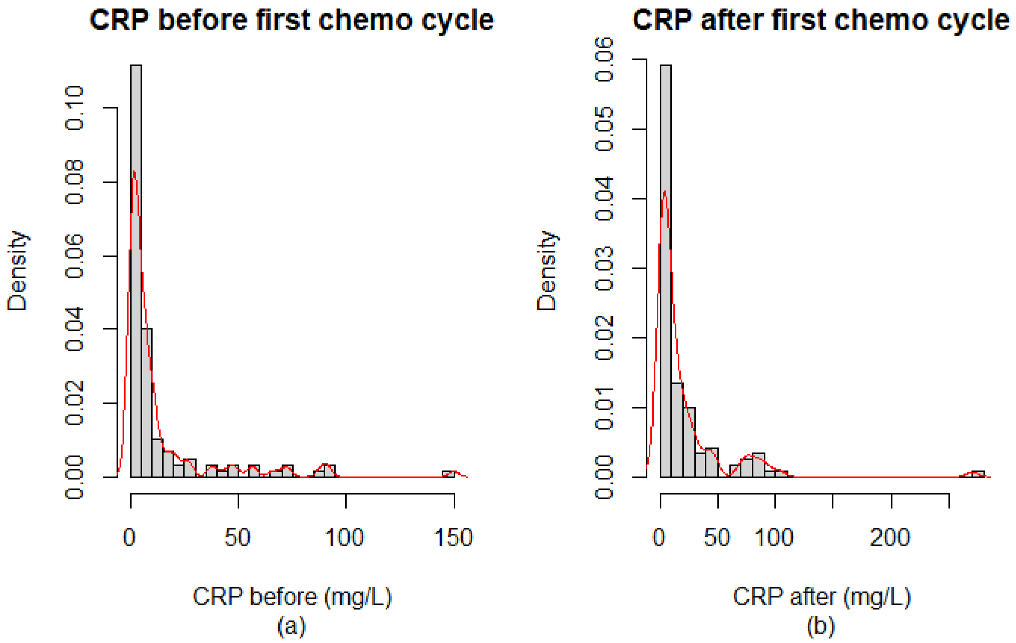


Figure B.260: Histogram of the C-Reactive Protein (CRP) (mg/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=96), 1 = Progressive Disease (n=24), total n=120)

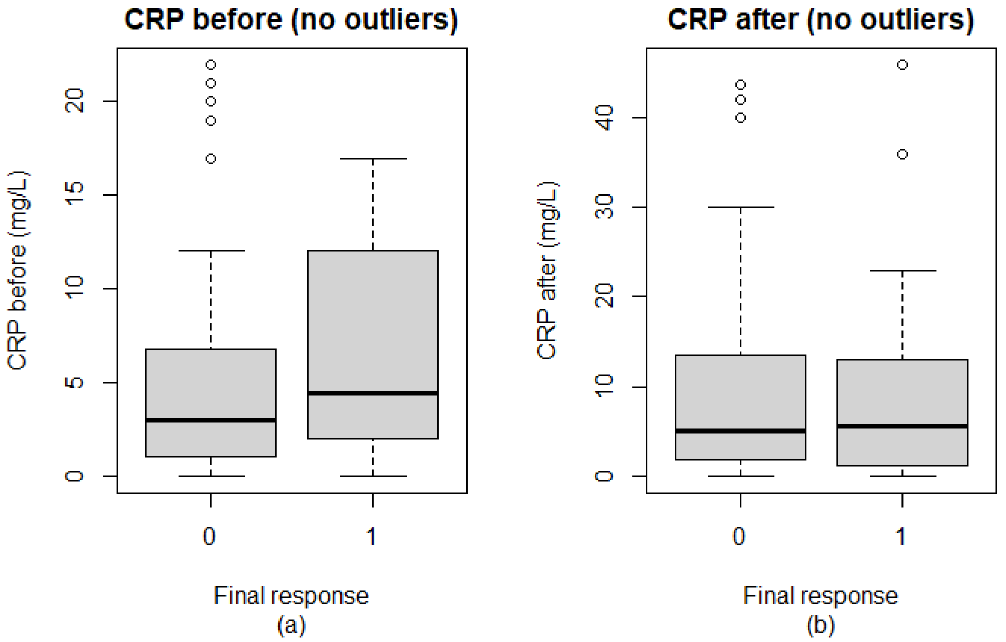


Figure B.261: Boxplot of the C-Reactive Protein (CRP) (mg/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=80), 1 = Progressive Disease (n=18), total n=98)

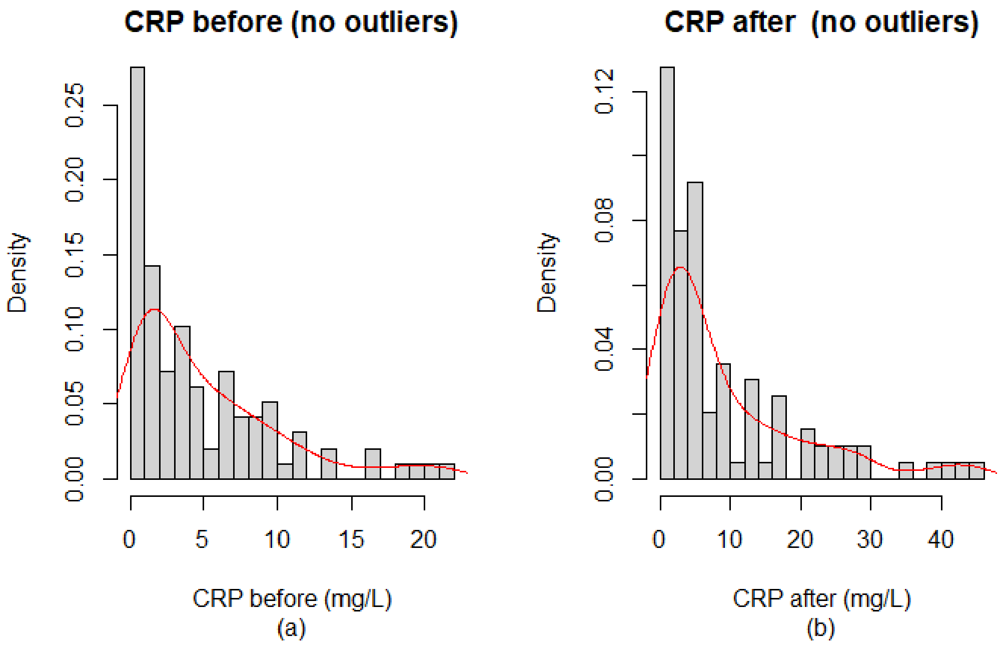


Figure B.262: Histogram of the C-Reactive Protein (CRP) (mg/L) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=80), 1 = Progressive Disease (n=18), total n=98)

B.2.2.17 Systemic Inflammation Index

SII before		**SII after**	
Patient ID	*Value*	*Patient ID*	*Value*
001PANC0036	7246.800	001PANC0015	5563.158
002PP20011	2112.892	001PANC0023	3188.976
078PANC0002	2469.190	001PANC0004	2706.022
001PANC0043	2388.971	001PANC0008	2851.294
002PP20046	2278.044	001PP20009	2865.977
001PANC0061	2879.677	078PANC0002	3067.107
		151PP20004	5220.000
		151PP20013	4091.769

Table B.84: Overview of the outlier values determined using the IQR-method for the Systemic Inflammation Index measured before and after the first chemotherapy cycle.

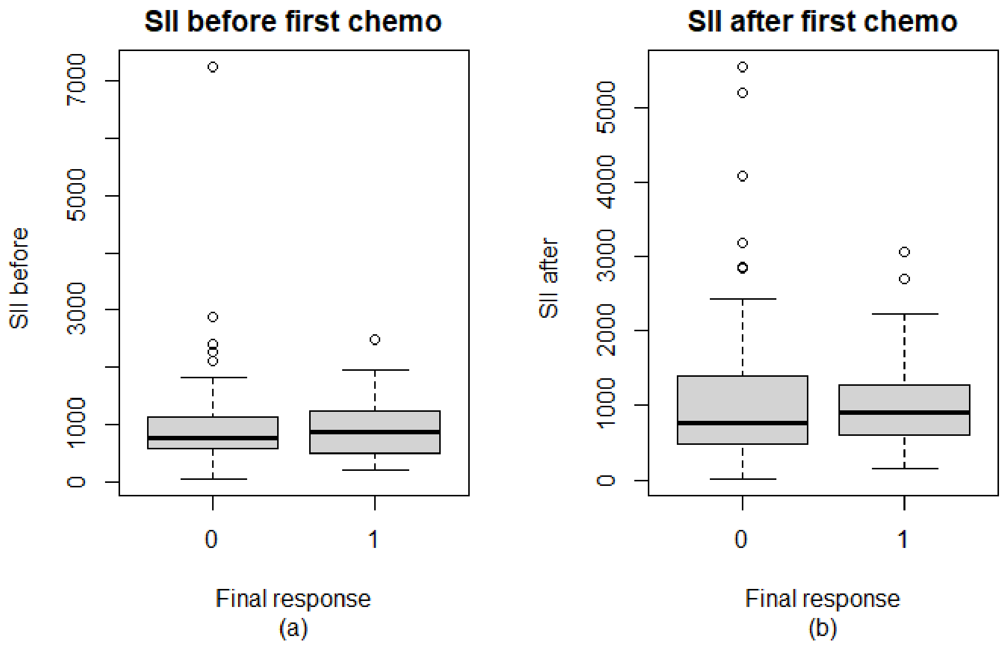


Figure B.263: Boxplot of the System Inflammation Index (SII) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=86), 1 = Progressive Disease (n=24), total n=110)

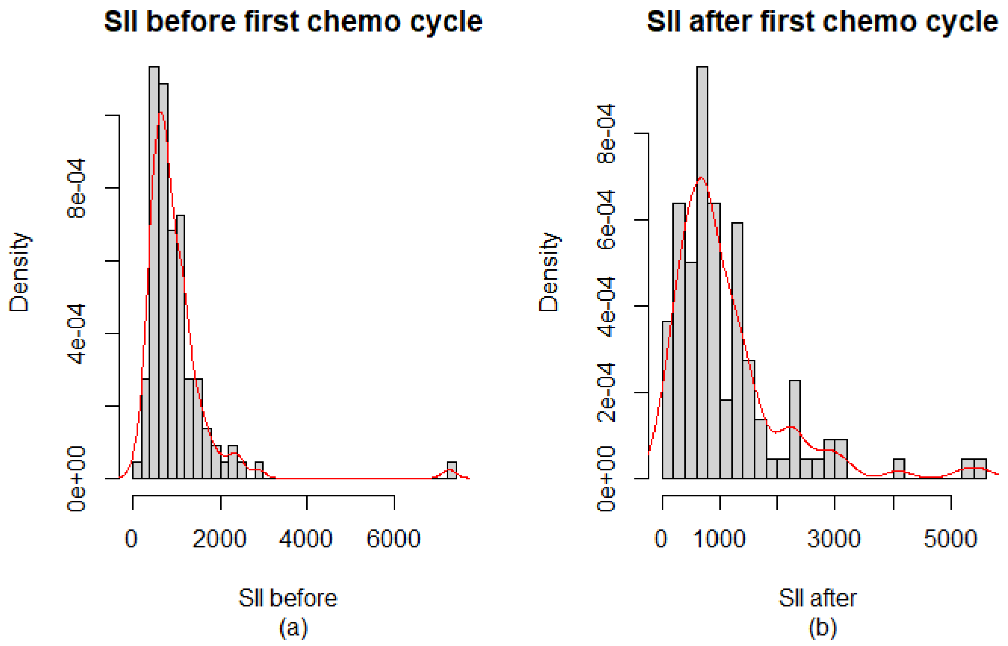


Figure B.264: Histogram of the System Inflammation Index (SII) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=86), 1 = Progressive Disease (n=24), total n=110)

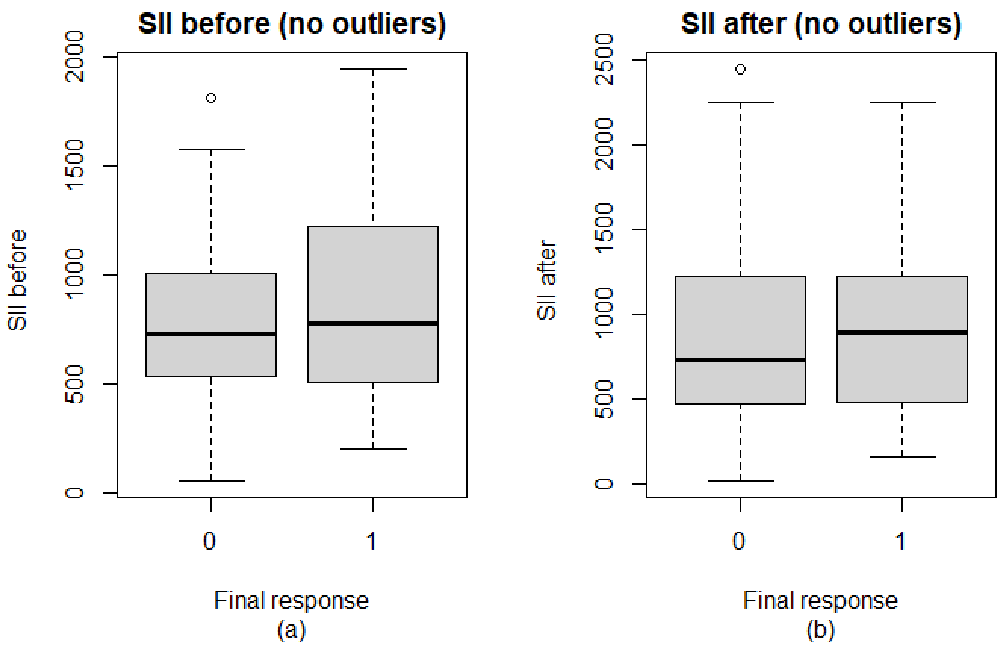


Figure B.265: Boxplot of the System Inflammation Index (SII) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=75), 1 = Progressive Disease (n=22), total n=97)

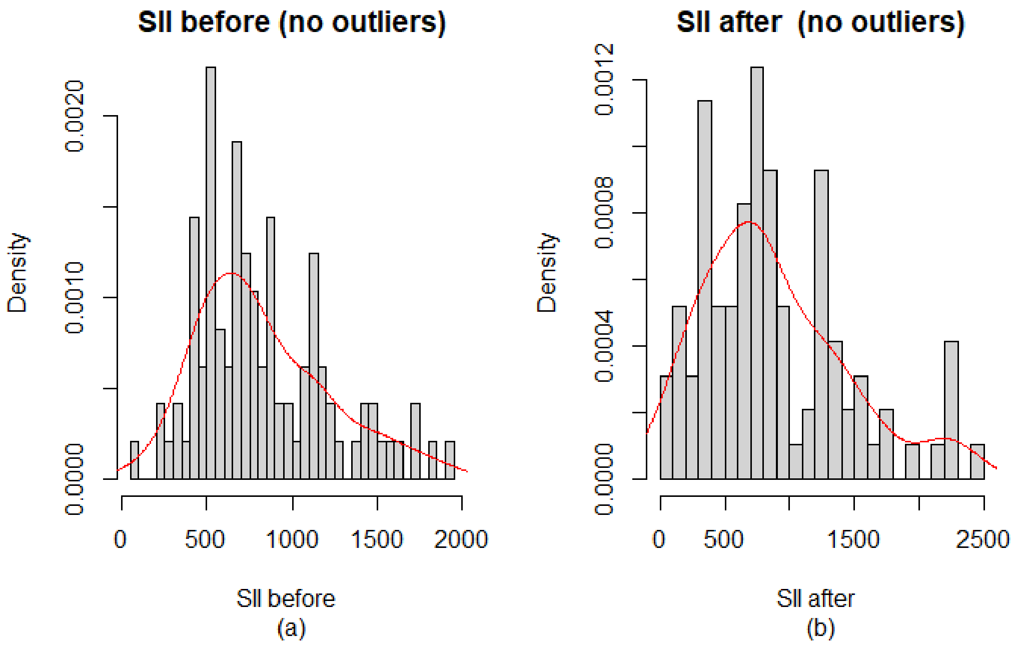


Figure B.266: Histogram of the System Inflammation Index (SII) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=75), 1 = Progressive Disease (n=22), total n=97)

B.2.2.18 Neutrophil-to-Lymphocyte Ratio

NLR before		**NLR after**	
Patient ID	*Value*	*Patient ID*	*Value*
001PANC0036	19.80	001PANC0015	36.84
001PANC0043	8.38	001PANC0035	19.49
001PP20032	8.84	001PANC0023	19.69
148PP20022	7.82	001PANC0024	13.79
001PANC0061	9.11	001PANC0017	12.22
		001PANC0026	13.04
		001PANC0011	15.79
		001PP20007	13.01
		151PP20004	36.00

Table B.85: Overview of the outlier values determined using the IQR-method for the Neutrophil-to-Lymphocyte Ratio measured before and after the first chemotherapy cycle.

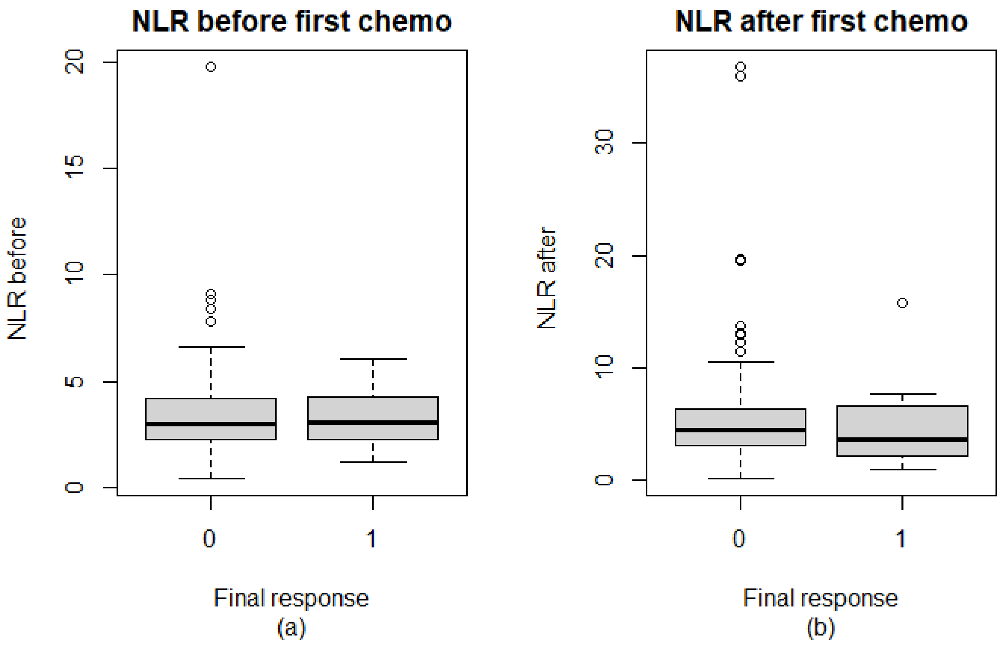


Figure B.267: Boxplot of the Neutrophil-to-Lymphocyte Ratio (NLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=87), 1 = Progressive Disease (n=24), total n=111)

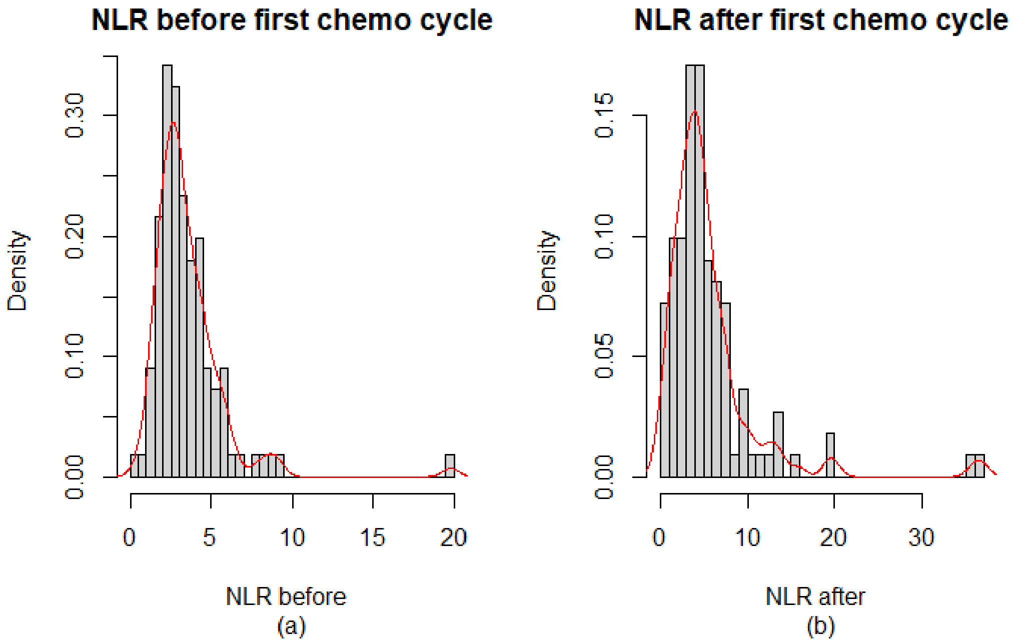


Figure B.268: Histogram of the Neutrophil-to-Lymphocyte Ratio (NLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=87), 1 = Progressive Disease (n=24), total n=111)

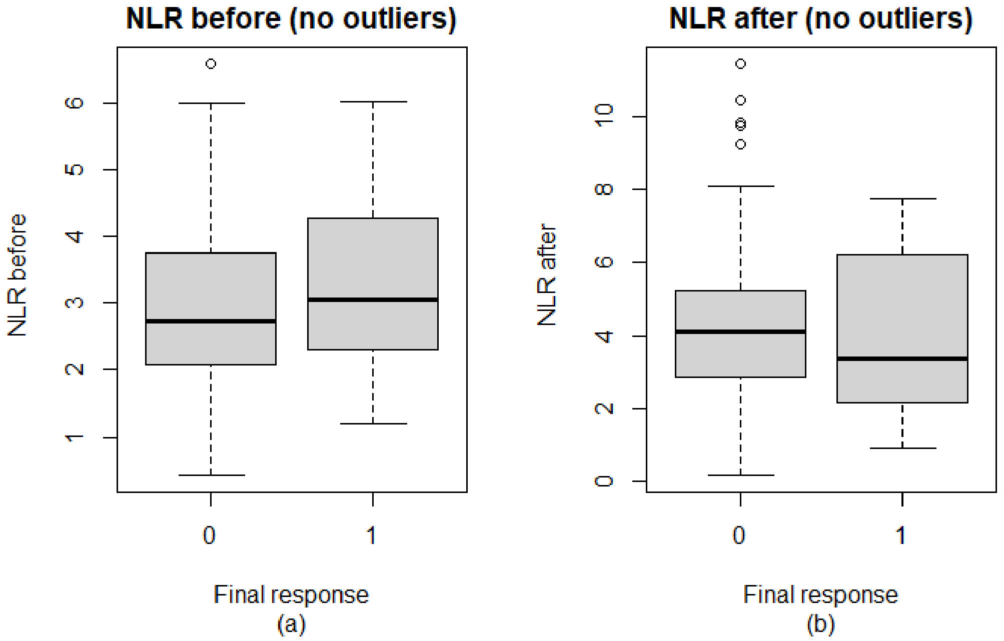


Figure B.269: Boxplot of the Neutrophil-to-Lymphocyte Ratio (NLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=74), 1 = Progressive Disease (n=23), total n=97)

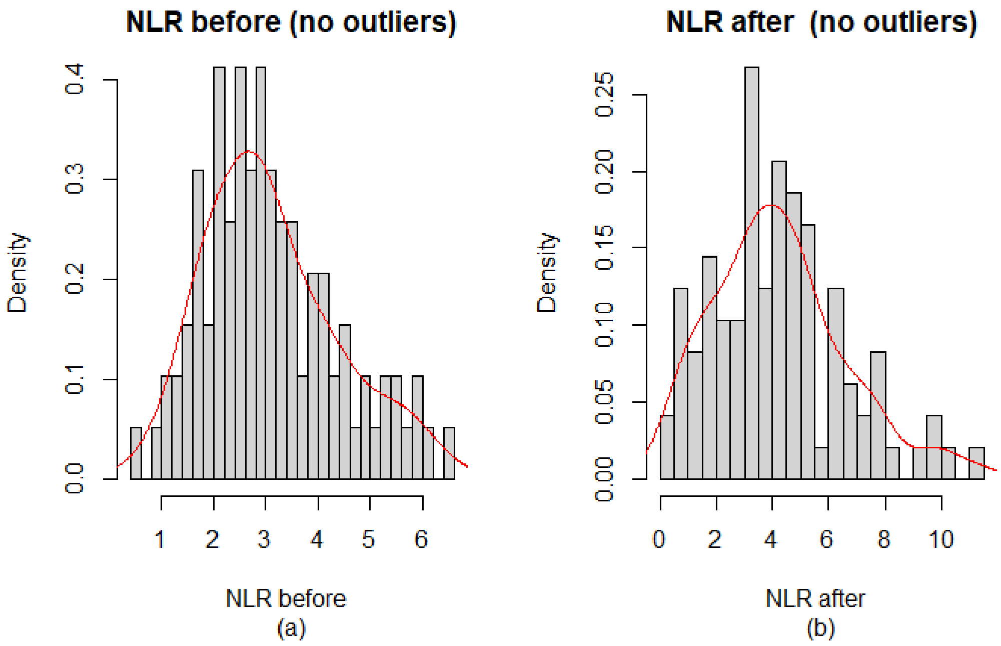


Figure B.270: Histogram of the Neutrophil-to-Lymphocyte Ratio (NLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=74), 1 = Progressive Disease (n=23), total n=97)

B.2.2.19 Platelet-to-Lymphocyte Ratio

PLR before		**PLR after**	
Patient ID	*Value*	*Patient ID*	*Value*
001PANC0002	412.79	001PANC0004	291.91
001PANC0036	915.00	078PANC0002	294.63
001PANC0005	587.32	151PP20002	330.00
001PANC0051	395.33	151PP20004	725.00
001PANC0043	419.12	151PP20017	330.00
001PP20032	465.91	151PP20013	274.62
		151PP20011	323.33
		165PP20008	273.79
		005PP20010	287.78
		151PP20029	296.25

Table B.86: Overview of the outlier values determined using the IQR-method for the Platelet-to-Lymphocyte Ratio measured before and after the first chemotherapy cycle.

B.2.3 Difference before and after first cycle outlier analysis

The difference between a variable before and after the first chemotherapy cycle might also be indicative for the final response to chemotherapy. Therefore, the outlier analysis is also performed on these variables by grouping them together based on their properties (e.g. tumor marker, blood cells).

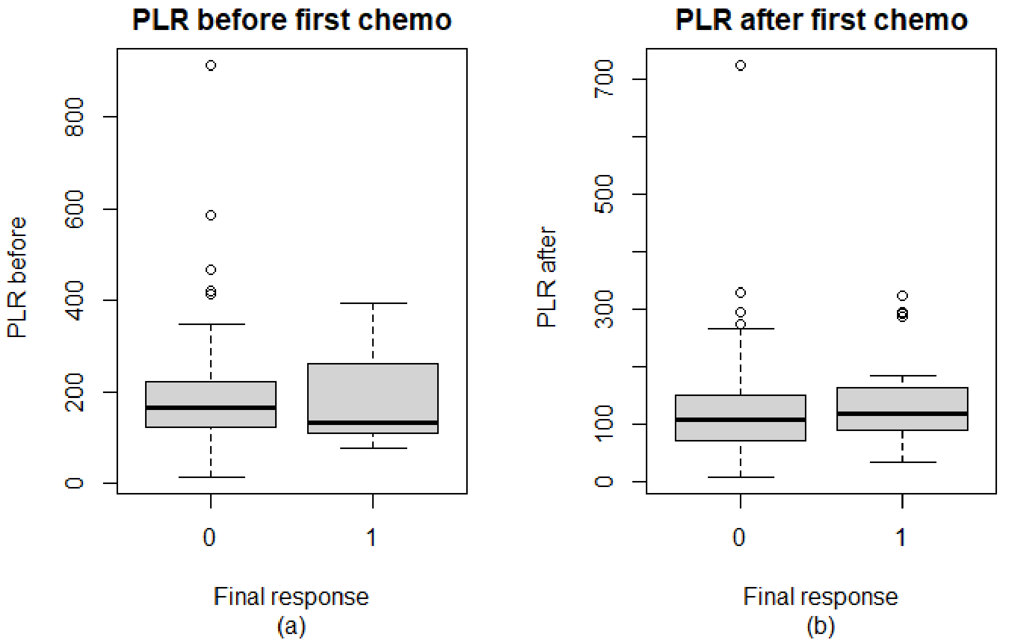


Figure B.271: Boxplot of the Platelet-to-Lymphocyte Ratio (PLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=89), 1 = Progressive Disease (n=24), total n=113)

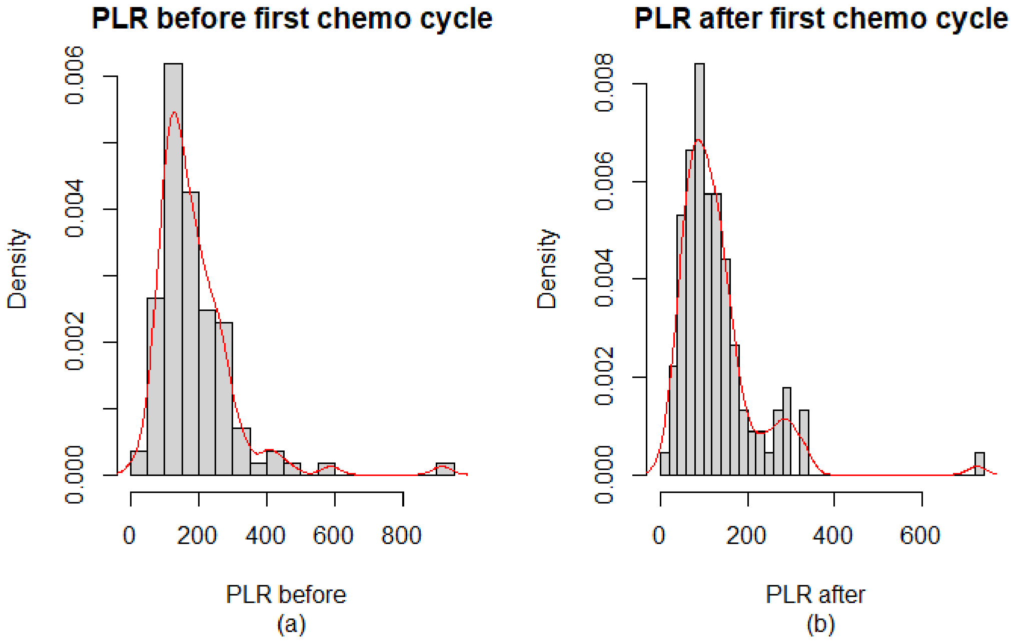


Figure B.272: Histogram of the Platelet-to-Lymphocyte Ratio (PLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=89), 1 = Progressive Disease (n=24), total n=113)

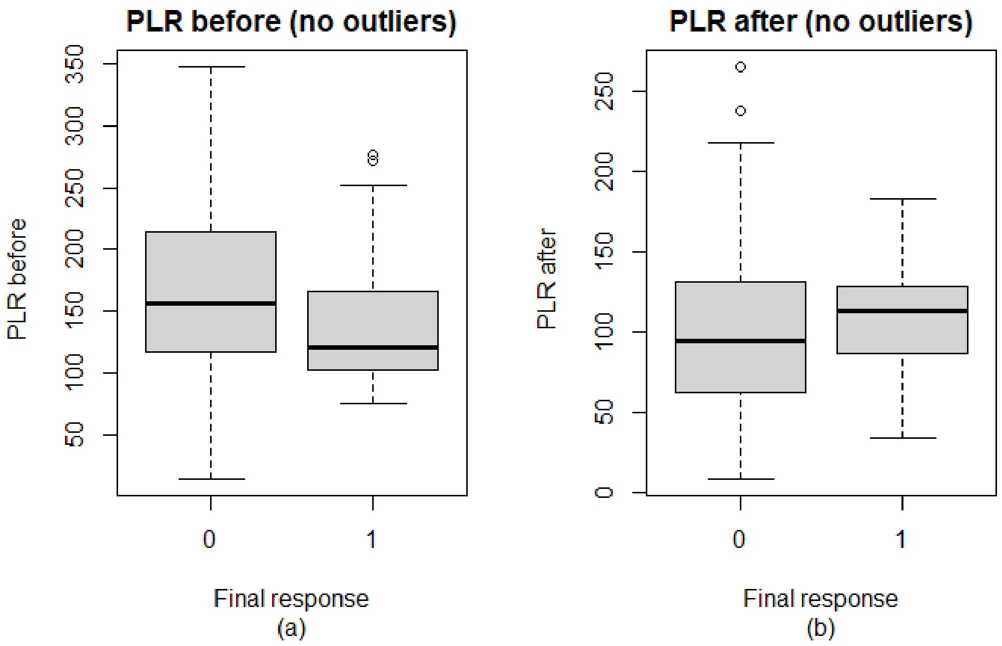


Figure B.273: Boxplot of the Platelet-to-Lymphocyte Ratio (PLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=78), 1 = Progressive Disease (n=19), total n=97)

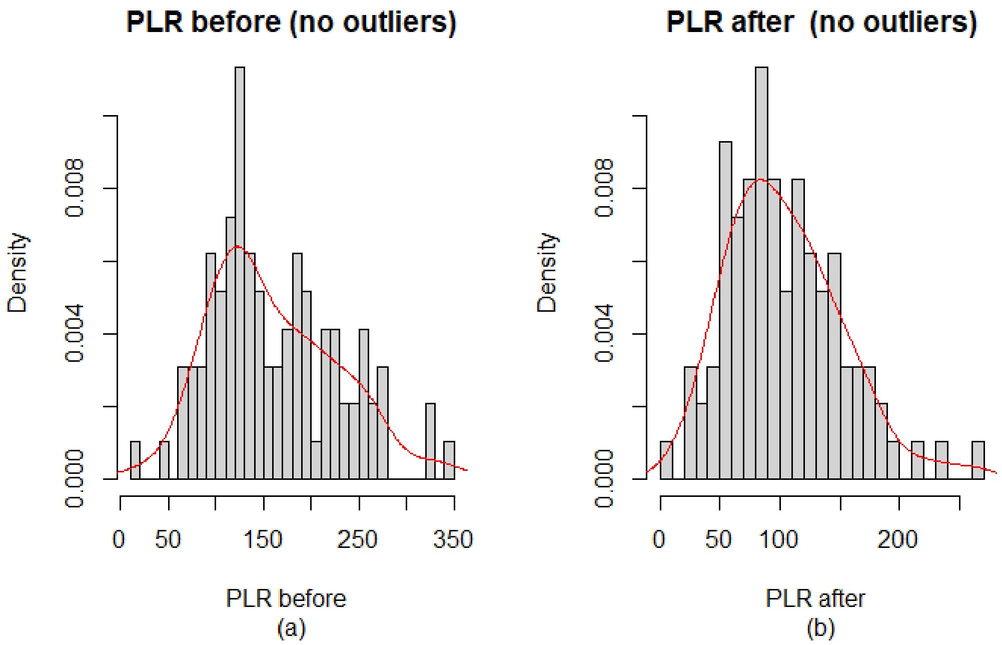


Figure B.274: Histogram of the Platelet-to-Lymphocyte Ratio (PLR) values (a) before the first chemotherapy cycle and (b) after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=78), 1 = Progressive Disease (n=19), total n=97)

B.2.3.1 Tumor markers CA19-9 and CEA difference

CA19-9 difference		**CEA difference**	
Patient ID	*Value (kU/L)*	*Patient ID*	*Value ($\mu g/L$)*
001PANC0015	722	001PANC0015	3.89
001PANC0002	702	001PANC0035	-5.60
001PANC0035	-9258	001PANC0009	5.30
001PANC0012	542	001PANC0025	5.90
001PANC0009	722	001PANC0023	8.63
001PANC0025	4833	001PANC0032	12.60
001PANC0037	5860	001PANC0013	-3.94
001PANC0004	3670	001PANC0033	9.40
001PANC0010	-483	001PANC0019	-15.00
001PANC0033	-507	001PANC0011	5.20
001PANC0019	-1912	001PP20022	-9.80
001PANC0011	1393	002PP20011	20.30
001PP20007	-649	065PP20005	-59.60
001PP20022	-4069	078PANC0001	97.60
078PP20007	-511	078PP20005	3.70
165PP20003	398	078PP20008	5.10
001PANC0038	321	001PANC0058	-5.15
001PANC0050	291	001PANC0050	4.86
005PP20004	-932	078PP20028	8.90
020PP20006	-300	001PP20045	-20.00
065PP20014	-373	078PP20038	3.80
151PP20011	360		
001PANC0062	-918		
001PP20045	-887		
001PP20048	-752		
005PP20020	377		
005PP20010	-778		

Table B.87: Overview of the outlier values determined using the IQR-method for the CA19-9 (kU/L) and CEA ($\mu g/L$) difference values before and after the first chemotherapy cycle.

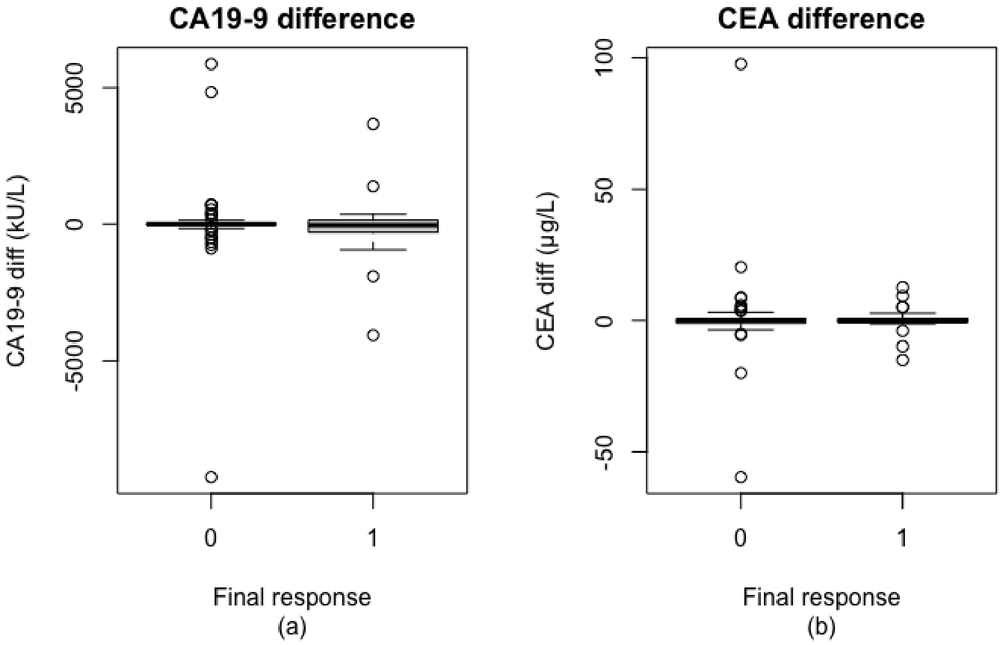


Figure B.275: Boxplot of the tumor markers (a) CA19-9 and (b) CEA difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=90), 1 = Progressive Disease (n=25), total n=115)

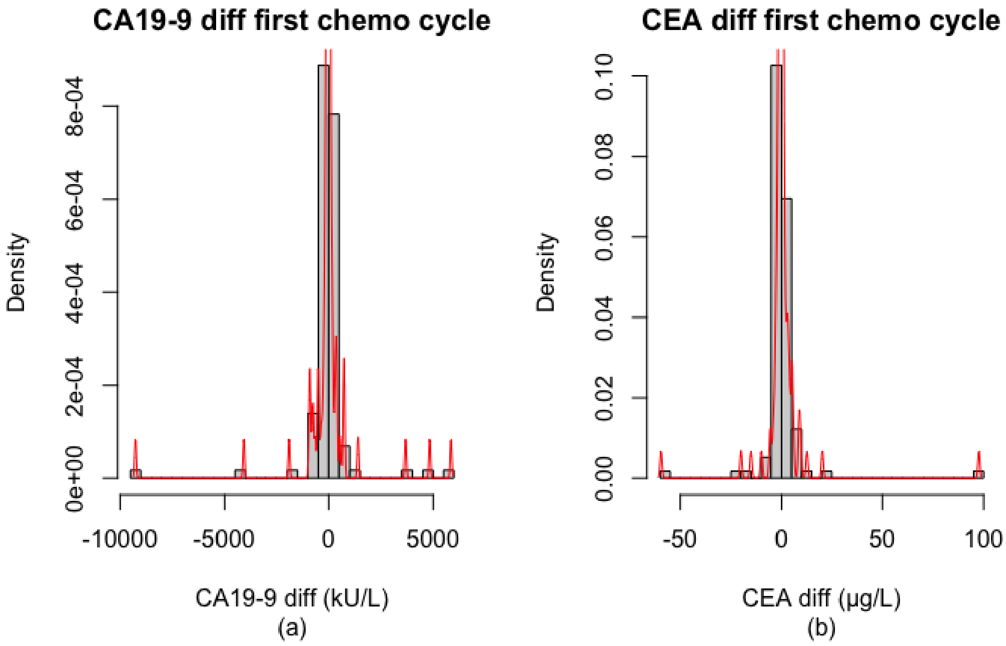


Figure B.276: Histogram of the tumor markers (a) CA19-9 and (b) CEA difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=90), 1 = Progressive Disease (n=25), total n=115)

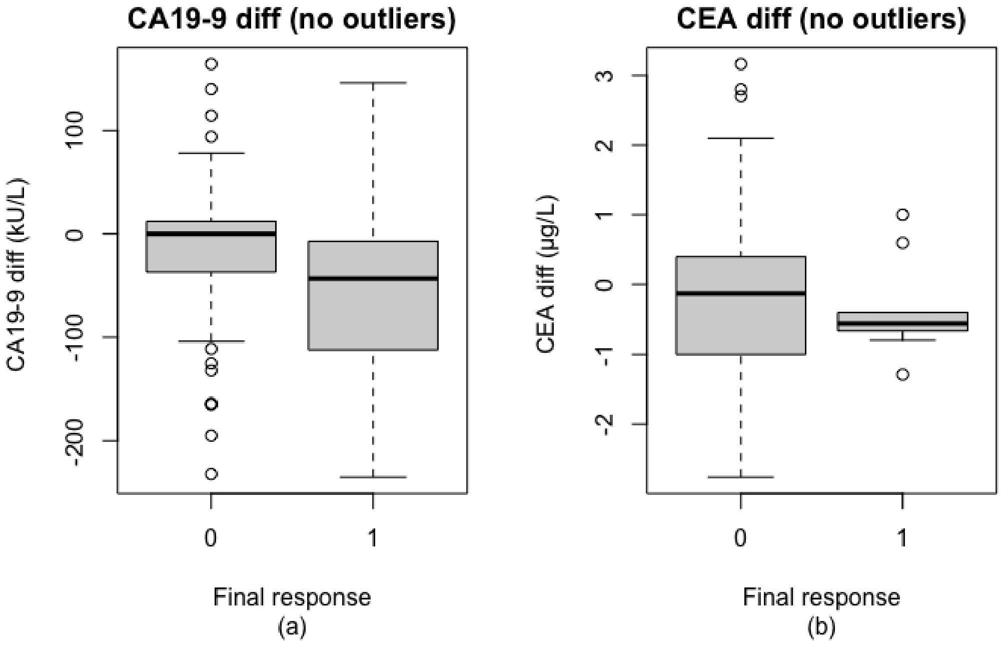


Figure B.277: Boxplot of the tumor markers (a) CA19-9 and (b) CEA difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=66), 1 = Progressive Disease (n=11), total n=77)

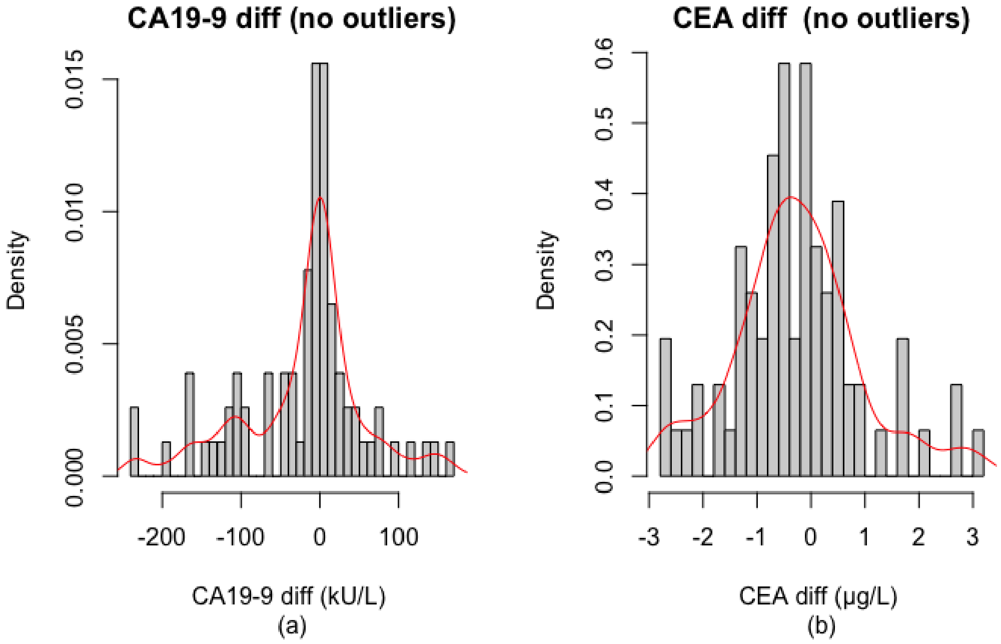


Figure B.278: Histogram of the tumor markers (a) CA19-9 and (b) CEA difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=66), 1 = Progressive Disease (n=11), total n=77)

B.2.3.2 Blood variables: Hemoglobin, Thrombocytes, Leukocytes, Neutrophils, Lymphocytes

HB difference		**TB difference**	
Patient ID	*Value (mmol/L)*	*Patient ID*	*Value ($10^9/L$)*
001PP20036	-1.4	001PANC0004	-228
078PP20028	-1.7	002PP20008	-250
133PP20008	-1.7	002PP20005	291
165PP20018	-1.4	059PP20001	-230
		065PP20005	-132
		001PANC0051	370
		001PANC0038	-105
		005PP20004	247
		078PP20016	-113
		133PP20007	398
		148PP20025	247
		151PP20018	-192
		151PP20013	-158
		165PP20006	-304
		065PP20017	-192
		165PP20014	-168
		165PP20018	-155

Table B.88: Overview of the outlier values determined using the IQR-method for the Hemoglobin (mmol/L) and Thrombocyte ($10^9/L$) difference values before and after the first chemotherapy cycle.

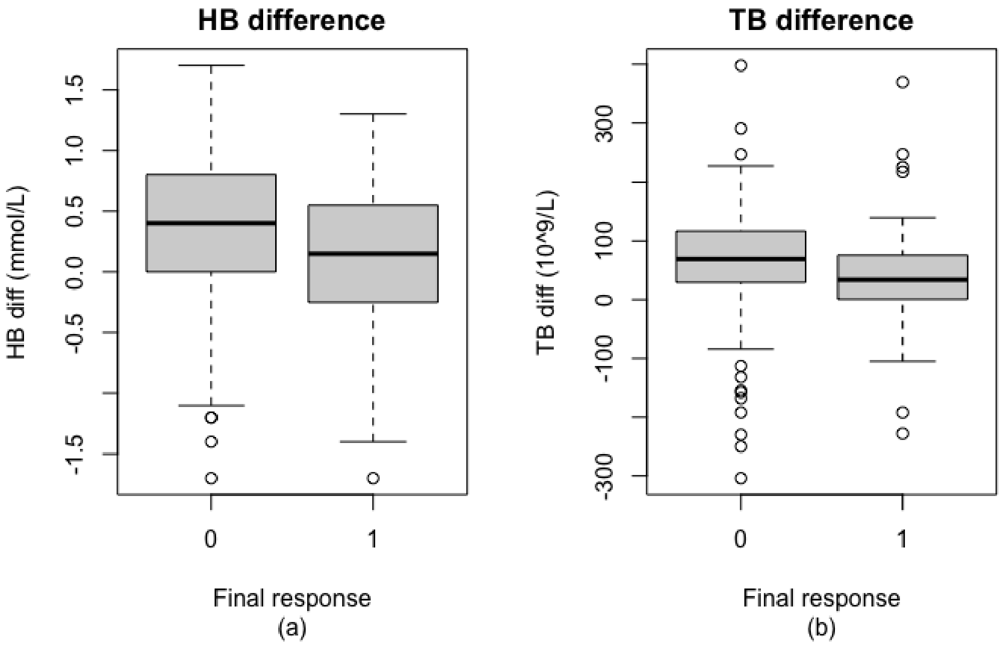


Figure B.279: Boxplot of the (a) Hemoglobin values and (b) Thrombocyte count difference before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=166), 1 = Progressive Disease (n=40), total n=206)

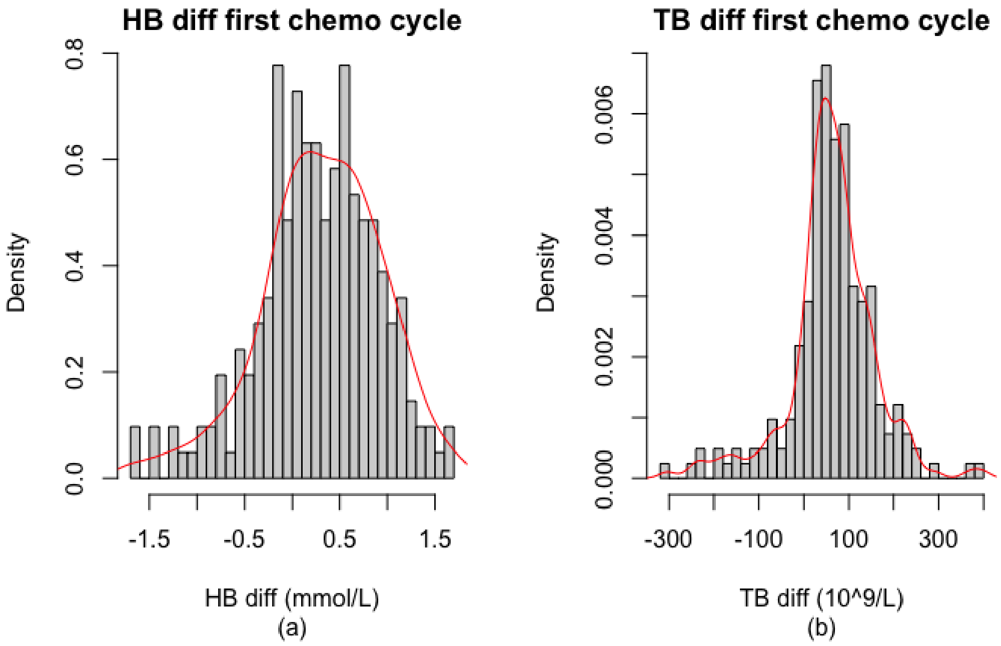


Figure B.280: Histogram of the (a) Hemoglobin values and (b) Thrombocyte count difference before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=166), 1 = Progressive Disease (n=40), total n=206)

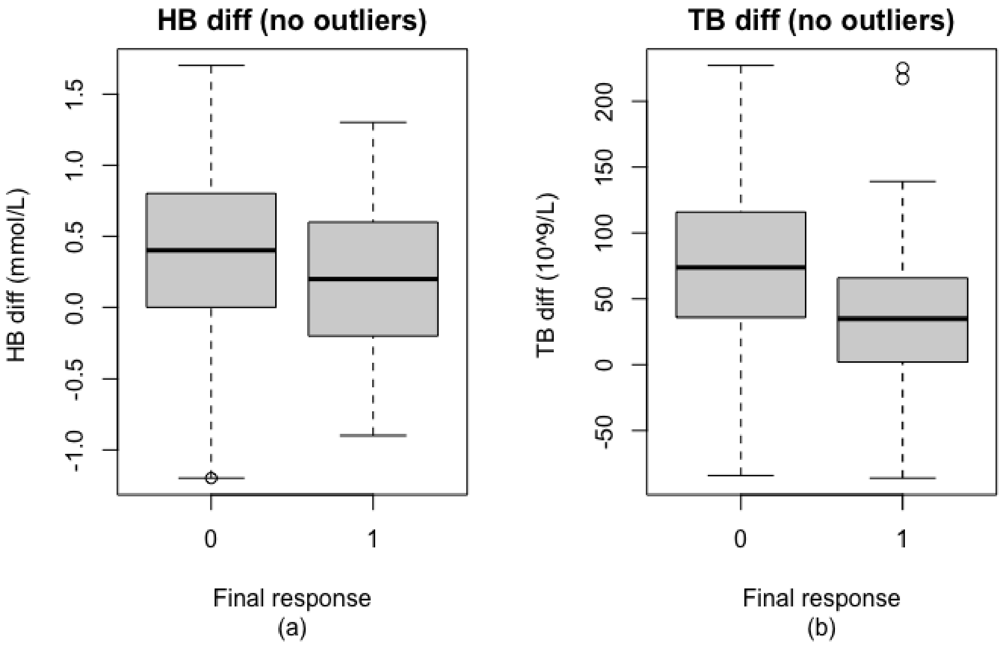


Figure B.281: Boxplot of the (a) Hemoglobin values and (b) Thrombocyte count difference before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=153), 1 = Progressive Disease (n=33), total n=186)

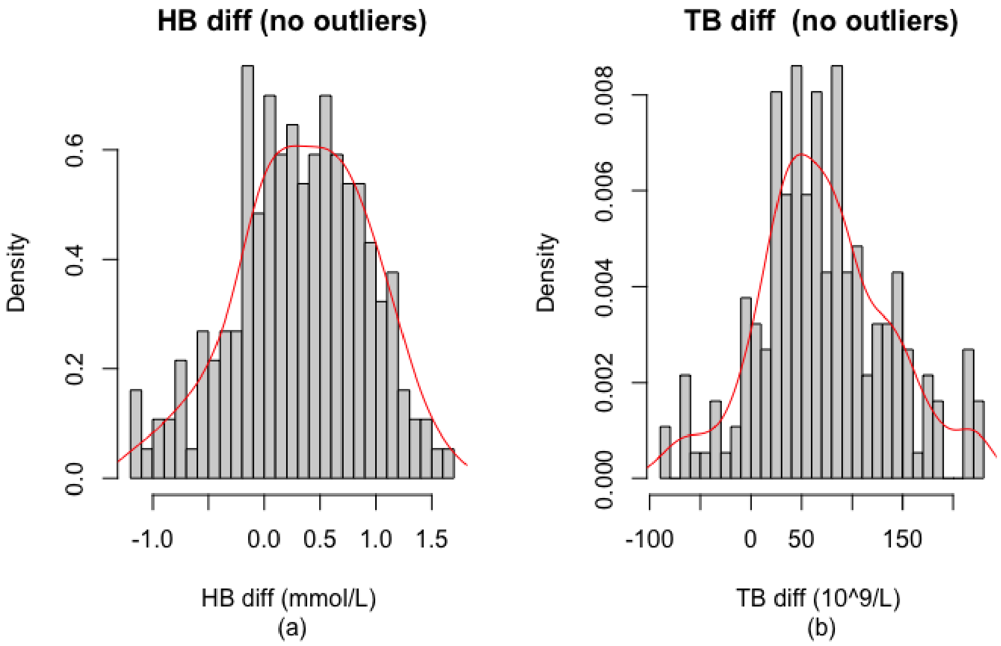


Figure B.282: Histogram of the (a) Hemoglobin values and (b) Thrombocyte count difference before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=153), 1 = Progressive Disease (n=33), total n=186)

LK difference		**NP difference**		**LC difference**	
Patient ID	*Value* $(10^9/L)$	*Patient ID*	*Value* $(10^9/L)$	*Patient ID*	*Value* $(10^9/L)$
001PANC0035	-22.2	001PANC0014	-22.6	001PANC0029	-2.5
001PANC0014	-30.9	001PANC0023	-21.4	001PANC0014	-2.8
001PANC0023	-26.6	001PP20017	-21.2	148PP20019	-3.1
001PP20017	-29.1	148PP20006	-18.2	151PP20034	2.1
148PP20006	-22.5				

Table B.89: Overview of the outlier values determined using the IQR-method for the Leukocytes, Neutrophils and Lymphocyte count $(10^9/L)$ difference before and after the first chemotherapy cycle.

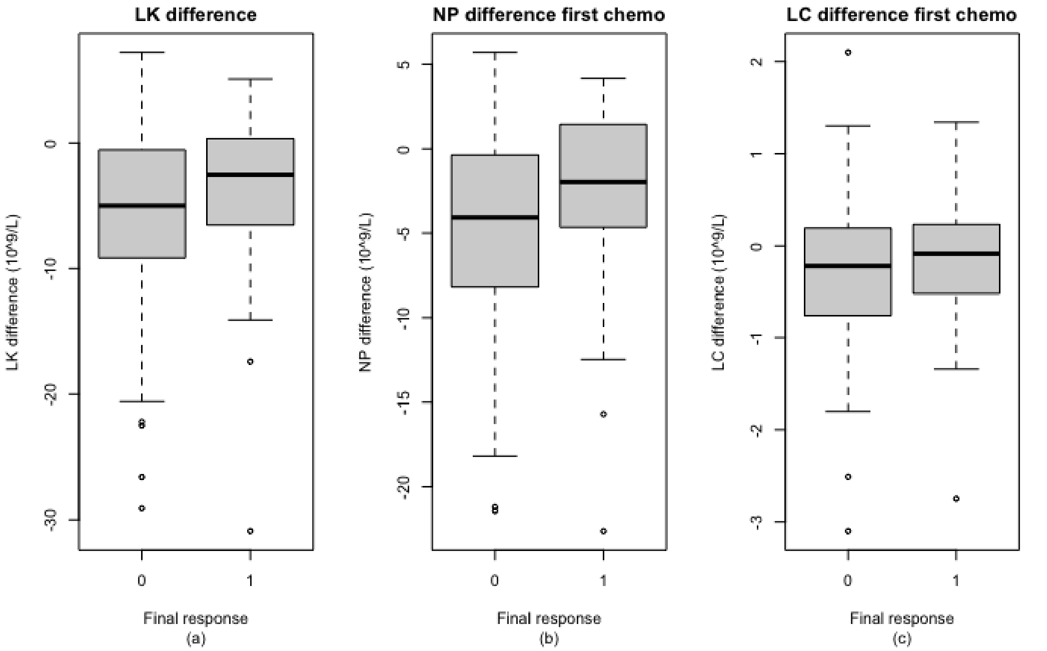


Figure B.283: Boxplot of the (a) Leukocytes, (b) Neutrophils and (c) Lymphocytes count difference before after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=87), 1 = Progressive Disease (n=24), total n=111)

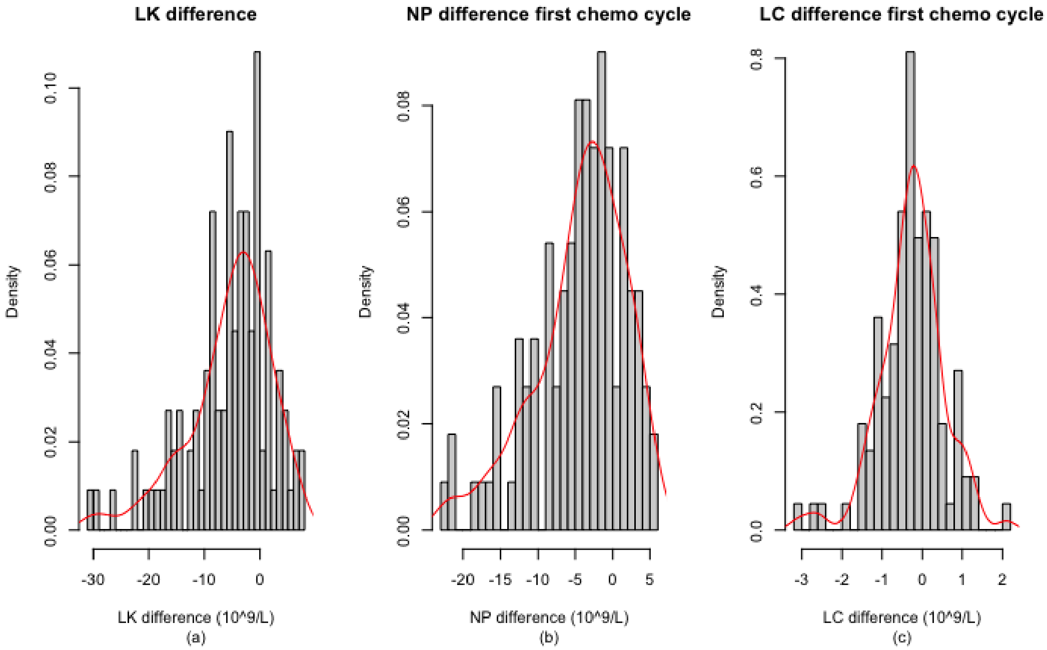


Figure B.284: Histogram of the (a) Leukocytes, (b) Neutrophils and (c) Lymphocytes count difference before after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=87), 1 = Progressive Disease (n=24), total n=111)

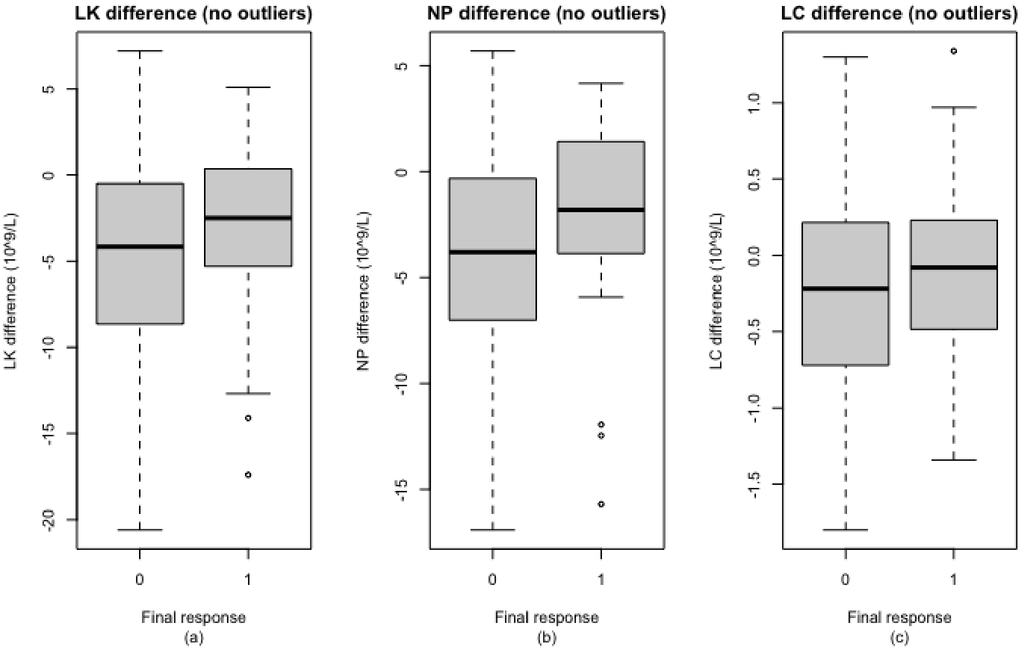


Figure B.285: Boxplot of the (a) Leukocytes, (b) Neutrophils and (c) Lymphocytes count difference before after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=80), 1 = Progressive Disease (n=23), total n=103)

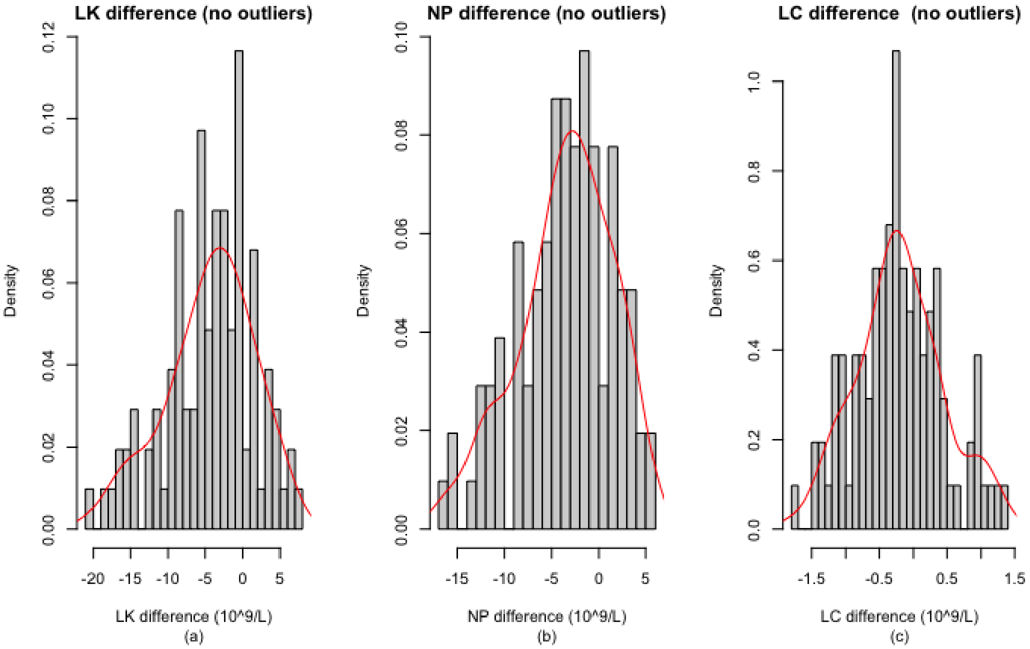


Figure B.286: Histogram of the (a) Leukocytes, (b) Neutrophils and (c) Lymphocytes count difference before after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=80), 1 = Progressive Disease (n=23), total n=103)

B.2.3.3 Kidney function: Creatinin and Glomular Filtration Rate

CR **difference**		**GFR** **difference**	
Patient ID	*Value* *($\mu mol/L$)*	*Patient ID*	*Value* *(mL/min)*
001PANC0015	-23	001PANC0028	12
001PANC0011	61	001PANC0015	23
001PP20008	-82	001PANC0024	12
078PP20007	22	001PANC0017	12
151PP20009	-20	001PANC0011	-61
001PANC0058	18	001PP20007	-13
001PANC0039	20	001PP20008	42
001PANC0050	26	002PANC0002	10
001PP20032	-23	002PP20011	11
078PP20026	-104	002PP20008	13
078PP20028	-20	059PP20001	-16
148PP20013	22	078PP20007	-19
001PP20045	22	078PP20006	-10
001PP20040	-843	151PP20004	13
151PP20034	-25	151PP20009	15
		001PANC0053	14
		001PANC0039	-13
		001PANC0041	20
		001PANC0038	-13
		001PANC0050	-22
		001PP20031	-12
		001PP20032	23
		001PP20030	11
		002PP20037	-12
		002PP20063	10
		065PP20015	12
		078PP20026	54
		078PP20028	21
		133PP20006	13
		151PP20017	-16
		151PP20013	-10
		151PP20021	-14
		001PP20040	-13
		065PP20017	-8
		078PP20038	16
		151PP20025	-9
		151PP20034	20

Table B.90: Overview of the outlier values determined using the IQR-method for the Creatinin ($\mu mol/L$) and Glomular Filtration Rate (mL/min) difference values before and after the first chemotherapy cycle.

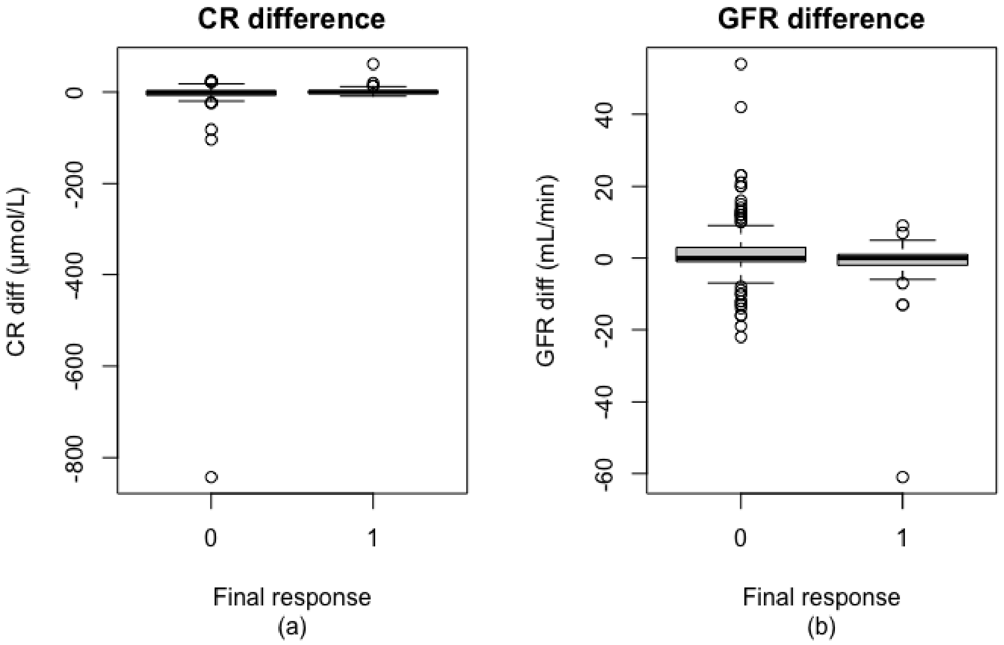


Figure B.287: Boxplot of the (a) Creatinin and (b) Glomular Filtration rate difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=168), 1 = Progressive Disease (n=41), total n=209)

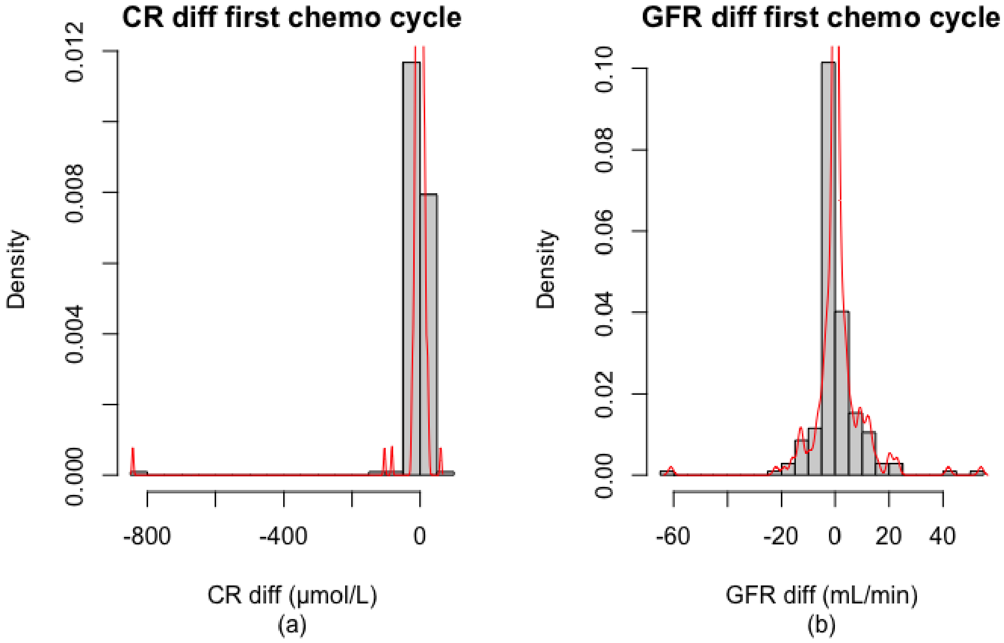


Figure B.288: Histogram of the (a) Creatinin and (b) Glomular Filtration rate difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=168), 1 = Progressive Disease (n=41), total n=209)

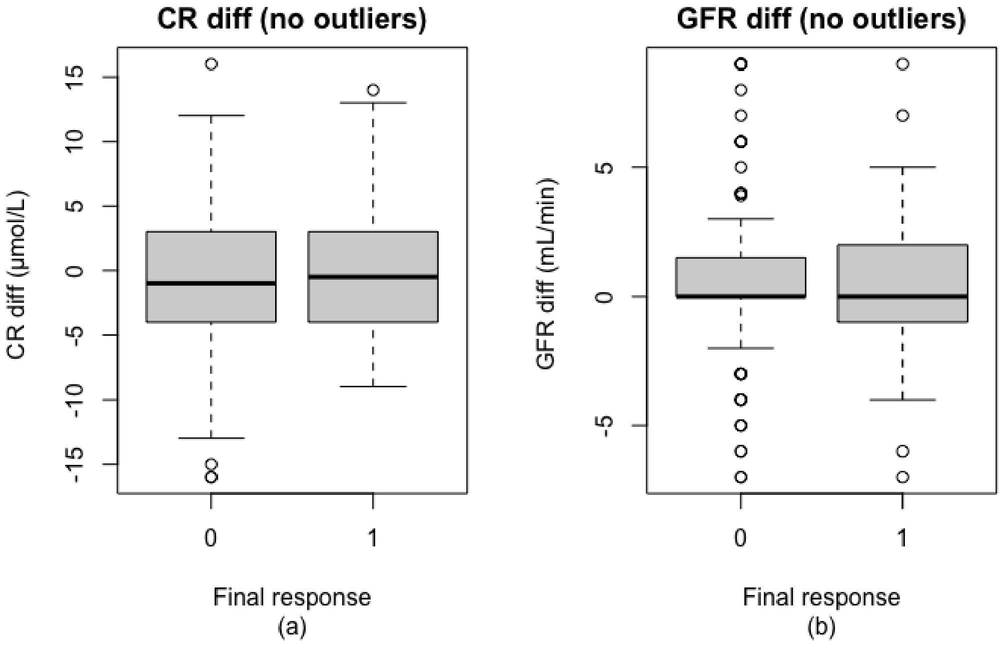


Figure B.289: Boxplot of the (a) Creatinin and (b) Glomular Filtration rate difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=131), 1 = Progressive Disease (n=38), total n=169)

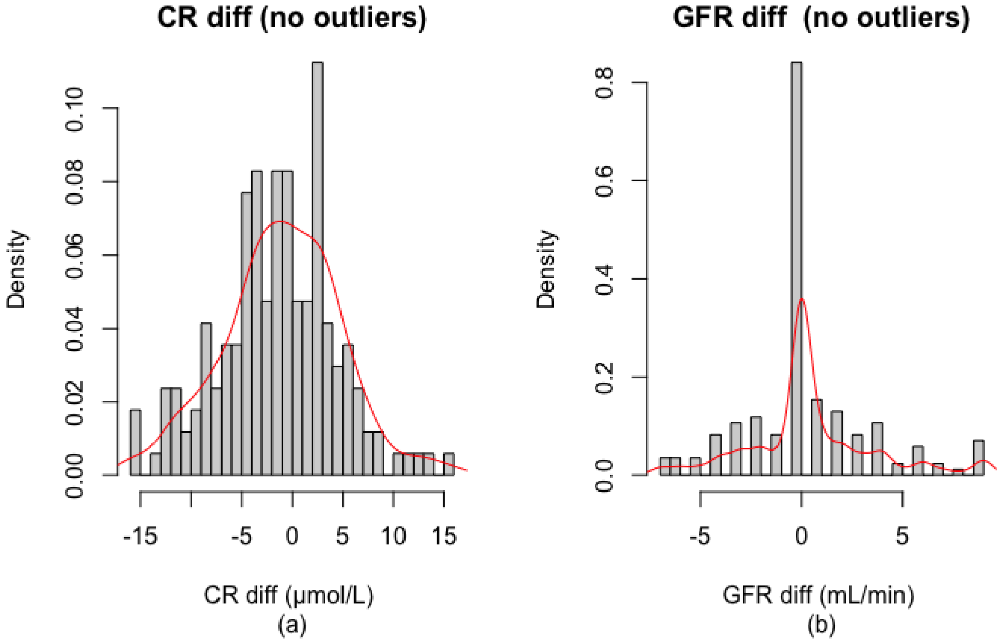


Figure B.290: Histogram of the (a) Creatinin and (b) Glomular Filtration rate difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=131), 1 = Progressive Disease (n=38), total n=169)

B.2.3.4 Nutrition status: Sodium, Potassium and Albumin

Na difference		**K difference**		**Alb difference**	
Patient ID	*Value (mmol/L)*	*Patient ID*	*Value (mmol/L)*	*Patient ID*	*Value (g/L)*
		001PANC0015	1.3	001PANC0043	-6
		001PANC0037	1.8	001PANC0050	14
		001PANC0011	1.5	165PP20012	34
		001PANC0051	-0.8		
		001PANC0061	-1.0		
		001PP20045	1.4		

Table B.91: Overview of the outlier values determined using the IQR-method for the Sodium (mmol/L), Potassium (mmol/L) and Albumin (g/L) difference before and after the first chemotherapy cycle.

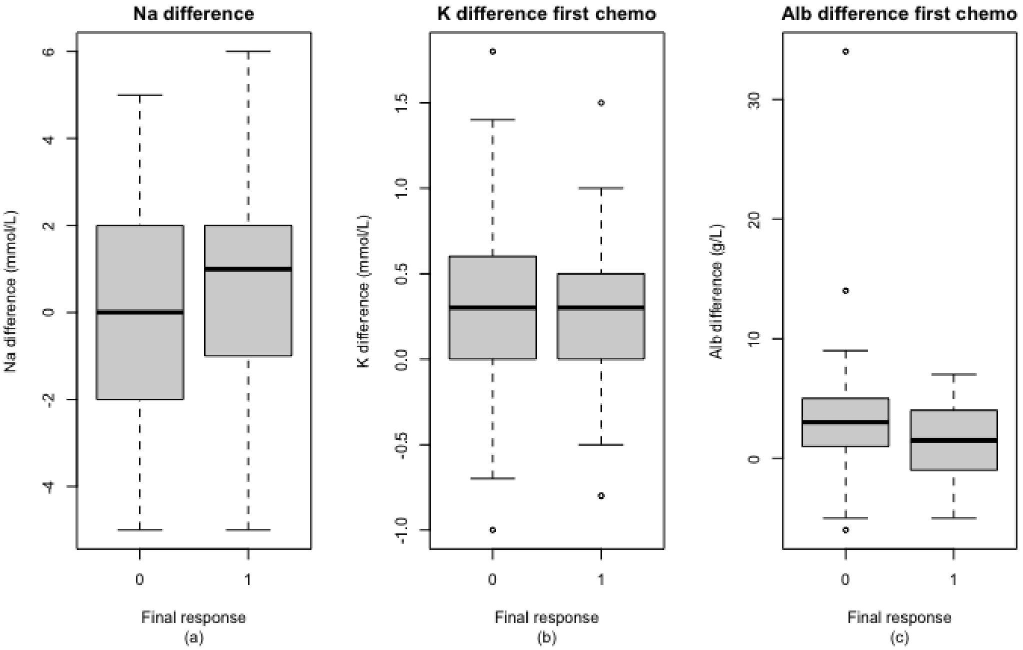


Figure B.291: Boxplot of the (a) Sodium, (b) Potassium and (c) Albumin difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=80), 1 = Progressive Disease (n=22), total n=102)

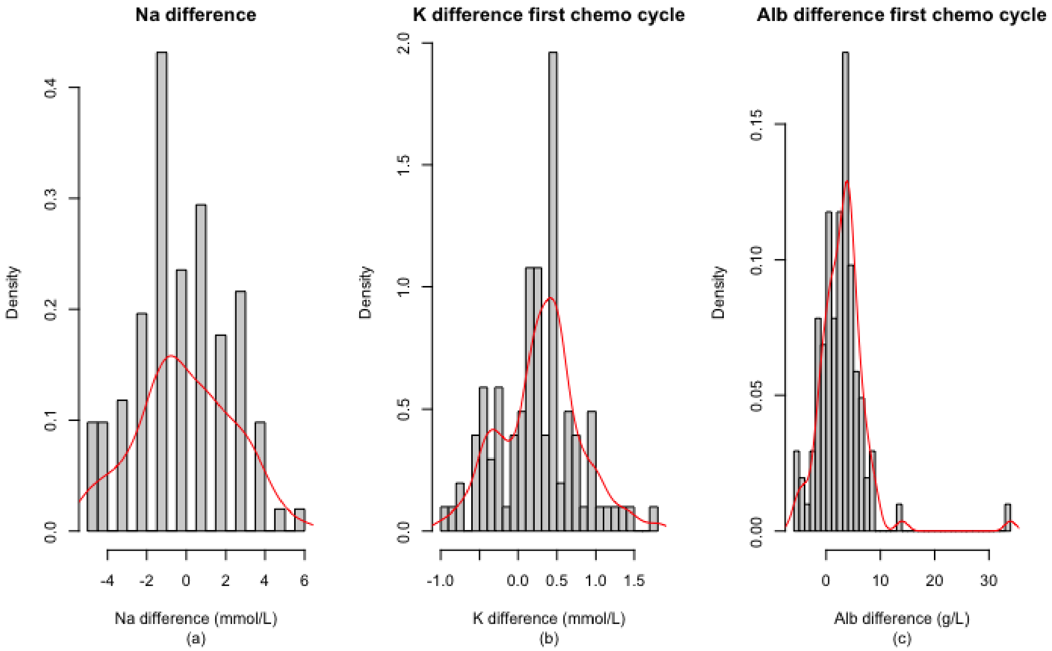


Figure B.292: Histogram of the (a) Sodium, (b) Potassium and (c) Albumin difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=80), 1 = Progressive Disease (n=22), total n=102)

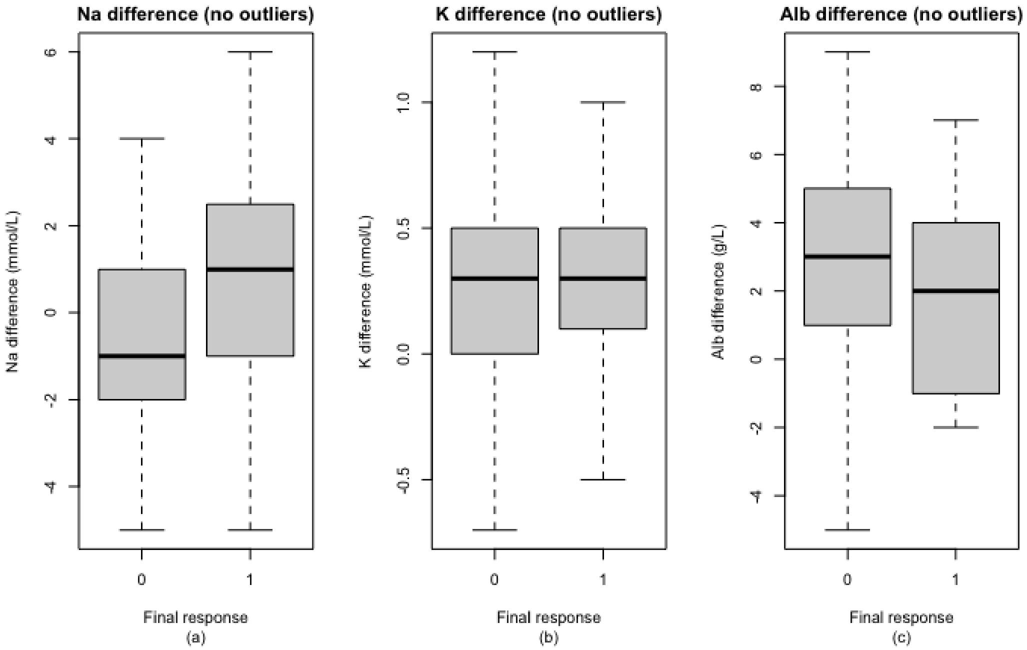


Figure B.293: Boxplot of the (a) Sodium, (b) Potassium and (c) Albumin difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=73), 1 = Progressive Disease (n=20), total n=93)

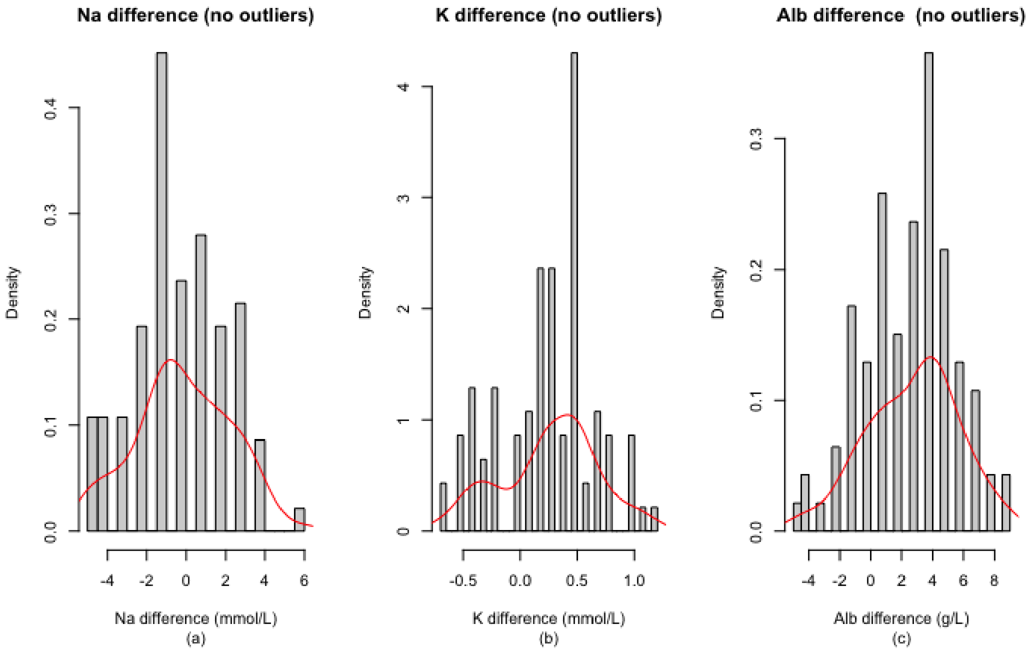


Figure B.294: Histogram of the (a) Sodium, (b) Potassium and (c) Albumin difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=73), 1 = Progressive Disease (n=20), total n=93)

B.2.3.5 Liver function: ASAT, ALAT, Alkaline Phosphatase, γ-Glutamyl Transferase and Bilirubin

ASAT difference		**ALAT difference**	
Patient ID	*Value (U/L)*	*Patient ID*	*Value (U/L)*
001PANC0001	83	001PANC0001	94
001PANC0029	-48	001PANC0029	-125
001PANC0025	148	001PANC0025	345
001PANC0032	-191	001PANC0032	-221
001PANC0004	-177	001PANC0004	-159
001PANC0013	261	001PANC0013	609
001PP20007	-43	001PP20002	84
018PP20003	-137	148PP20002	-125
151PP20009	50	148PP20009	93
001PANC0052	1079	151PP20009	92
001PP20031	-43	001PANC0052	1183
001PP20032	185	001PANC0045	68
002PANC0013	79	001PP20031	-114
002PP20051	200	001PP20032	161
078PP20016	50	002PANC0013	128
133PP20001	243	002PP20046	51
133PP20005	275	002PP20047	75
148PANC0004	40	020PP20006	76
148PANC0008	-53	065PP20015	61
148PP20013	38	078PANC0003	-68
148PP20019	-36	078PP20029	-85
151PP20017	78	078PP20016	116
148PANC0009	103	133PP20001	97
151PP20028	81	133PP20005	265
		148PANC0004	-114
		148PP20013	99
		148PP20027	61
		148PP20019	-75
		151PP20017	147
		165PP20006	-65
		078PP20038	-64
		148PANC0009	312
		151PP20028	157

Table B.92: Overview of the outlier values determined using the IQR-method for the ASAT (U/L) and ALAT (U/L) difference values before and after the first chemotherapy cycle.

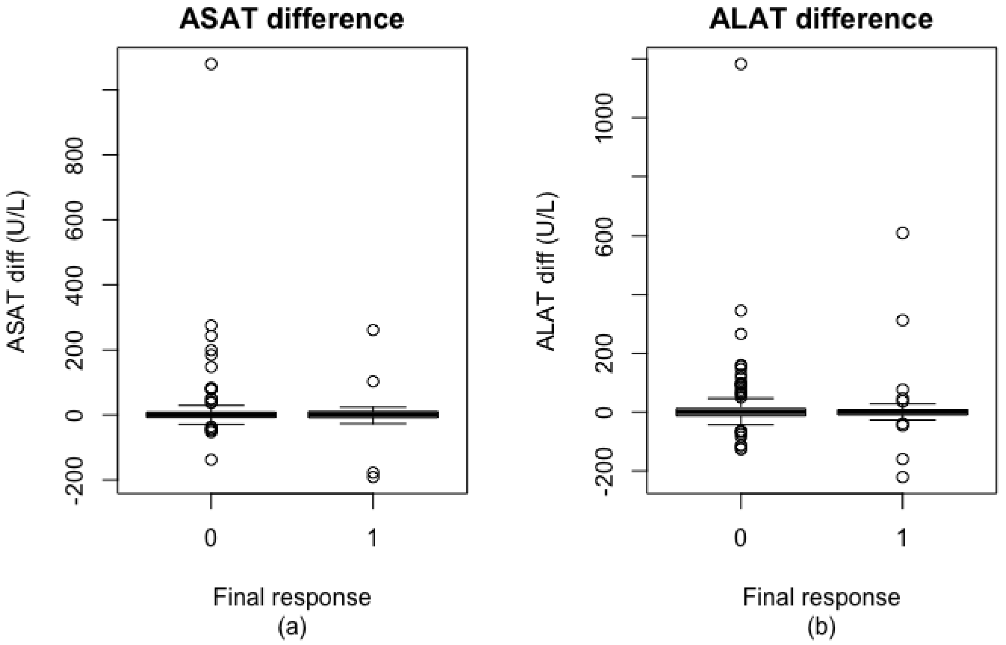


Figure B.295: Boxplot of the (a) ASAT and (b) ALAT difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=144), 1 = Progressive Disease (n=33), total n=177)

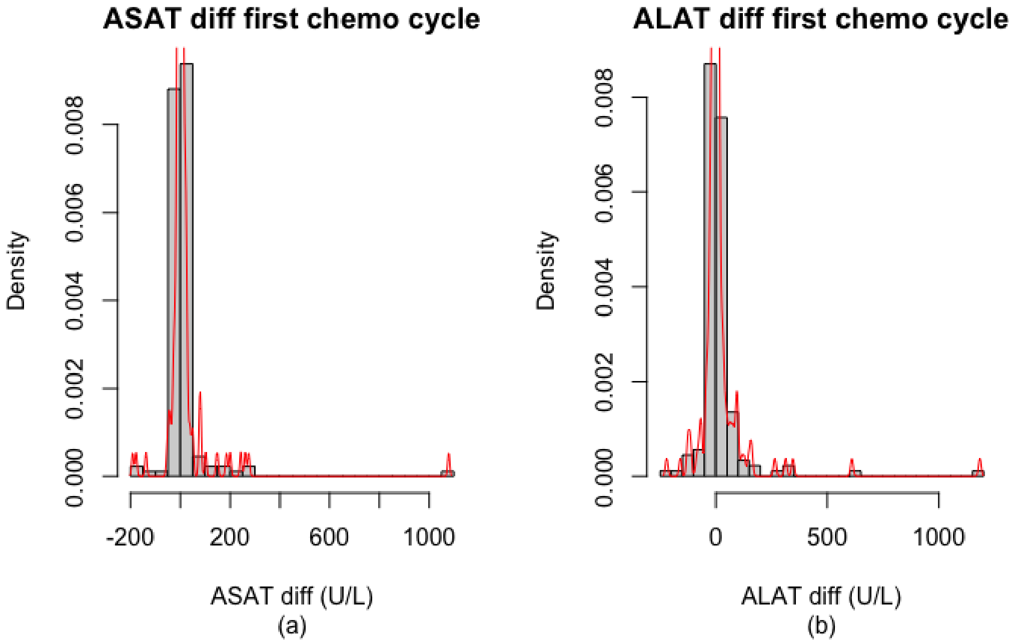


Figure B.296: Histogram of the (a) ASAT and (b) ALAT difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=144), 1 = Progressive Disease (n=33), total n=177)

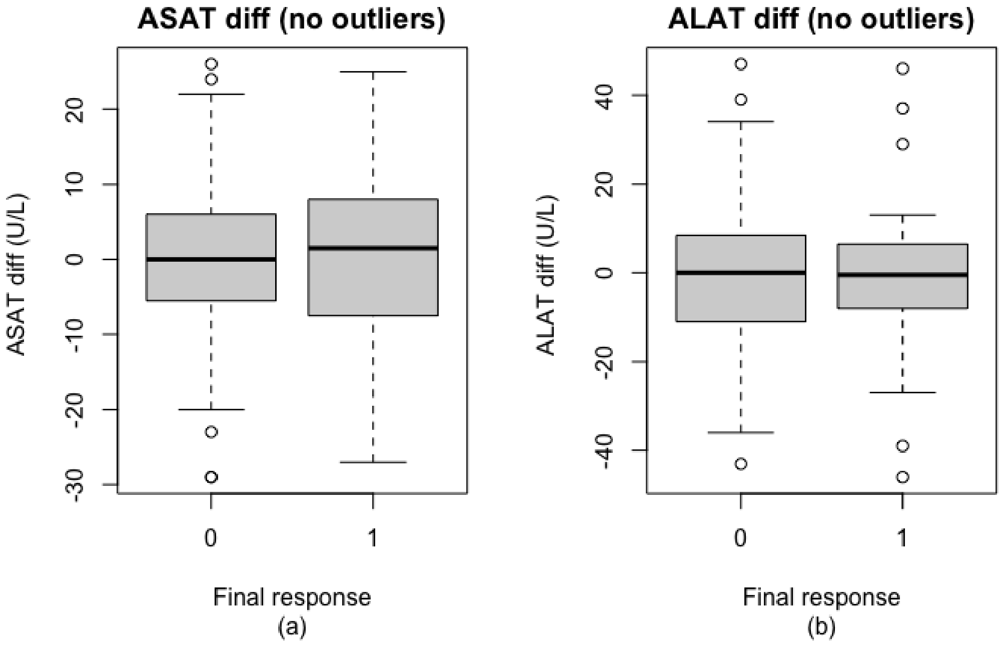


Figure B.297: Boxplot of the (a) ASAT and (b) ALAT difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=112), 1 = Progressive Disease (n=28), total n=140)

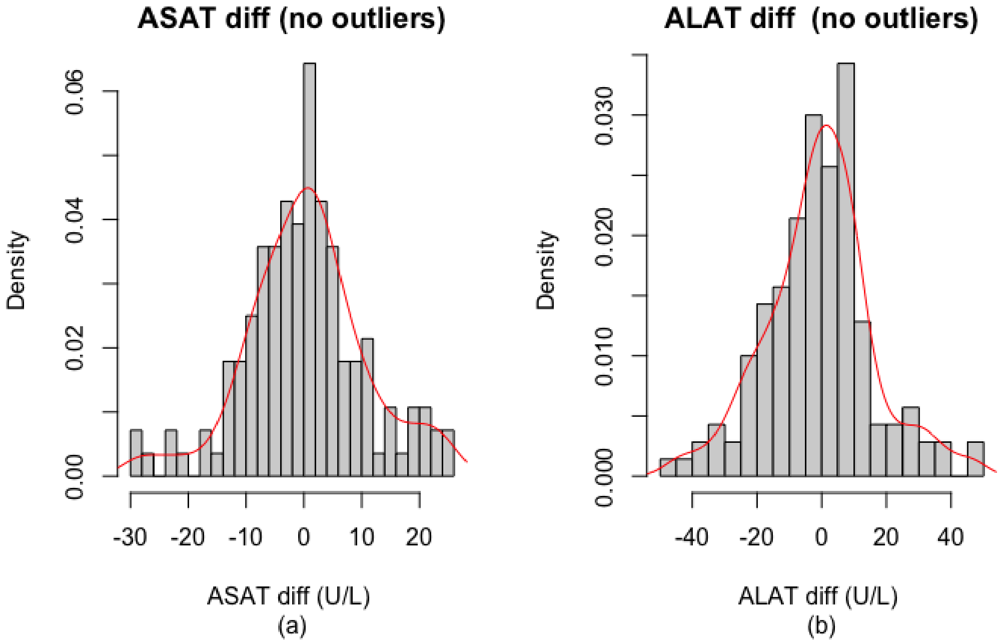


Figure B.298: Histogram of the (a) ASAT and (b) ALAT difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=112), 1 = Progressive Disease (n=28), total n=140)

AF difference		**GGT difference**		**BR difference**	
Patient ID	*Value (U/L)*	*Patient ID*	*Value (U/L)*	*Patient ID*	*Value ($\mu mol/L$)*
001PANC0014	-217	001PANC0015	244	001PANC0025	118
001PANC0004	-485	001PANC0035	-166	001PANC0013	77
001PANC0018	162	001PANC0009	-161	001PANC0052	35
001PANC0033	157	001PANC0025	625	133PP20001	39
148PP20006	173	001PANC0032	-463	133PP20006	39
151PP20004	291	001PANC0004	-275	133PP20005	33
001PP20032	159	001PANC0018	174	148PP20013	68
078PP20016	175	001PANC0013	458	165PP20012	-21
133PP20006	147	001PANC0033	177	165PP20005	32
133PP20005	235	001PP20002	182	151PP20028	43
078PP20040	-351	001PP20012	154		
		002PP20005	170		
		078PP20007	-117		
		148PP20006	240		
		151PP20002	256		
		151PP20004	848		
		001PANC0055	-115		
		001PANC0052	499		
		001PANC0039	-217		
		001PP20031	-123		
		001PP20032	591		
		002PP20047	317		
		020PP20006	332		
		078PP20016	257		
		133PP20005	749		
		148PP20015	208		
		148PP20019	-220		
		078PP20040	-266		
		151PP20034	213		

Table B.93: Overview of the outlier values determined using the IQR-method for the Alkaline Phosphatase (U/L), γ-Glutamyl Transferase (U/L) and Bilirubin ($\mu mol/L$) difference values before and after the first chemotherapy cycle.

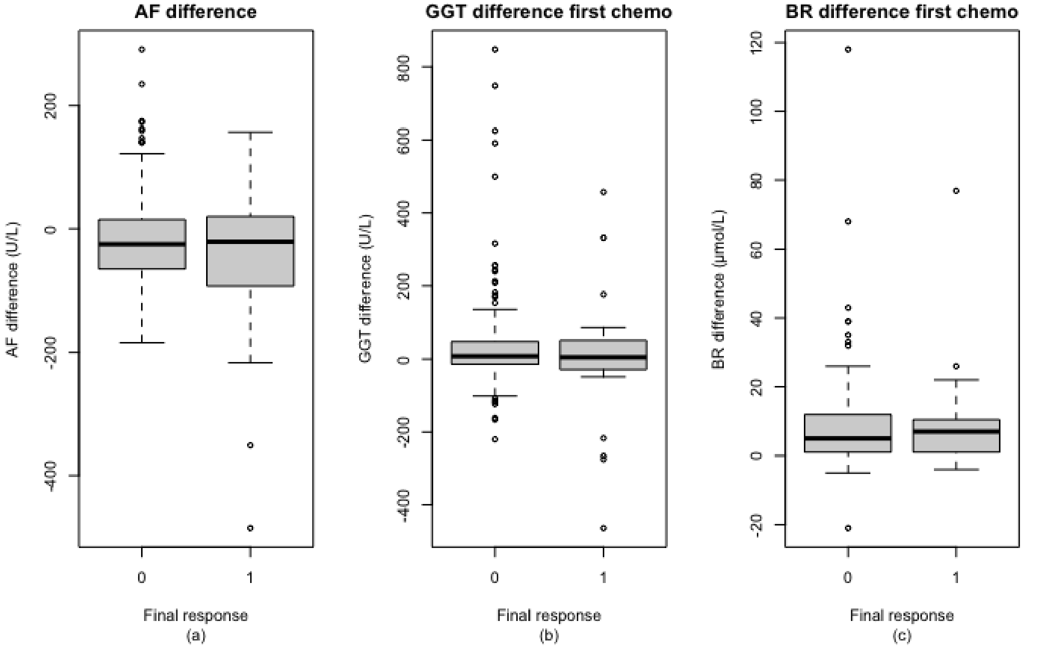


Figure B.299: Boxplot of the (a) Alkaline Phosphatase, (b) γ-Glutamyl Transferase and (c) Bilirubin difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=142), 1 = Progressive Disease (n=32), total n=174)

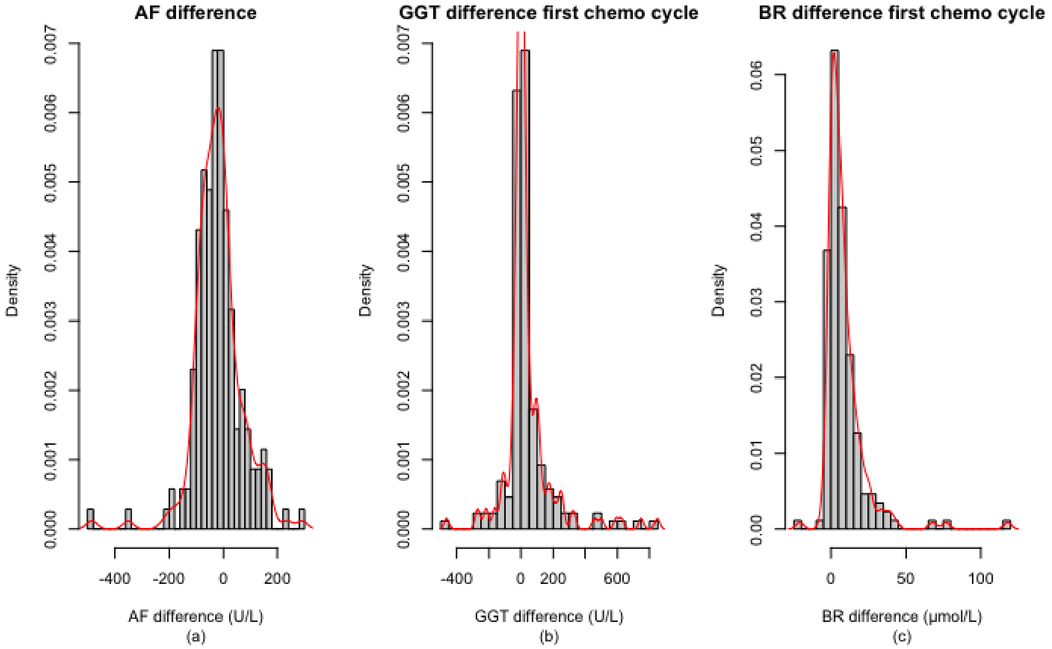


Figure B.300: Histogram of the (a) Alkaline Phosphatase, (b) γ-Glutamyl Transferase and (c) Bilirubin difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=142), 1 = Progressive Disease (n=32), total n=174)

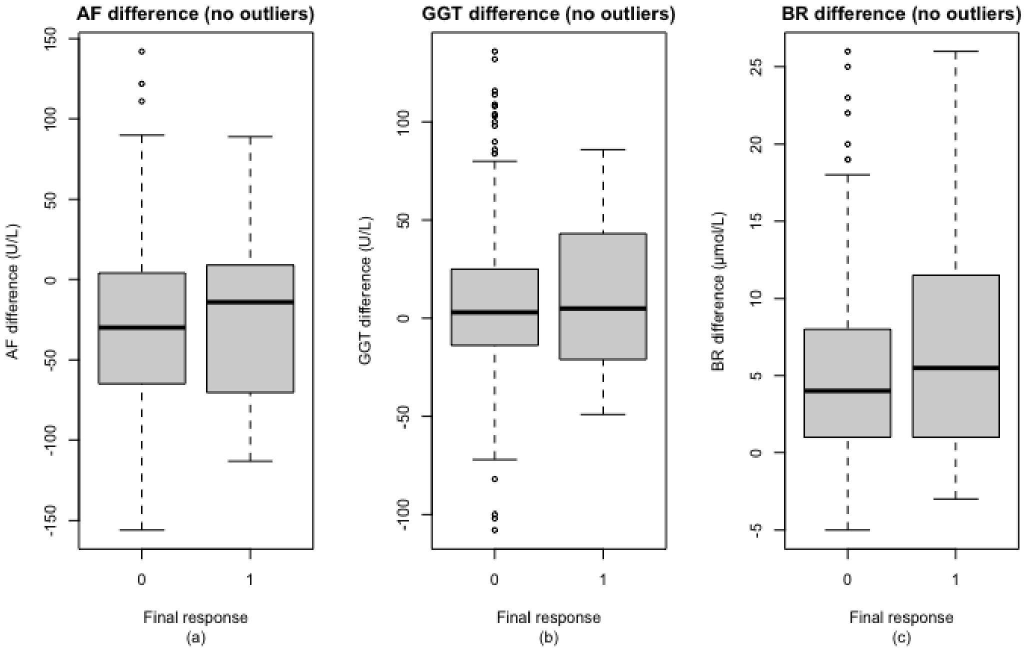


Figure B.301: Boxplot of the (a) Alkaline Phosphatase, (b) γ-Glutamyl Transferase and (c) Bilirubin difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=114), 1 = Progressive Disease (n=24), total n=138)

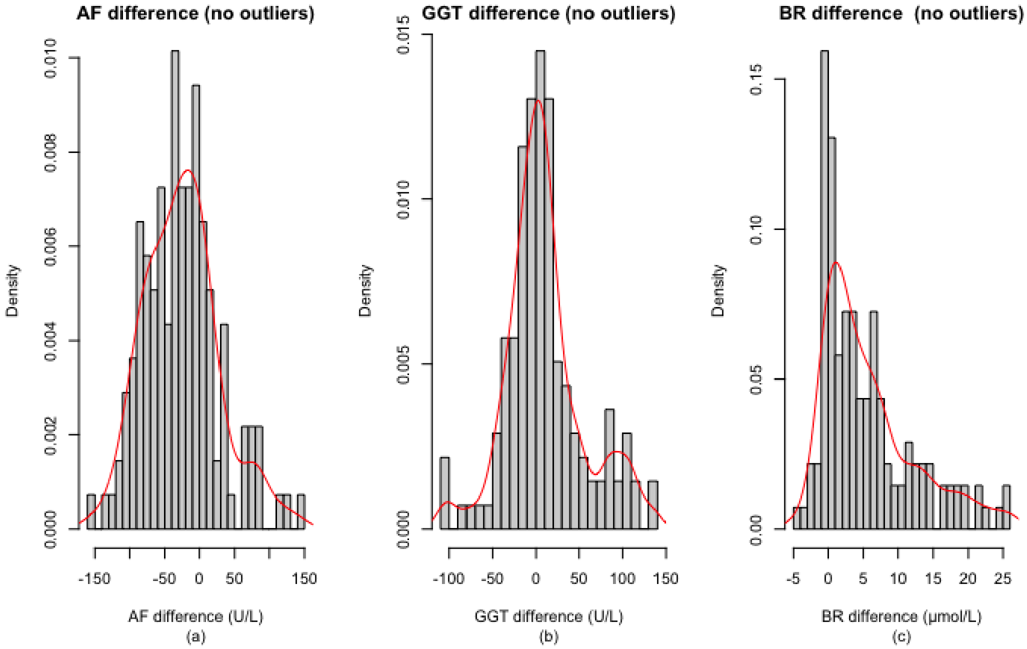


Figure B.302: Histogram of the (a) Alkaline Phosphatase, (b) γ-Glutamyl Transferase and (c) Bilirubin difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=114), 1 = Progressive Disease (n=24), total n=138)

B.2.3.6 Inflammation: C-Reactive Protein, Leukocytes, Systemic Inflammation Index, Neutrophil-to-Lymphocyte Ratio and Platelet-to-Lymphocyte Ratio

CRP difference		**SII difference**	
Patient ID	*Value (mg/L)*	*Patient ID*	*Value*
001PANC0028	30.0	001PANC0015	4301.6
001PANC0026	-41.0	001PANC0035	-1738.6
001PANC0011	-60.0	001PANC0036	6178.6
001PP20012	47.6	001PANC0023	2511.2
065PP20005	-76.0	001PANC0004	-1688.5
078PANC0002	-86.9	001PANC0011	-1798.5
151PP20004	143.0	001PANC0008	-1675.9
001PANC0039	56.0	001PP20015	-1747.6
001PANC0050	-64.8	151PP20004	-4357.6
005PP20004	-36.0	002PP20046	1805.6
078PP20026	-270.3	151PP20013	-3013.9
148PP20013	-68.0		
148PP20015	-63.0		
001PANC0061	-31.0		
005PP20020	58.0		
078PP20038	-34.8		

Table B.94: Overview of the outlier values determined using the IQR-method for the C-Reactive Protein (mg/L) and Systemic Inflammation Index difference values before and after the first chemotherapy cycle.

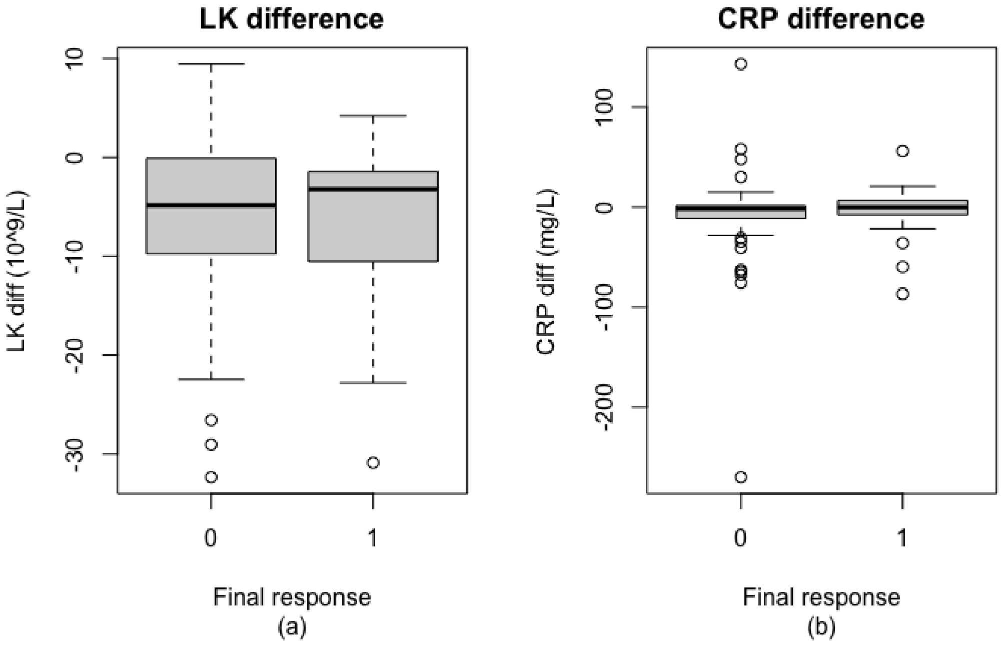


Figure B.303: Boxplot of the (a) Leukocytes and (b) C-Reactive Protein difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=96), 1 = Progressive Disease (n=23), total n=119)

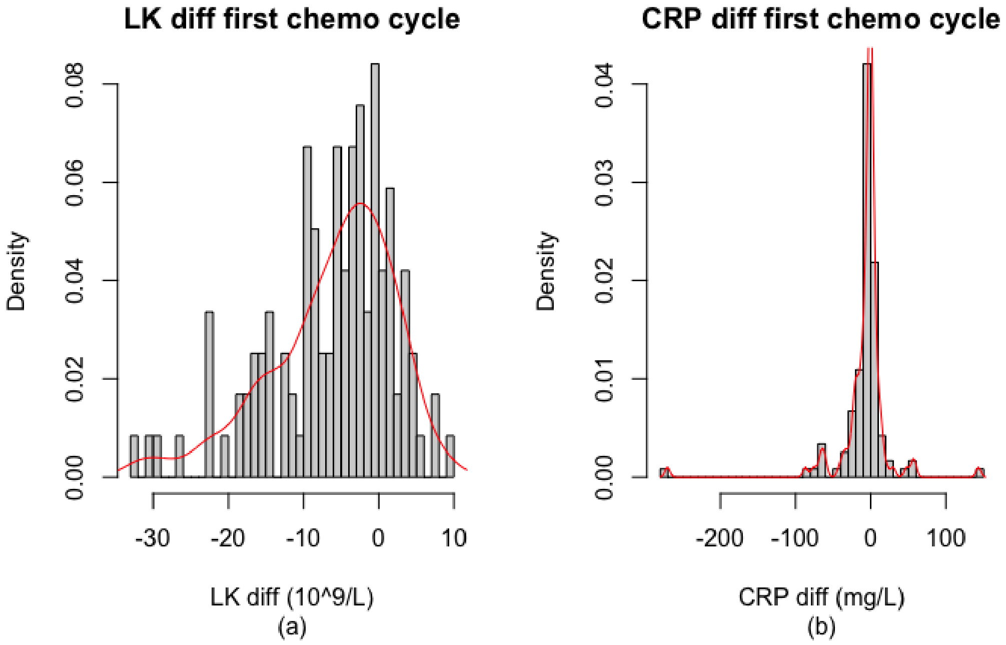


Figure B.304: Histogram of the (a) Leukocytes and (b) C-Reactive Protein difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=96), 1 = Progressive Disease (n=23), total n=119)

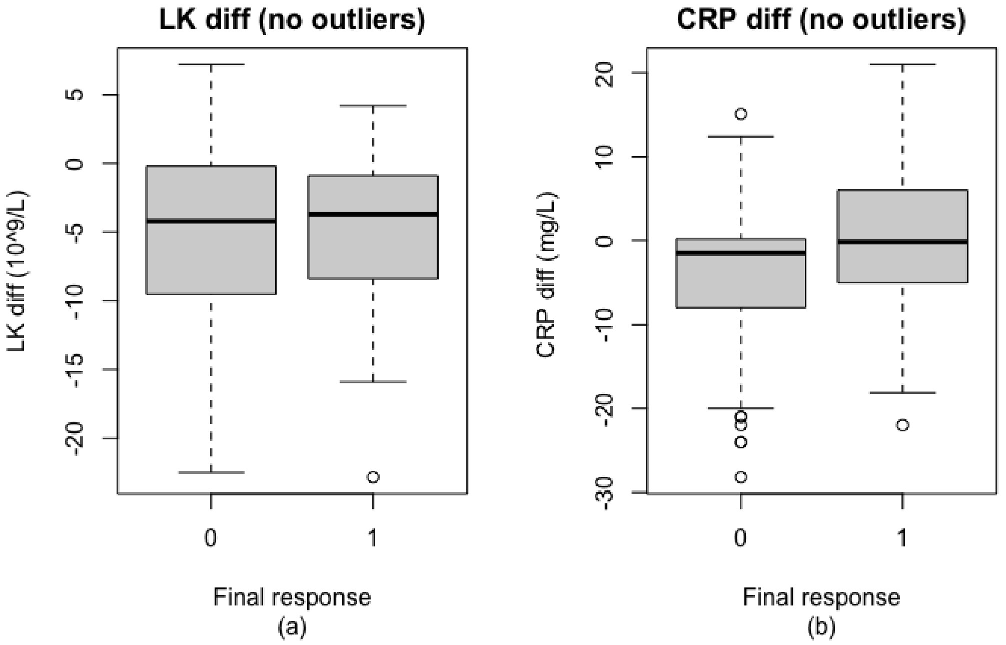


Figure B.305: Boxplot of the (a) Leukocytes and (b) C-Reactive Protein difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=81), 1 = Progressive Disease (n=18), total n=99)

NLR difference		**PLR difference**	
Patient ID	*Value*	*Patient ID*	*Value*
001PANC0015	-33.0	001PANC0036	671.7
001PANC0035	-15.1	151PP20002	-127.1
001PANC0036	12.5	151PP20004	-585.9
001PANC0023	-16.9	001PANC0051	242.6
001PANC0024	-9.6	001PP20032	330.3
001PANC0026	-8.6	133PP20007	2280
001PANC0011	-14.2	151PP20029	-137.2
001PP20007	-10.6		
151PP20004	-30.4		
001PP20032	6.6		

Table B.95: Overview of the outlier values determined using the IQR-method for the Neutrophil-to-Lymphocyte Ratio and Platelet-to-Lymphocyte Ratio difference values before and after the first chemotherapy cycle.

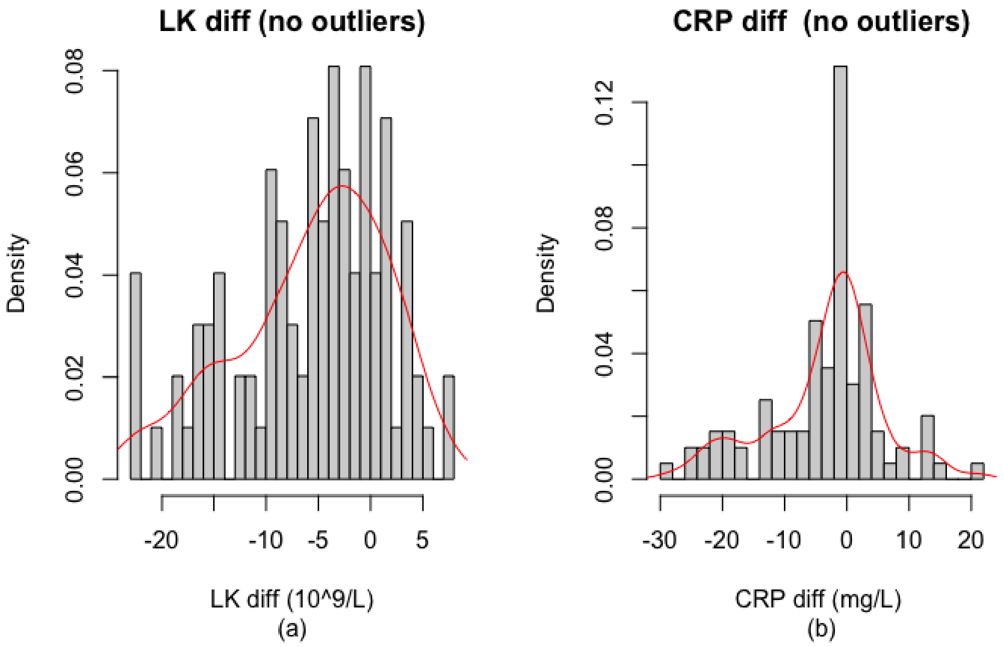


Figure B.306: Histogram of the (a) Leukocytes and (b) C-Reactive Protein difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=81), 1 = Progressive Disease (n=18), total n=99)

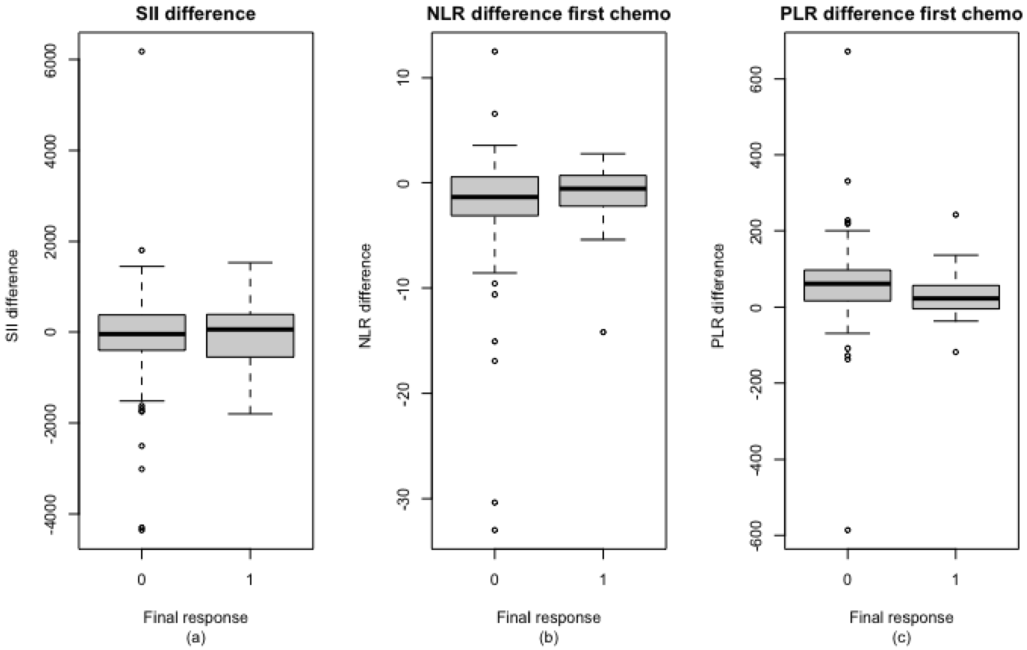


Figure B.307: Boxplot of the (a) Systemic inflammation index, (b) Neutrophil-to-Lymphocyte Ratio and (c) Platelet-to-Lymphocyte Ratio difference values before and after the first chemotherapy cycle categorised by final response. (0 = Disease control (n=86), 1 = Progressive Disease (n=24), total n=110)

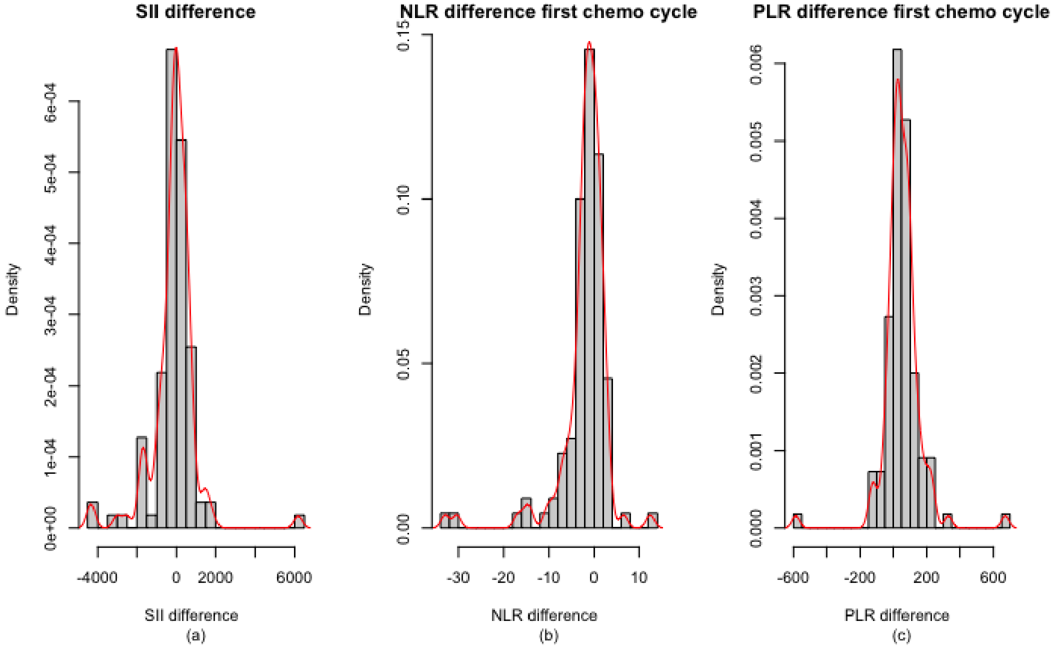


Figure B.308: Histogram of the (a) Systemic inflammation index, (b) Neutrophil-to-Lymphocyte Ratio and (c) Platelet-to-Lymphocyte Ratio difference values before and after the first chemotherapy cycle categorised by final response with a fitted density estimate in red. (0 = Disease control (n=86), 1 = Progressive Disease (n=24), total n=110)

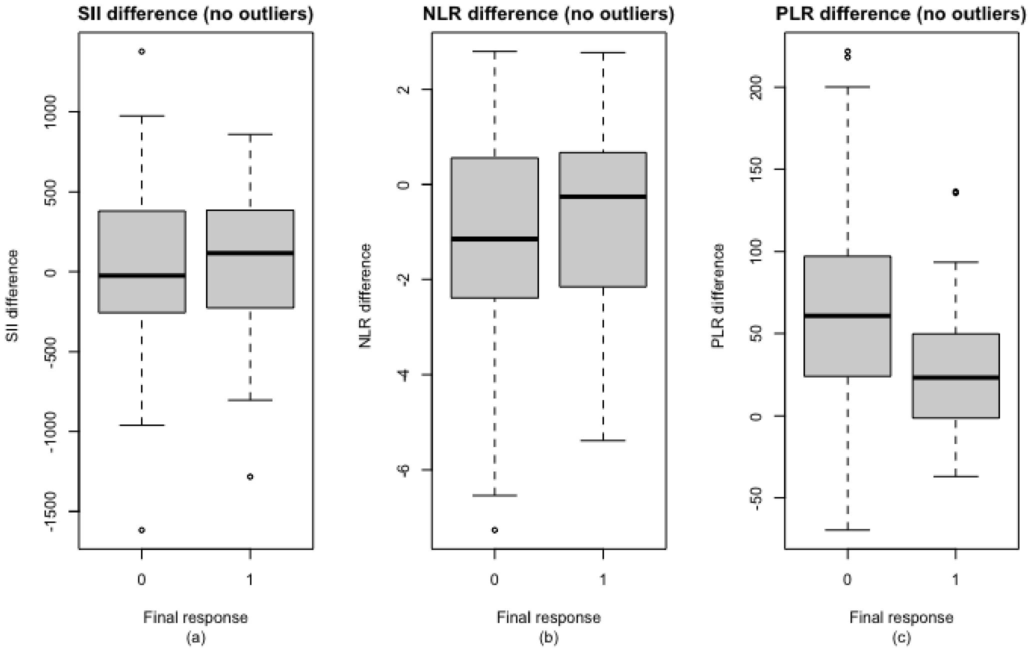


Figure B.309: Boxplot of the (a) Systemic inflammation index, (b) Neutrophil-to-Lymphocyte Ratio and (c) Platelet-to-Lymphocyte Ratio difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method. (0 = Disease control (n=70), 1 = Progressive Disease (n=21), total n=91)

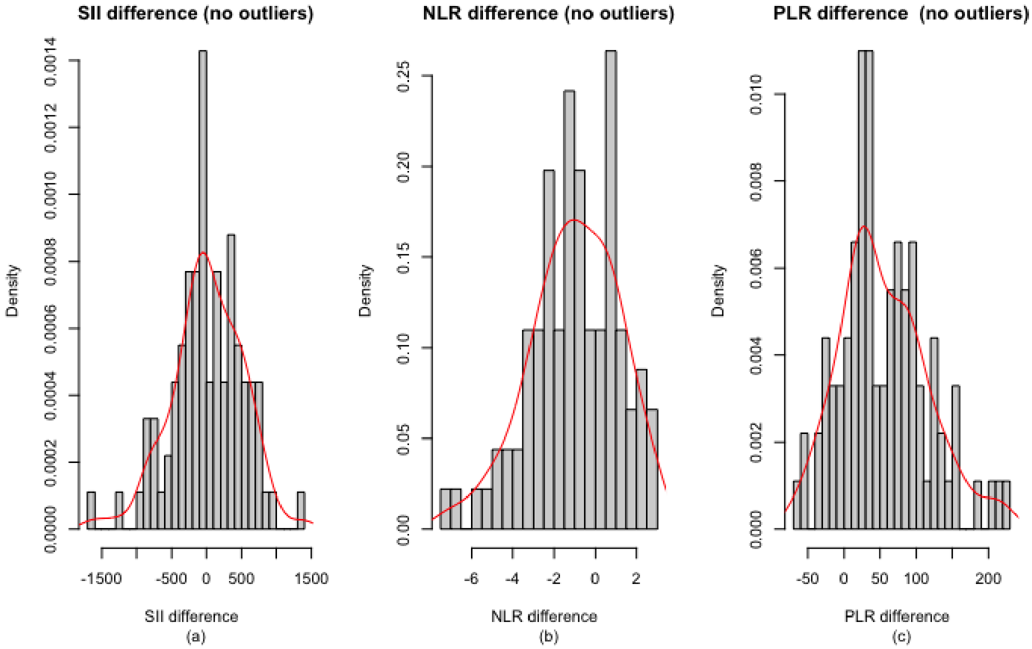


Figure B.310: Histogram of the (a) Systemic inflammation index, (b) Neutrophil-to-Lymphocyte Ratio and (c) Platelet-to-Lymphocyte Ratio difference values before and after the first chemotherapy cycle categorised by final response after removal of outliers based on the IQR-method with a fitted density estimate in red. (0 = Disease control (n=70), 1 = Progressive Disease (n=21), total n=91)

B.2.4 Overview of number of outliers per patient

Patient ID	Number of Outliers	Outlier Variables
001PANC0004	20	CA199diag, CA199before, CA199after, CA199diff, CEAbefore, CEAafter, TBafter, TBdiff, ASATafter, ASATdiff, ALATafter, ALATdiff, AFafter, AFdiff, GGTafter, GGTdiff, CRPbefore, SIIafter, SIIdiff, PLRafter
001PANC0011	18	CA199before, CA199diff, CEAdiff, CRdiff, GFRbefore, GFRafter, GFRdiff, Naafter, Kdiff, AFbefore, AFafter, BRafter, CRPbefore, CRPafter, CRPdiff, SIIdiff, NLRafter, NLRdiff
001PP20032	15	CRdiff, GFRdiff, ASATbefore, ASATdiff, ALATbefore, ALATdiff, AFdiff, GGTbefore, GGTdiff, CRPbefore, CRPafter, NLRbefore, NLRdiff, PLRbefore, PLRdiff
001PANC0025	13	CA199diag, CA199before, CA199after, CA199diff, CEAdiff, ASATbefore, ASATdiff, ALATbefore, ALATdiff, GGTbefore, GGTdiff, BRbefore, BRdiff
151PP20004	13	GFRdiff, AFbefore, AFdiff, GGTbefore, GGTdiff, CRPbefore, CRPdiff, SIIafter, SIIdiff, NLRafter, NLRdiff, PLRafter, PLRdiff
133PP20005	13	NPbefore, ASATbefore, ASATdiff, ALATbefore, ALATafter, ALATdiff, AFbefore, AFafter, AFdiff, GGTbefore, GGTafter, GGTdiff, BRdiff
001PANC0015	12	CA199diff, CEAdiff, CRdiff, GFRafter, GFRdiff, Kafter, Kdiff, GGTdiff, SIIafter, SIIdiff, NLRafter, NLRdiff
001PANC0035	12	CA199diag, CA199before, CA199after, CA199diff, CEAdiag, CEAbefore, CEAafter, CEAdiff, GGTdiff, SIIdiff, NLRafter, NLRdiff
001PANC0002	11	CA199diag, CA199before, CA199after, CA199diff, CEAdiag, CEAbefore, CEAafter, GFRbefore, Nabefore, CRPbefore, PLRbefore
001PANC0033	11	CA199diff, CEAbefore, CEAafter, CEAdiff, AFbefore, AFafter, AFdiff, GGTbefore, GGTafter, GGTdiff, CRPbefore
148PP20013	11	LKbefore, CRdiff, ASATbefore, ASATdiff, ALATbefore, ALATdiff, AFbefore, BRbefore, BRdiff, CRPafter, CRPdiff
148PP20019	11	LKbefore, NPbefore, LCbefore, LCafter, LCdiff, ASATafter, ASATdiff, ALATafter, ALATdiff, GGTafter, GGTdiff
001PANC0032	10	CEAbefore, CEAafter, CEAdiff, ASATafter, ASATdiff, ALATafter, ALATdiff, AFafter, GGTafter, GGTdiff
001PANC0013	10	CEAdiff, GFRbefore, ASATbefore, ASATdiff, ALATbefore, ALATdiff, GGTbefore, GGTdiff, BRbefore, BRdiff

Patient ID	**Number of Outliers**	**Outlier Variables**
001PP20022	10	CA199diag, CA199before, CA199after, CA199diff, CEAdiag, CEAbefore, CEAafter, CEAdiff, GFRbefore, GFRafter
001PANC0052	10	LKbefore, GFRbefore, GFRafter, ASATbefore, ASATdiff, ALATbefore, ALATdiff, GGTbefore, GGTdiff, BRdiff
078PP20016	10	TBdiff, ASATbefore, ASATdiff, ALATbefore, ALATdiff, AFbefore, AFafter, AFdiff, GGTbefore, GGTdiff
001PANC0061	10	LKbefore, NPbefore, Kdiff, GGTbefore, GGTafter, CRPbefore, CRPafter, CRPdiff, SIIbefore, NLRbefore
078PP20040	10	CA199before, CA199after, ASATafter, ALATafter, AFbefore, AFafter, AFdiff, GGTbefore, GGTafter, GGTdiff
001PANC0023	9	CEAdiff, LKdiff, NPdiff, GFRbefore, GFRafter, SIIafter, SIIdiff, NLRafter, NLRdiff
001PANC0019	9	CA199diag, CA199before, CA199after, CA199diff, CEAdiag, CEAbefore, CEAafter, CEAdiff, ASATafter
001PP20017	9	CEAdiag, CEAbefore, CEAafter, LKafter, LKdiff, NPafter, NPdiff, LCbefore, LCafter
001PANC0039	9	LKbefore, NPbefore, CRdiff, GFRdiff, GGTafter, GGTdiff, Albafter, CRPbefore, CRPdiff
001PANC0036	8	CRPbefore, CRPafter, SIIbefore, SIIdiff, NLRbefore, NLRdiff, PLRbefore, PLRdiff
001PANC0014	8	LKafter, LKdiff, NPafter, NPdiff, LCafter, LCdiff, AFafter, AFdiff
148PP20006	8	NPdiff, ASATafter, AFbefore, AFafter, AFdiff, GGTbefore, GGTafter, GGTdiff
001PANC0050	8	CA199diff, CEAdiff, CRdiff, GFRdiff, Naafter, Albdiff, CRPafter, CRPdiff
151PP20013	8	TBdiff, CRbefore, GFRbefore, GFRafter, GFRdiff, SIIafter, SIIdiff, PLRafter
001PP20045	8	CA199diff, CEAdiag, CEAbefore, CEAafter, CEAdiff, LCbefore, CRdiff, Kdiff
001PANC0028	7	GFRdiff, AFbefore, AFafter, GGTbefore, GGTafter, CRPbefore, CRPdiff
001PANC0009	7	CA199diag, CA199before, CA199after, CA199diff, CEAdiff, GGTafter, GGTdiff

Patient ID	**Number of Outliers**	**Outlier Variables**
001PANC0016	7	CEAdiag, CEAbefore, CEAafter, AFbefore, AFafter, CRPbefore, CRPafter
001PANC0018	7	LKbefore, NPbefore, AFbefore, AFdiff, GGTbefore, GGTafter, GGTdiff
001PP20007	7	CA199diff, GFRdiff, ASATafter, ASATdiff, ALATafter, NLRafter, NLRdiff
151PP20009	7	CRdiff, GFRbefore, GFRdiff, ASATbefore, ASATdiff, ALATbefore, ALATdiff
001PANC0051	7	TBbefore, TBdiff, GFRbefore, GFRafter, Kdiff, PLRbefore, PLRdiff
001PANC0043	7	GFRbefore, GFRafter, Albbefore, Albdiff, SIIbefore, NLRbefore, PLRbefore
001PP20031	7	GFRdiff, ASATafter, ASATdiff, ALATafter, ALATdiff, GGTafter, GGTdiff
078PP20026	7	CRafter, CRdiff, GFRafter, GFRdiff, BRafter, CRPafter, CRPdiff
078PP20028	7	CEAdiag, CEAdiff, HBafter, HBdiff, CRdiff, GFRafter, GFRdiff
133PP20001	6	ASATbefore, ASATafter, ASATdiff, ALATbefore, ALATdiff, BRbefore, BRdiff
133PP20007	6	TBbefore, TBdiff, ALATbefore, ALATafter, AFbefore, AFafter, PLRdiff
151PP20017	6	TBbefore, GFRdiff, ASATbefore, ASATdiff, ALATbefore, ALATdiff, PLRafter
001PANC0001	6	ASATbefore, ASATdiff, ALATbefore, ALATdiff, GGTbefore, GGTafter
001PANC0029	6	LCdiff, Naafter, ASATafter, ASATdiff, ALATafter, ALATdiff
078PANC0002	6	TBafter, CRPafter, CRPdiff, SIIbefore, SIIafter, PLRafter
078PP20007	6	CA199diff, HBbefore, LKbefore, CRdiff, GFRdiff, GGTdiff
001PANC0058	6	CEAdiff, CRbefore, CRafter, CRdiff, GFRbefore, GFRafter
002PANC0013	6	ASATbefore, ASATafter, ASATdiff, ALATbefore, ALATafter, ALATdiff
151PP20034	6	LCbefore, LCdiff, CRdiff, GFRafter, GFRdiff, GGTdiff
151PP20028	6	ASATbefore, ASATdiff, ALATbefore, ALATdiff, BRbefore, BRdiff

Patient ID	**Number of Outliers**	**Outlier Variables**
002PANC0013	6	ASATbefore, ASATafter, ASATdiff, ALATbefore, ALATafter, ALATdiff
151PP20034	6	LCbefore, LCdiff, CRdiff, GFRafter, GFRdiff, GGTdiff
151PP20028	6	ASATbefore, ASATdiff, ALATbefore, ALATdiff, BRbefore, BRdiff
001PANC0037	5	CA199diag, CA199before, CA199after, CA199diff, Kdiff
001PANC0010	5	CA199diff, GFRbefore, GFRafter, GGTbefore, GGTafter
065PP20005	5	CEAdiff, TBbefore, TBdiff, CRPafter, CRPdiff
001PP20036	5	HBdiff, TBbefore, TBafter, LKbefore, NPbefore
002PP20047	5	ALATbefore, ALATdiff, AFbefore, GGTbefore, GGTdiff
002PP20063	5	GFRafter, GFRdiff, ASATafter, ALATbefore, ALATafter
005PP20004	5	CA199diff, TBdiff, GGTbefore, GGTafter, CRPdiff
020PP20006	5	CA199diff, Naafter, ALATdiff, GGTbefore, GGTdiff
133PP20006	5	GFRafter, GFRdiff, ASATbefore, AFdiff, BRdiff
148PANC0004	5	NPafter, ASATbefore, ASATdiff, ALATafter, ALATdiff
165PP20006	5	TBafter, TBdiff, Naafter, ALATafter, ALATdiff
148PANC0009	5	ASATbefore, ASATdiff, ALATbefore, ALATdiff, AFbefore
001PP20008	4	CRafter, CRdiff, GFRafter, GFRdiff
002PANC0007	4	CRbefore, CRafter, GFRbefore, GFRafter
002PP20011	4	CEAdiff, TBbefore, GFRdiff, SIIbefore
002PP20005	4	TBbefore, TBdiff, ALATbefore, GGTdiff
001PANC0038	4	CA199diff, TBdiff, GFRbefore, GFRdiff
148PANC0003	4	LKbefore, LCbefore, LCafter, ASATafter
148PP20015	4	GGTdiff, CRPbefore, CRPafter, CRPdiff
165PP20012	4	BRafter, BRdiff, Albafter, Albdiff
001PP20040	4	CRafter, CRdiff, GFRbefore, GFRdiff

Patient ID	**Number of Outliers**	**Outlier Variables**
078PP20038	4	CEAdiff, GFRdiff, ALATdiff, CRPdiff
165PP20018	4	HBdiff, TBafter, TBdiff, Naafter
001PANC0024	3	GFRdiff, NLRafter, NLRdiff
001PANC0026	3	CRPdiff, NLRafter, NLRdiff
001PP20002	3	ALATbefore, ALATdiff, GGTdiff
001PP20013	3	GFRbefore, GGTafter, CRPbefore
001PP20012	3	GGTdiff, CRPbefore, CRPdiff
002PP20008	3	TBafter, TBdiff, GFRdiff
059PP20001	3	TBafter, TBdiff, GFRdiff
078PANC0001	3	CEAdiag, CEAbefore, CEAdiff
078PP20004	3	LKafter, LKdiff, NPafter
151PP20002	3	GGTdiff, PLRafter, PLRdiff
001PANC0047	3	LCbefore, GFRbefore, Naafter
002PP20046	3	ALATdiff, SIIbefore, SIIdiff
078PP20029	3	GFRbefore, ALATafter, ALATdiff
148PP20027	3	ASATbefore, ALATbefore, ALATdiff
151PP20011	3	CA199diag, CA199diff, PLRafter
165PP20010	3	CA199diag, CA199before, CA199after
001PP20048	3	CA199diff, CRPbefore, CRPafter
002PP20069	3	CRbefore, GFRbefore, GFRafter
005PP20020	3	CA199diff, CRPbefore, CRPdiff
001PANC0017	2	GFRdiff, NLRafter
001PANC0003	2	GFRbefore, GFRafter
001PANC0008	2	SIIafter, SIIdiff

Patient ID	**Number of Outliers**	**Outlier Variables**
001PP20015	2	CRPbefore, SIIdiff
018PP20003	2	ASATafter, ASATdiff
065PP20003	2	GFRbefore, GFRafter
148PP20002	2	ALATafter, ALATdiff
148PP20009	2	ALATbefore, ALATdiff
001PANC0042	2	GFRbefore, GFRafter
001PP20029	2	GFRbefore, GFRafter
002PP20051	2	ASATbefore, ASATdiff
002PP20037	2	GFRdiff, Albafter
065PP20015	2	GFRdiff, ALATdiff
148PANC0008	2	ASATafter, ASATdiff
151PP20018	2	TBafter, TBdiff
151PP20021	2	GFRbefore, GFRdiff
001PP20046	2	GFRbefore, GFRafter
005PP20016	2	ASATbefore, ASATafter
005PP20010	2	CA199diff, PLRafter
065PP20017	2	TBdiff, GFRdiff
151PP20029	2	PLRafter, PLRdiff
165PP20014	2	TBafter, TBdiff
001PANC0012	1	CA199diff
001PANC0005	1	PLRbefore
001PP20009	1	SIIafter
002PANC0002	1	GFRdiff
002PANC0005	1	ALATbefore

Patient ID	**Number of Outliers**	**Outlier Variables**
002PP20024	1	ALATbefore
002PP20007	1	GFRbefore
002PP20002	1	BRafter
018PP20004	1	TBafter
078PP20005	1	CEAdiff
078PP20008	1	CEAdiff
078PP20006	1	GFRdiff
165PP20003	1	CA199diff
001PANC0053	1	GFRdiff
001PANC0055	1	GGTdiff
001PANC0045	1	ALATdiff
001PANC0041	1	GFRdiff
001PP20030	1	GFRdiff
002PP20048	1	GFRafter
002PP20055	1	ALATafter
065PP20014	1	CA199diff
078PANC0003	1	ALATdiff
148PP20025	1	TBdiff
148PP20022	1	NLRbefore
165PP20008	1	PLRafter
165PP20005	1	BRdiff
001PANC0062	1	CA199diff
005PP20021	1	GFRafter
078PP20039	1	NPafter
133PP20008	1	HBdiff
133PP20009	1	BRafter
151PP20025	1	GFRdiff
166PP20006	1	Albbefore

Table B.96: Number of outliers per patients with their corresponding outlier variables sorted from the most outliers to the least.

B.3. PCA (background further)

In order to understand some more mathematical concepts and the interpretation of the plots in more detail, some more explanation on the generated plots is provided in this section.

B.3.1 Correlation vs Covariance

Prior to applying PCA, it is important to have an understanding of the concepts of correlation and covariance. These measures both relate to the linear relationship between two variables, but differ in how they normalize the data and the range of their values.

B.3.1.1 Covariance

Covariance quantifies the joint variability of two random variables and indicates the extend to which they 'co-vary'. Covariance can take on positive or negative values depending on the direction of the relationship. However, its magnitude is not standardized and is influenced by the scales of the variables. A positive covariance implies that both variables tend to exhibit similar high or low values simultaneously, while a negative covariance suggests an inverse relationship, wherein one variable tends to be high while the other tends to be low. Mathematically, covariance is defined as the expected value of the product of the deviations of the two variables from their respective means. It represents the average of the product of the differences between each value and its mean:

$$Cov(X,Y) = \mathbb{E}\Big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\Big). \tag{B.7}$$

B.3.1.2 Correlation

Correlation is a standardized measure that assesses the linear relationship between two variables, essentially the normalized version of the covariance. It quantifies the strength and direction of the linear relationship between two variables and ranges between -1 and 1. A value of -1 indicates a perfect negative correlation, 0 signifies no correlation, and 1 denotes a perfect positive correlation. Correlation can be calculated as given in Equation (B.8), dividing the covariance by the product of the standard deviations of the two variables.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}, \tag{B.8}$$

where $\rho(X,Y)$ is the correlation between X and Y, $Cov(X,Y)$ is the covariance between X and Y as given in Equation (B.7), $\sigma(X)$ is the standard deviation of X and $\sigma(Y)$ is the standard deviation of Y. When using PCA, the correlation matrix is used instead of the covariance matrix when the variables are measured on different scales. The use of the correlation matrix ensures that the PCs are uncorrelated and have equal variances. However, if all the variables are measured on the same scale, then the covariance matrix is sufficient. In the following subsections both the correlation as well as the covariance matrix are used in the PCA analysis to show the difference in the outcome.

B.3.2 Loadings and Scores in PCA

Two other important concepts that will be discussed in the PCA analysis are loadings and scores.

B.3.2.1 Loadings

Loadings are the coefficients or weights (ϕ_{ij}) that represent the relationship (correlation) between the original variables and the PCs. Each PC has a corresponding set of loadings, and they jointly make up the principal component loading vector. Loadings are calculated by taking the dot product of the standardized data matrix and the eigenvectors of the covariance matrix. Often loadings are represented as a matrix or a table, with rows corresponding to the original variables and columns corresponding to the PCs. The loadings provide information about which variables give the largest contribution to the components and range from -1 to +1. A high absolute value shows that the variable strongly influences the principal component and a value close to 0 indicate that the variable has a small influence on the principal component. The sign of the loading (+ or -) indicate whether a variable and a principal component are positively or negatively correlated. Furthermore, the eigenvectors of the covariance matrix represent the direction of the PCs in the original data space. Each eigenvector corresponds to one principal component, and the order of the eigenvectors is determined by the corresponding eigenvalues, which represent the amount of variance explained by each principal component

$$loading = eigenvector \times \sqrt{eigenvalue} \tag{B.9}$$

The purpose of this calculation is to incorporate the concept of “load", which provides information about the amount of variance. By multiplying the eigenvalue with the eigenvector, which represents the variance of the PCs, each coefficient becomes “loaded" by the amount of variance. Consequently, these coefficients provide information about the co-variability among variables.

In plots, loadings are often depicted as arrows with specific angles and lengths relative to the axes. The angle of the arrow with respect to the principal axis signifies the contribution of a particular variable to the direction of the PCs in which it is influential. The length of the arrow reflects the strength of the contribution of that variable to the corresponding direction. For example, in a 2D plot with PC1 on the x-axis and PC2 on the y-axis, a horizontal arrow represents the variable that contributes predominantly to PC1, while a vertical arrow represents a feature that primarily contributes to PC2. The length of the arrow indicates the magnitude of its contribution. Thus, a longer arrow indicates a better representation of that particular variable in the new principal component space. Additionally, the angles between the loading arrows provide insights into the correlation between variables or features. A small angle indicates a positive correlation, a large angle (between 90 and 180 degrees) indicates a negative correlation, and an angle of 90 degrees denotes no correlation. However, as will be explained later, depending on how well the variables have been projected from the original high dimensional space to this lower dimensional space, this correlation is indeed present or not.

B.3.2.2 Scores

Scores, in the context of PCA, are the projections of the original dataset onto the principal component space. Each observation is associated with a set of scores, which are obtained by multiplying the standardized data matrix by the matrix of loadings. Typically, scores are presented in the form of a matrix or a table, where the rows correspond to the observations and the columns correspond to the PCs. Mathematically, the scores of the PCs represent the coordinates of each observation relative to the axes defined by the PCs. By calculating the dot product of the standardized data matrix and the loading matrix, the coordinates of each observation in the principal component space are determined. These scores provide a representation of how each observation contributes to the PCs and how they are positioned in relation to each other within the new coordinate system.

B.3.3 PCA plots

Before the analysis on the data is given, first a brief explanation on the various plots and measures used is given in this subsection.

B.3.3.1 Screeplot

The screeplot is a graphical representation that illustrates the proportion of variance explained by each PC in order to aid in determining the number of PCs to retain for further analysis. The plot consists of a line or bar chart, where the x-axis represents the PCs in ascending order of magnitude, and the y-axis represents the proportion of variance explained by each PC. The bars or points in the screeplot indicate the amount of variance explained by each PC, with the PC1 on the left side (explaining the highest amount of variation) and the subsequent components following in order. The eigenvalues associated with the PCs are typically used to measure the amount of variation captured by each component. Larger eigenvalues correspond to PCs that explain a higher proportion of the total variance in the dataset. In practice, a screeplot is used to visually identify the “elbow point" or the point in the plot where the curve starts to level off. This point serves as a criterion for determining the number of PCs to retain. Additionally, a predetermined cutoff value (e.g., 70%, 80%, or 90% cumulative variance) may be used to determine this. By examining the screeplot, informed decisions can be made about the number of PCs to include in subsequent analyses, balancing the amount of variance explained with the desired level of dimensionality reduction.

B.3.3.2 Scatter plot

A scatter plot in the principal component space represents the original observations from the dataset based on their scores along the PCs. Each observation is plotted according to its scores, which are the coordinates of the observation in the principal component space. To obtain the scores, the original data matrix is multiplied by the loading matrix. In a 2D scatter plot, the location of each observation is determined by its scores along the PC1 and PC2. In this thesis. observations that are well-represented by the PCs are highlighted in colors such as red or yellow. On the other hand, poorly represented observations, which do not align well with the PCs, are represented in shades of blue. Altogether, a scatter plot provides insights into the grouping, patterns, and relationships among observations in the principal component space, helping to identify clusters or patterns that may not be apparent in the original data space. It serves as a visual tool to explore and interpret the data from a dimensionality-reduced perspective.

B.3.3.3 Loading Plot

A loading plot is a graph that displays the correlation between the original variables and the PCs and helps to understand relationships between the original variables and the PCs. The variables are displayed as arrows or vectors. The direction and the length of these arrows represent the contribution and relative importance of the variable in determining the direction of the PC. Each PC is represented by an axis. In the 2D case, the x-axis will be represented by PC1 and the y-axis by PC2. Variables that are close to the axis of a PC have a high contribution to that PC while variables that are far away have lower contributions. Note that in this case the original variables are projected from a high dimensional feature space onto this 2D space. In order to understand the magnitude of the information preserved a correlation circle can be used. This circle reflects 100% of the variance captured. The closer an arrow is to the circle, the more information is captured and the shorter the arrow, the more information is lost. This is indicated in color as well, the more red the colour, the better the representation of that variable in the principal component space and the more blue the arrow, the smaller the contribution. Note that only for the well-projected variables (the variables pointing close to the correlation circle) the cosine of the angle can be interpreted as the correlation coefficient. On the contrary, when variables are not well-projected the angle might be small in the projected principal component space (e.g. 2D), while the angle in the original space (e.g. 15D) can be very large. Therefore, the cosine of the angle in the original space is nowhere near the cosine of the angle in the projected space. Positively correlated variables are grouped together while negatively correlated variables are positioned on opposite sides of the plot. A more clear explanation with illustration will be provided in detail during the analysis of the data.

B.3.3.4 Biplot

A biplot shows the scores as well as the loadings all in one plot. Just like in the other plots the axes are the PCs and each point represents the transformed observation to the principal component space. The arrows or vectors are the loadings, just as in the loading plot. The interpretation is similar as previously explained.

B.3.3.5 Quality of Representation

Another way to visualize the quality of representation of the original variables as well as the correlation between the original variables and the PCs is by plotting them against each other. The quality of representation is determined using the $\cos^2$, where a high value indicates a good representation of the variable on the PC. Consequently, this variable is positioned close to the circumference of the correlation circle in the loading plot. Conversely, a low $\cos^2$ value indicates a poorly represented variable, which can be seen as having a small arrow in loading plot. The better represented variables are the most important in interpreting the PCs. This is because the variables that are correlated with PC1 and PC2 are the most important in explaining the variability in the dataset as PC1 and PC2 capture the most information. Variables that barely correlate with any PC or correlate in the last dimensions have low contribution and might be removed to simplify the overall analysis.

To be more specific, in PCA, the cos2 plot provides insights into the contribution of each variable to the PCs. It represents the squared cosines of the variables, indicating the proportion of variance in each variable that is captured by a particular principal component. Interpreting the cos2 plot involves examining the values associated with each variable in the PCs. Higher cos2 values indicate that the variable is well-represented by the corresponding principal component, meaning it contributes significantly to that component's variation. Conversely, lower cos2 values suggest that the variable has less influence on the specific principal component. Some important key points to consider when interpreting a cos2 plot are:

- Variable Importance: Variables with higher cos2 values (close to 1) are more important in determining the specific principal component. They contribute more to the overall variability explained by that component.
- Dimension Representation: By examining the cos2 values across different PCs, one can determine which variables have a stronger association with specific dimensions. Variables with high cos2 values in multiple PCs may exhibit multidimensional or have shared relationships with different components.
- Total Variance: The sum of cos2 values for each variable across all PCs provides an indication of the overall contribution of that variable to the total variance explained by the PCA. Variables with higher cumulative cos2 values are more influential in explaining the overall dataset variation.

It is important to note that the interpretation of the cos2 plot should be done in conjunction with other PCA diagnostic tools, such as scree plots, loading plots, and biplots, to gain a comprehensive understanding of the relationship between variables and PCs. Additionally, to highlight the most contributing variables in each dimension a contribution plot is made too, in which a bigger darker coloured dot represents a higher contribution to that principal component dimension. To go into more detail, a contribution plot in PCA provides information about the relative contribution of each observation (data point) to the PCs. It helps identify which observations have the most influence on the formation of the PCs and can reveal outliers or influential cases. When interpreting a contribution plot, consider the following important details:

- Observation Importance: The higher the contribution value of an observation, the more it contributes to the formation of the PCs. Observations with larger contributions have a stronger impact on the overall variability explained by the PCs.
- Outlier Detection: Outliers or influential cases that deviate significantly from the majority of observations will have high contribution values. These observations can be identified as they stand out from the rest in the contribution plot.
- Dimension Representation: By examining the contribution values across different PCs, one can understand which dimensions are primarily influenced by specific observations. Observations with high contributions in multiple PCs may have a more substantial influence on the overall dataset representation.

The contribution plot provides insights into the individual data points' impact on the PCs, allowing for the identification of influential cases and the assessment of their influence on the overall analysis. It complements the cos2 plot, which focuses on the variables' contribution to the PCs, providing a holistic understanding of the relationships between observations and variables in PCA. Together, the cos2 plot and contribution plot offer a comprehensive view of the PCA results, allowing for the identification of key variables and influential observations in the dataset.

B.3.3.6 PCA plot with prediction ellipses

Another informative plot is the PCA plot with 95% confidence prediction ellipses. In this PCA plot the x-axis and y-axis correspond to PC1 and PC2 respectively. In contrast to the previous scatter plots, the scores are now visualized using dots and triangles coloured in blue and red, respectively, to indicate the two final responses (DC and PD). The coloured ellipses represent the prediction regions around the mean scores for each group. These ellipses can be interpreted as follows: with 95% confidence, the predicted average value of an observation categorized by final response will fall within the shaded area of the ellipse for that specific category.

To determine the 95% prediction ellipses, various distributions can be taken. In this case the student-t distribution is used. The t-distribution is often used when the sample size is small (in this case a sample size of 100-250 is regarded as small in statistics) and the standard deviation is unknown. It is a probability distribution that is similar to the normal distribution but has heavier tails, which means that it assigns more probability to extreme values. This often happens in the case outliers or rare events, which in turn can heavily influence the analysis and assigning more probability to extreme values means that these are more likely to occur. To interpret the ellipse plots it is necessary to look at the overlap of the ellipses. When the ellipses in the PCA plot overlap, it indicates that the groups represented (in this case the final dichotomized response) have a similar variance-covariance structure. This means that the PCs do not provide much discrimination power to separate those groups and there may not be much difference between them in terms of original variables.

On the other hand, when one ellipse is entirely inside another, it indicates that the group represented by the smaller ellipse has a lower variance-covariance structure than the group represented by the larger ellipse. Thus, the PCs are able to discriminate well between these groups and that there is indeed a significant difference between them in terms of original variables. Additionally, the variance-covariance structure refers to the pattern of variances and covariances among the variables in a dataset. In PCA, the goal is to summarise this structure using a smaller number of variables (PCs), which are linear combinations of the original variables. It determines how much variance each PC captures in the data. Finally, since in many cases it is seen that the first two PCs were not able to make a clear distinguishment between the two final response groups, a scatter matrix plot of the first four PCs is provided. In this plot each PC is plotted against each other and the final response is coloured.

B.3.3.7 3D plots

In addition to the 2D plots, where the variables are only projected onto the first two PCs, a 3D plot is made. In this case the variables are projected onto 3 PCs. These 3D scatter plots represent the scores of the first three PCs. Each point in the scatter plot corresponds to a sample and the coordinates of each point in the three-dimensional space are determined by the values of the first three PCs. In this paper, the colour of each point in the scatter plot represent the final response of the patient to chemotherapy, which is classified as either disease control or progressive disease. Interpreting these plots involves analyzing the relationships between the variables and the scores on the first three PCs. For example, points that cluster together in the scatter plot may indicate groups of patients with similar responses to treatment and the biplot (which is only plotted in the interactive 3D plot) can be used to identify which variables are most strongly associated with these groupings. Additionally, the biplot can be used to identify variables that have similar relationships with the first three PCs.

B.4. PCA (full analysis)

This section provides the full PCA analysis conducted on the provided dataset as explained in section 3.3. It contains a detailed principle component analysis of the various groups described before. These were:

1. Blood: HB, TB, LK, NP, LC
2. White Blood Cells: LK, NP, LC
3. Kidney: Na, K, CR, GFR
4. Liver: ASAT, ALAT, AF, GGT, BR and INR (INR is not taken into account due to lack of data)
5. Nutrition: Alb, Na, K
6. Inflammation: CRP, LK, SII, NLR, PLR
7. Patient Characteristics: Height, Weight, BMI, Age
8. All measured variables: HB, TB, LK, NP, LC, Na, K, CR, GFR, ASAT, ALAT, AF, GGT, BR, Alb, CRP, LK, SII, NLR, PLR
9. All measured variables and Age and BMI added as well as the difference of the measured variable before and after the first chemotherapy treatment
10. Age, BMI and the differences between the variables measured before and after the first chemotherapy treatment

The final response analyzed is the dichotomized final response, where 0 = Disease Control (CR, PR, SD) and 1 = Progressive Disease (PD).

B.4.1 White Blood Cells

The white blood cell group contains the data before and after the first chemotherapy treatment of the leukocytes (LK), Neutrophils (NP) and lymphocytes (LC) as well as their respective differences before and after the first chemotherapy treatment. It contains a total of 111 observations of 9 explanatory variables. The correlation matrix in Figure B.311 shows that almost all before and after values of the variables show some positive correlation. The LC before and after are very strongly correlated while the NP and LK show milder correlation. However, the differences between the LK and NP are very strongly correlated (correlation of 0.989), indicating that the difference in LK might be almost entirely determined by the difference in NP as NP is a type of LK, this was already observed earlier in the section B.1 in plots Figure 3.2 and Figure 3.3. Other interesting observations are that the differences of the variables show either negative or no correlation with the before and after values of the corresponding variables. For instance NP after is very negatively correlated with LKdiff and NPdiff as well as LKafter and NPafter. PCA is performed on the white blood cell group and the corresponding scree plot can be found in Figure B.312. This plot clearly shows that a bit more than half of the variation in the data is captured by PC1 and the 'elbow' point can be set at either PC3 or PC4. As the first three PCs explain a total of 90.5% of the total variation and if PC4 is included this will rise to 99.7% as given in Table B.97. The scatter plot of the scores of the white blood cell variables show that the well-represented observations tend to lie away from the origin and closer to the principal components (either PC1 or PC2) while the poorly represented observations tend to lie around the origin. Since PC1 explains a bit than half of the total variation, many observations seem to be correlated with PC2. However, there is no clear structure to be found.

The loading plot in Figure B.314 shows that almost all variables are either below the horizontal axis defined by PC1 or just above it. It also shows that NPafter and LKafter seem to be very positively correlated with PC1 and very well represented too. LKdiff and NPdiff show a strong negative correlation with PC1 and are represented almost entirely as well. This can be seen by the vectors pointing almost to the edge of the circumference of the correlation circle drawn. LKbefore is another well-explained variable and shows a stronger negative correlation with PC2. From this plot it can also be seen that as was seen in the correlation matrix Figure B.311 that NPafter, LKafter are strongly positively correlated and LKdiff and NPdiff as well. Furthermore it looks like all the before values (NPbefore, LCbefore and LKbefore) also show a positive correlation. However, since NPbefore and LCbefore are not so well-represented, the argument about the projection angle from a high dimensional space (in this case 9-dimensional) to the lower dimensional space (in this case 2-dimensional) is valid and clear conclusions cannot be drawn. This is seen back in the correlation matrix as the NPbefore and LKbefore values are strongly positively correlated while LCbefore only shows a slight positive correlation. Recall that the cosine of the angle between two variables is equal to the correlation coefficient between the variables if and only if the variables are well-represented.

The biplot illustrated in Figure B.315 shows that many observations tend to be correlated with PC1 and thus with the vectors: LKdiff, LCdiff, NPdiff, NPafter and LKafter. Nevertheless, no clear patterns are evident. The loading matrix displayed in Figure B.316 confirms that LKafter, NPafter, LKdiff, and NPdiff (in absolute values) are the most significant contributors to PC1, while LKbefore, NPbefore, LCbefore, and LCafter (all negative loading scores) mainly determine PC2. Moreover, the correlation plot in Figure B.317 and the cos2 plot in Figure B.318 reveal similar trends, suggesting that the white blood cell group is mainly determined by the LK and NP differences and after the first chemotherapy treatment values. The contribution plot in Figure B.319 confirms this observation and additionally indicates that PC2 is mainly determined by LKbefore, PC3 by NPbefore, and PC4 by PCdiff. Finally, the PCA plot with 95% prediction ellipses in Figure B.320 reveals significant overlap in the two dichotomized final response groups, suggesting that considering only the white blood cells is insufficient for grouping patients based on their final response in the first two dimensions. The 3D plots in Figures B.322 and B.323 provide a comprehensive visualization of the scores in the first three dimensions.

Finally, the fact that LK and NP differences and after chemotherapy values show the highest correlation in absolute values with PC1 suggests that these variables are the most important in explaining the variance in the white blood cell group. In other words, these variables contribute the most to the variability observed in the data set. This is supported by the fact that PC1 explains a significant portion of the total variance in the data, and that the loading matrix and contribution plot also indicate that these variables are the most significant contributors to PC1. However, it is important to note that there are no clear patterns evident in the biplot or correlation plot, which suggests that the relationships between these variables may be complex and nonlinear. Therefore, it may be necessary to use more advanced analytical techniques to further explore the relationships between these variables and identify any underlying patterns or trends, especially since it is evident that considering the white blood cells solely is insufficient for grouping patients based on their final response.

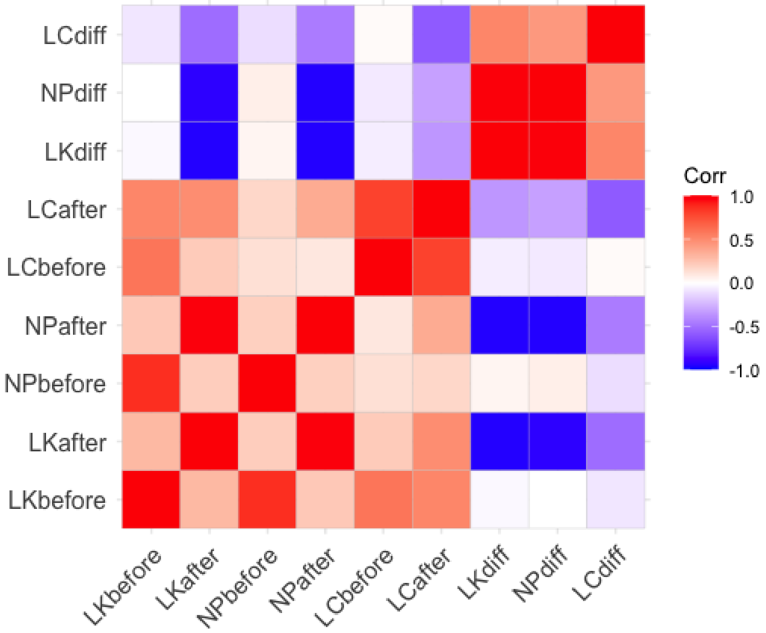


Figure B.311: Correlation Matrix of the various variables in the white blood cell group, n=111 (Disease control (n=87), Progressive disase (n=24)). Red = Positive correlation, Blue = negative correlation, White = No correlation.

WBC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	2.144	1.531	1.098	0.913	0.122
Proportion of Variance Explained	0.511	0.260	0.134	0.093	0.002
Cumulative Proportion	0.511	0.771	0.905	0.997	0.999

Table B.97: PCA summary values of the white blood cell group, n=111 (Disease control (n=87), Progressive disase (n=24)). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components.

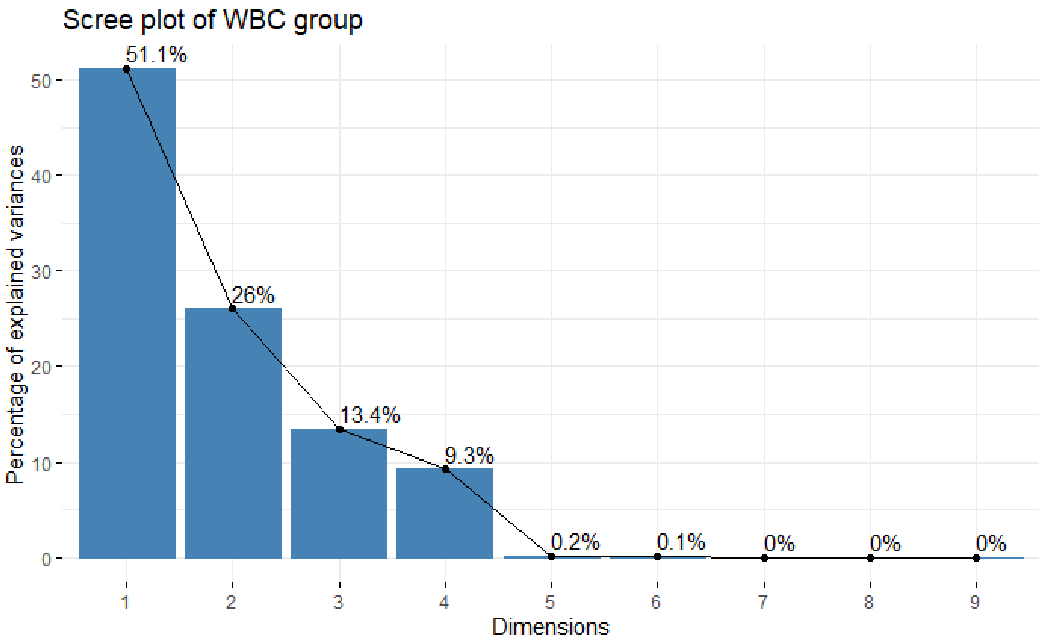


Figure B.312: Scree plot of the white blood cell variables, n=111 (Disease control (n=87), Progressive disase (n=24)).

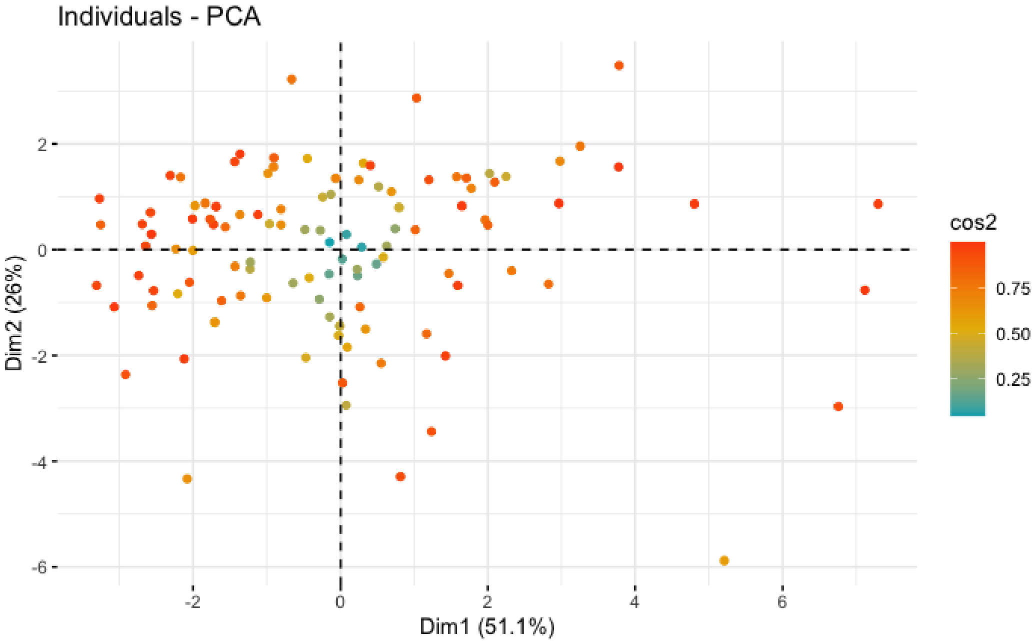


Figure B.313: Scatter plot of the white blood cell variables, n=111 (Disease control (n=87), Progressive disase (n=24)).

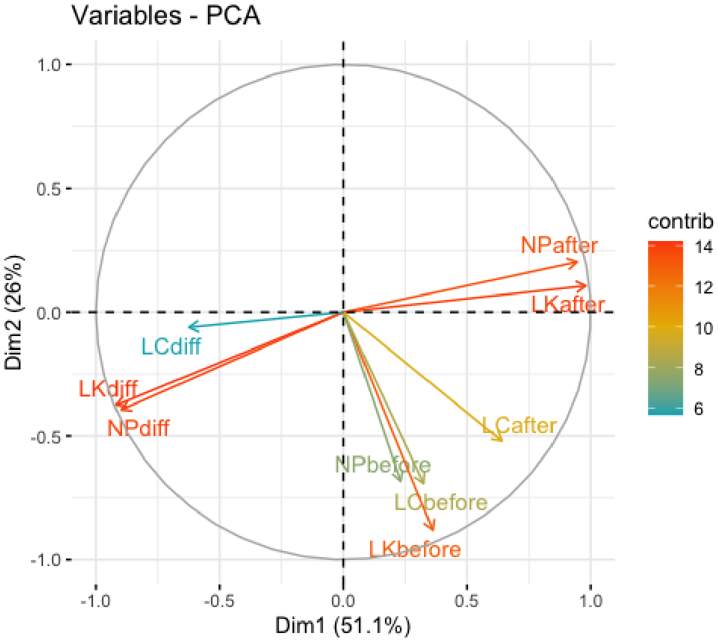


Figure B.314: Loading plot of the white blood cell variables, n=111 (Disease control (n=87), Progressive disase (n=24)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

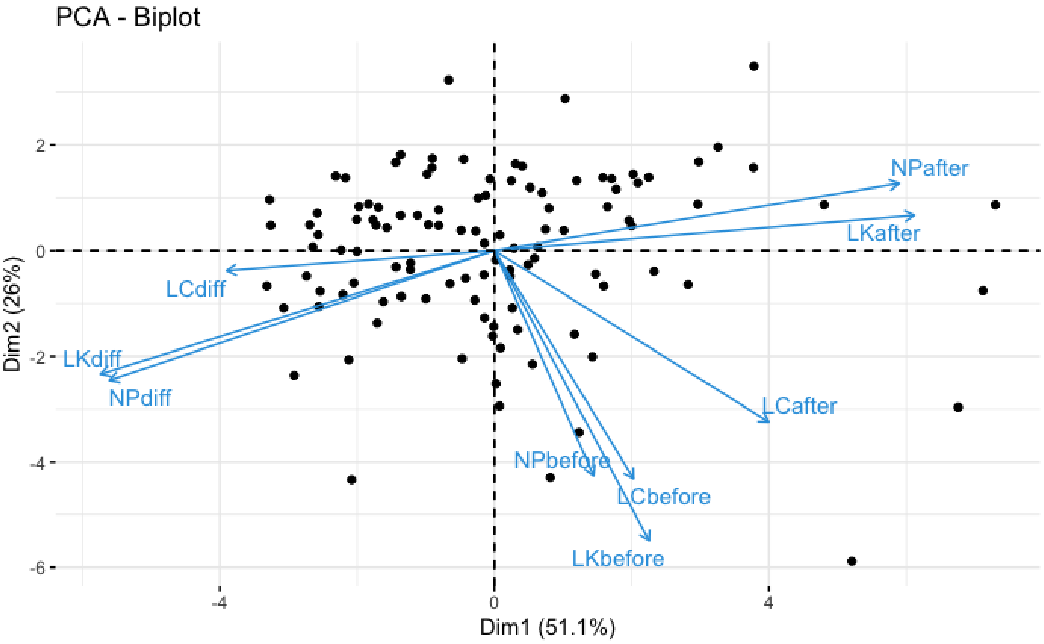
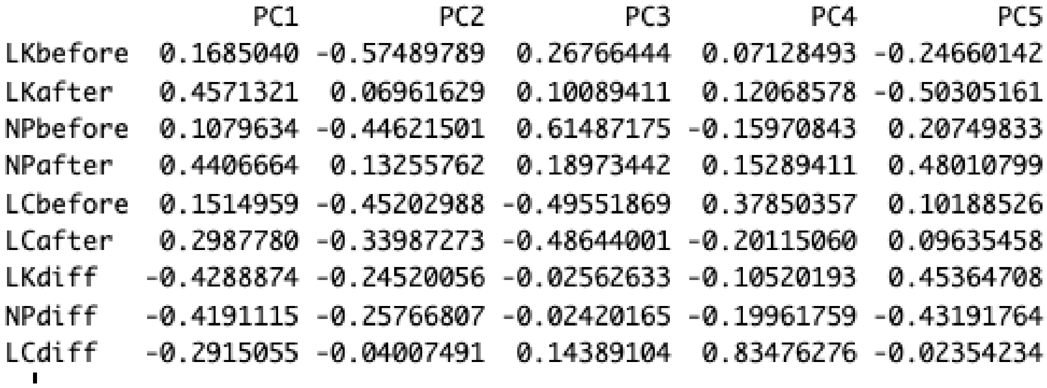


Figure B.315: Biplot plot of the white blood cell variables, n=111 (Disease control (n=87), Progressive disase (n=24)).



```
                PC1         PC2         PC3         PC4         PC5
LKbefore  0.1685040 -0.57489789  0.26766444  0.07128493 -0.24660142
LKafter   0.4571321  0.06961629  0.10089411  0.12068578 -0.50305161
NPbefore  0.1079634 -0.44621501  0.61487175 -0.15970843  0.20749833
NPafter   0.4406664  0.13255762  0.18973442  0.15289411  0.48010799
LCbefore  0.1514959 -0.45202988 -0.49551869  0.37850357  0.10188526
LCafter   0.2987780 -0.33987273 -0.48644001 -0.20115060  0.09635458
LKdiff   -0.4288874 -0.24520056 -0.02562633 -0.10520193  0.45364708
NPdiff   -0.4191115 -0.25766807 -0.02420165 -0.19961759 -0.43191764
LCdiff   -0.2915055 -0.04007491  0.14389104  0.83476276 -0.02354234
```

Figure B.316: Loading matrix of the white blood cell variables of the first 5 principal components, n=111 (Disease control (n=87), Progressive disase (n=24)).

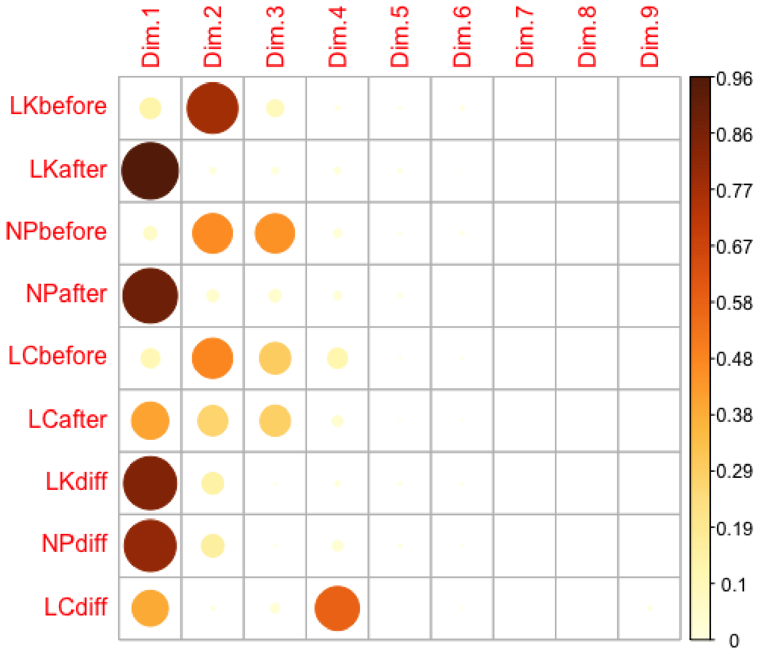


Figure B.317: Correlation of the original white blood cell variables with the principal components, n=111 (Disease control (n=87), Progressive disase (n=24)).

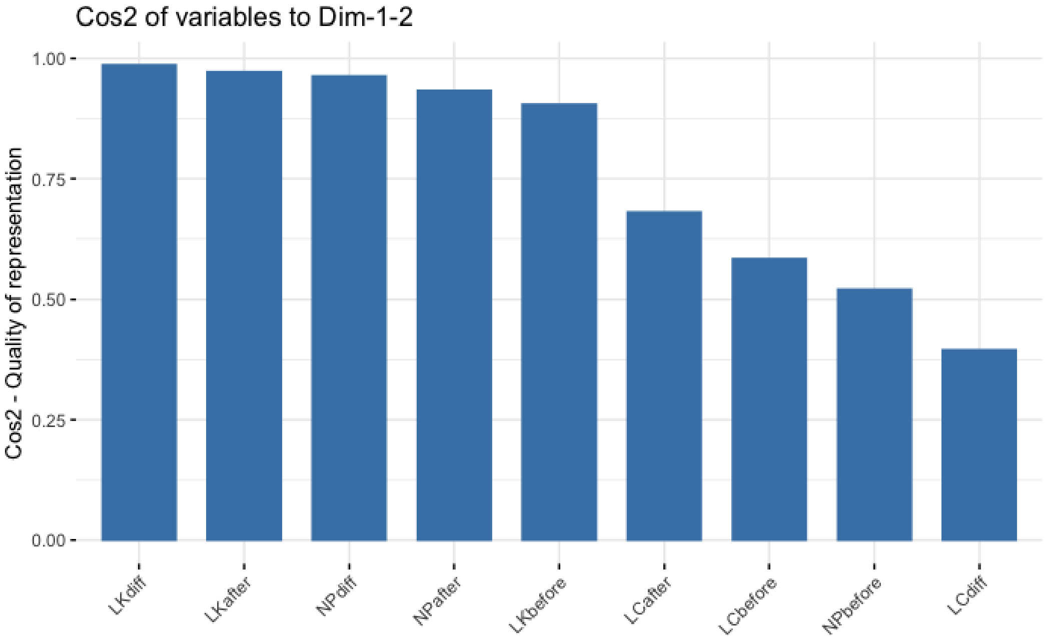


Figure B.318: Cos2 bar chart of the original white blood cell variables, n=111 (Disease control (n=87), Progressive disase (n=24)).

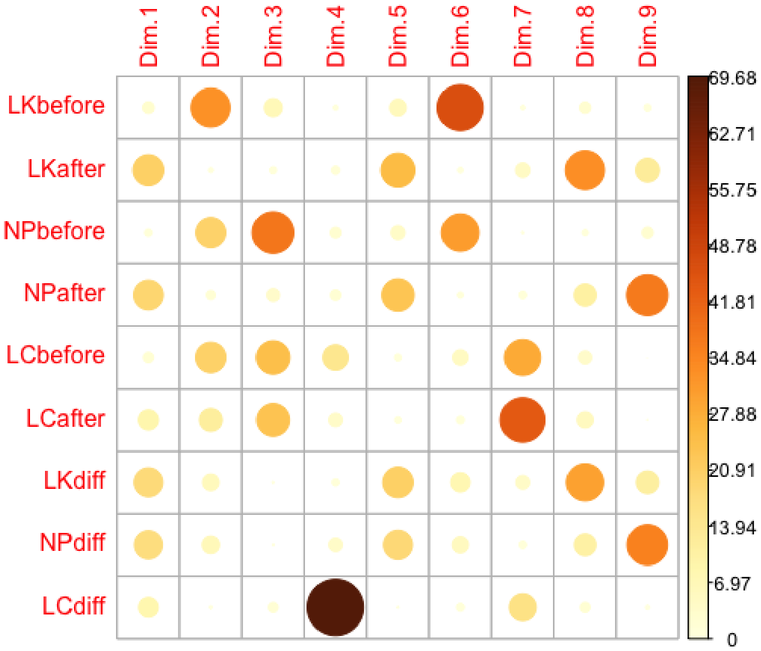


Figure B.319: Contribution of the original white blood cell variables to the principal components, n=111 (Disease control (n=87), Progressive disase (n=24)).

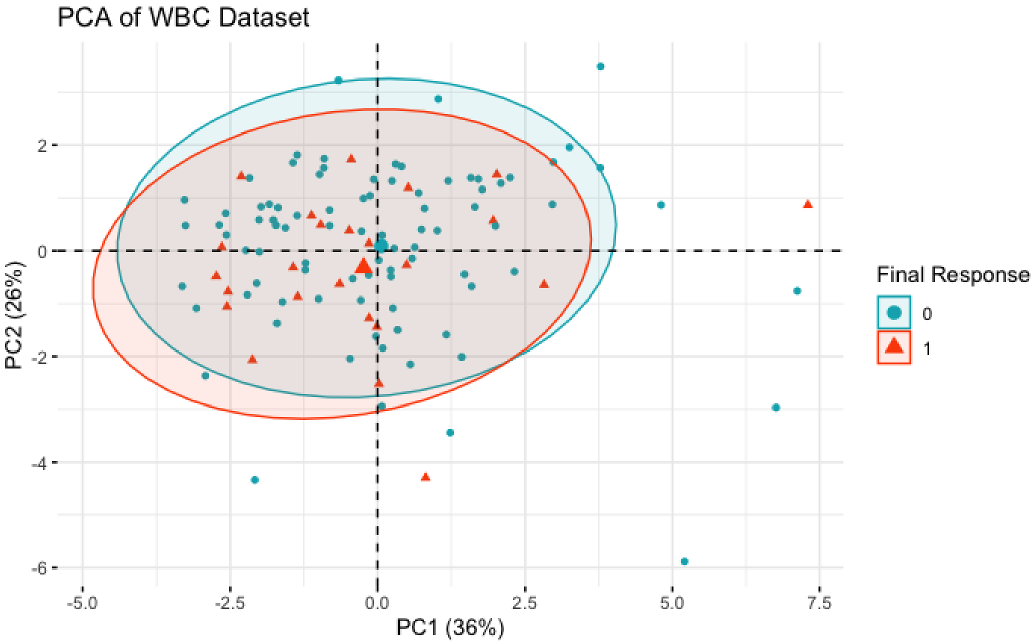


Figure B.320: PCA plot of the white blood cell group with 95% prediction ellipses, n=111 (Disease control (n=87), Progressive disease (n=24)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

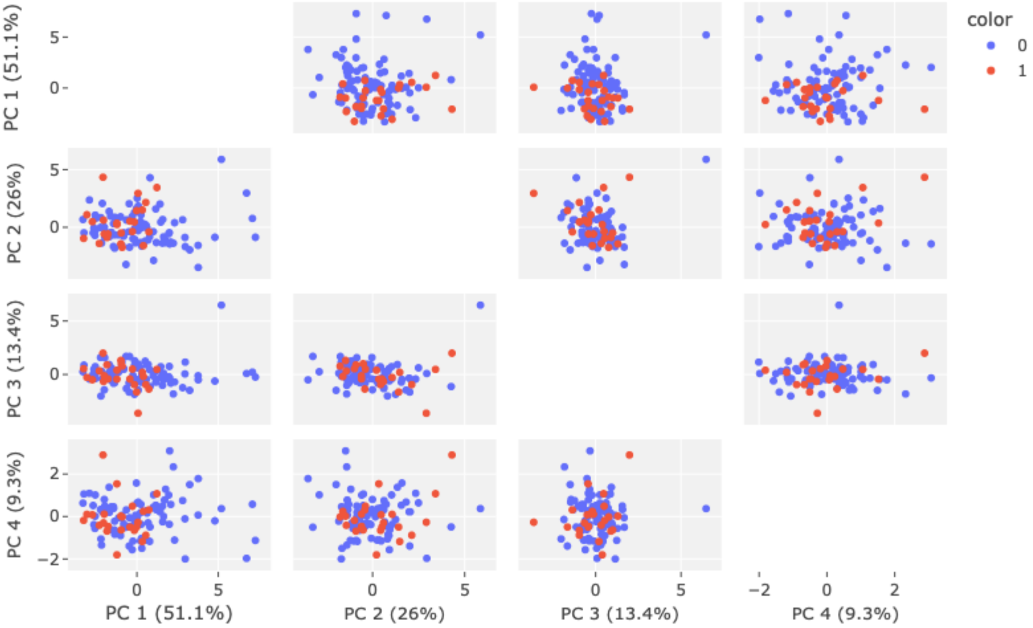


Figure B.321: Scatter matrix plot of the first four principal components, n=111 (Disease control (n=87), Progressive disease (n=24)). The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the white blood cell variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

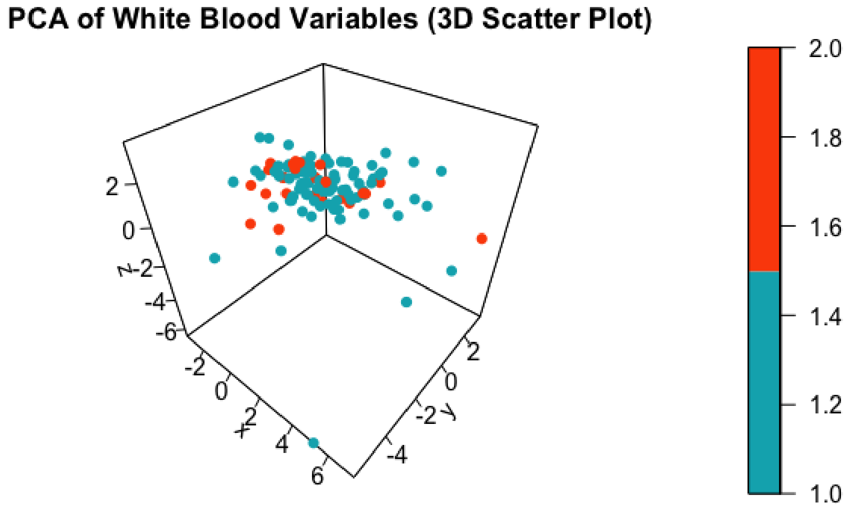


Figure B.322: 3D plot of the white blood cell blood variables projected onto the first three principal components, n=111 (Disease control (n=87), Progressive disase (n=24)).

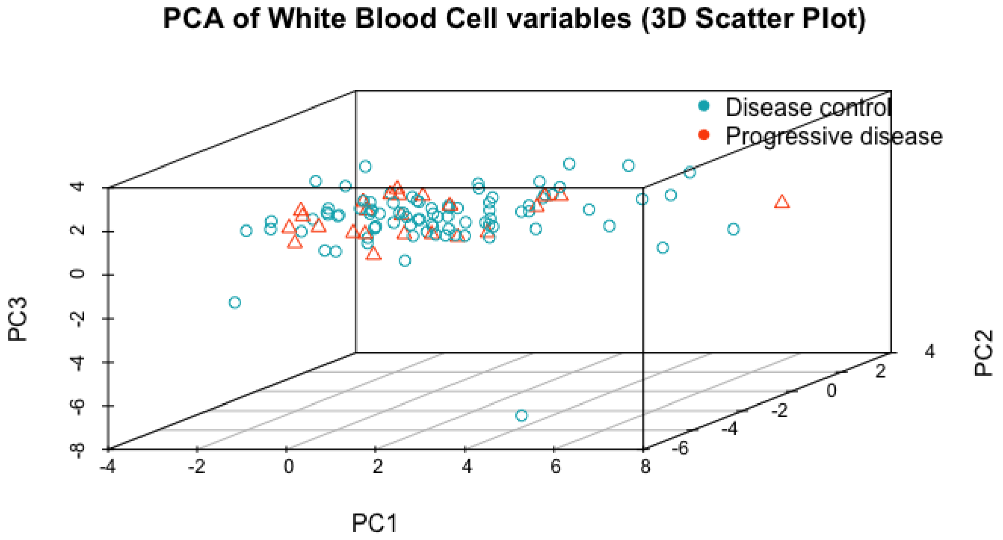


Figure B.323: 3D plot of the white blood cell variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease),, n=111 (Disease control (n=87), Progressive disase (n=24)).

B.4.2 Kidney Function

The kidney function group is composed of four variables, namely Sodium (Na), Potassium (K), Creatinin (CR), and Glomular Filtration Rate (GFR). The kidney dataset in consideration consists of 145 observations and 12 explanatory variables. After rescaling the variables, the correlation matrix displayed in Figure B.324 demonstrates that most variables exhibit no positive correlation, except for the before and after values of each variable respectively (especially the Na, CR and GFR show this). The differences between Na, K, CR, and GFR are clearly uncorrelated. Based on the screeplot presented in Figure B.325 and the values given in Table B.98, it is apparent that the first two principal components together do not account for the majority of the variation in the data. Even with the first five principal components, only 80.5% of the variance in the data can be explained. Given the limited number of explanatory variables, namely 12, this implies that the data may have a complex underlying structure that cannot be easily explained by a small number of variables. Such complexity could stem from various factors, including high dimensionality, non-linear relationships among variables, or the presence of noise in the data. Since it takes at least five principal components to explain a significant proportion of the variance, interpreting the results of the figures in this section, which only consider PC1 and PC2, or using the reduced set of variables for further analysis may prove challenging. Thus, it becomes necessary to explore alternative methods of dimensionality reduction or group the variables differently to simplify the underlying structure.

Despite the limited explanatory power of the first two principal components (PCs) in the kidney function dataset, exploring the 2D plots can reveal potential patterns and trends. The scatterplot in Figure B.326 shows that many observations are poorly represented in the principal component space defined by PC1 and PC2, suggesting weak correlation with either PC1 or PC2. However, the variables that are further away from the center of the observations tend to be better represented and more correlated with either PC1 or PC2. The loading plot in Figure B.327 indicates that the variables in the kidney function dataset point in different directions. In particular, PC1 is shows negative correlation with GFRafter. However, the projection of GFRafter onto PC1 is not well-represented, as indicated by the orange color and arrow being distant from the circumference of the correlation circle. Additionally, GFRbefore and CRbefore are important variables in the first dimension, implying that the crucial variables in the “kidney" group are GFR and CR. Additionally, the differences of all the variables, namely GFRdiff, Nadiff, Kdiff and CRdiff, in this group are poorly represented and do not contribute to the first two dimensions. The biplot in Figure B.328 summarizes the two plots described earlier, revealing no clear relationships between the observations and the variables. The loading matrix in Figure B.329, as well as the correlation, cos2, and contribution of the original variables to the PCs in Figure B.330, Figure B.331, and Figure B.332, respectively, confirm that the first dimension is primarily determined by CRbefore, CRafter, GFRbefore, and GFRafter. These variables exhibit the highest correlation with PC1 and thus contribute the most. The second dimension, PC2, is mainly determined by Nabefore, Naafter, and Kbefore, while the third dimension is mainly defined by Kafter and Kdiff. These findings suggest that potassium values after chemotherapy and the difference are unrelated to the sodium values before and after, defining two different PCs.

Finally, the response plotted in circles and triangles in Figure B.333 indicates that the overall mean is near the origin for both final response groups, and the majority of the ellipses overlap. However, the ellipse for the disease control patients is slightly smaller than the one for the progressive disease. To illustrate the observations in the alternate principal component space, a scatter matrix plot is displayed in Figure B.334. This plot exhibits the observations mapped onto the first four principal components. However, the scatter matrix merely indicates that the progressive disease group is concentrated within the disease control group when projected onto the different principal components. When the principal components are unable to separate the two different groups in any of the dimensions, it suggests that there is a considerable overlap between the two groups in terms of their underlying variables. This means that the variables that define the two groups have a similar pattern of variation, making it difficult to differentiate the groups based on those variables alone. In the context of the analysis, this could mean that the underlying kidney variables that define the disease control group and the progressive disease group are either not significant enough to separate the groups or have a similar pattern of variation. It could also mean that other factors, beyond the measured kidney variables, are contributing to the separation between the two groups. Unfortunately, the 3D plots in Figure B.335 and Figure B.336 do not provide more additional insights in the three dimensional space either.

Kidney	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	1.714	1.375	1.364	1.233	1.205
Proportion of Variance Explained	0.245	0.158	0.155	0.127	0.121
Cumulative Proportion	0.245	0.403	0.558	0.684	0.805

Table B.98: PCA summary values of the kidney group, n=145 (Disease control (n=114), Progressive disease (n=31)). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components.

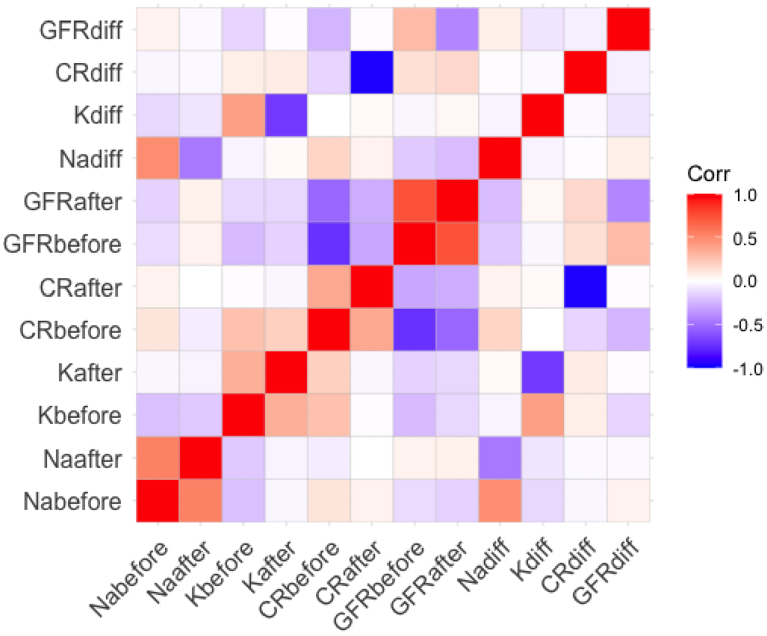


Figure B.324: Correlation Matrix of the various variables in the Kidney group, n=145 (Disease control (n=114), Progressive disease (n=31)). Red = Positive correlation, Blue = negative correlation, White = No correlation.

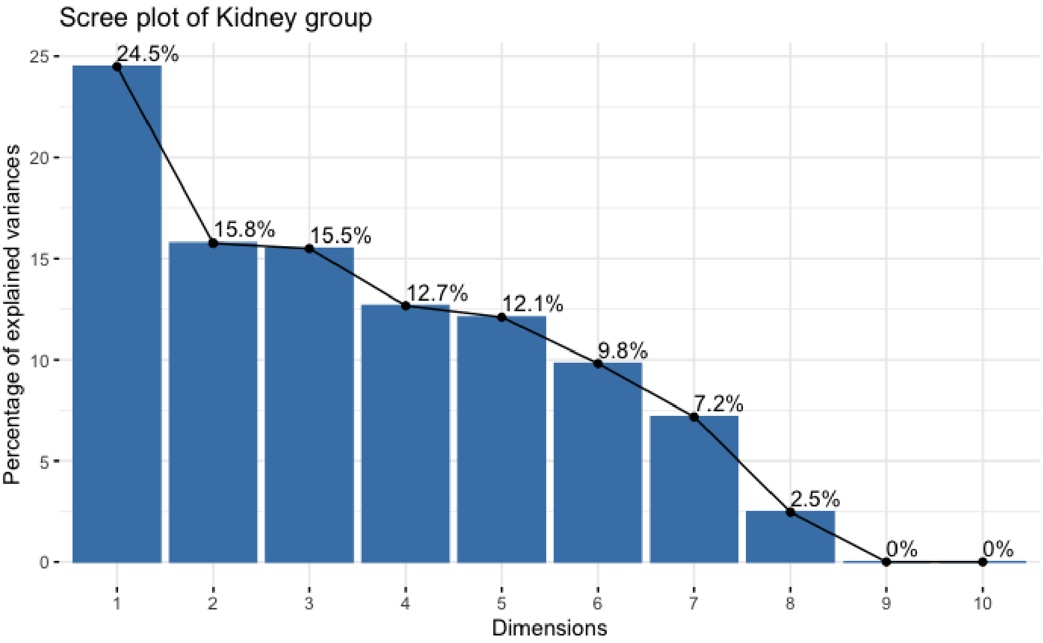


Figure B.325: Scree plot of the kidney variables, n=145 (Disease control (n=114), Progressive disease (n=31)).

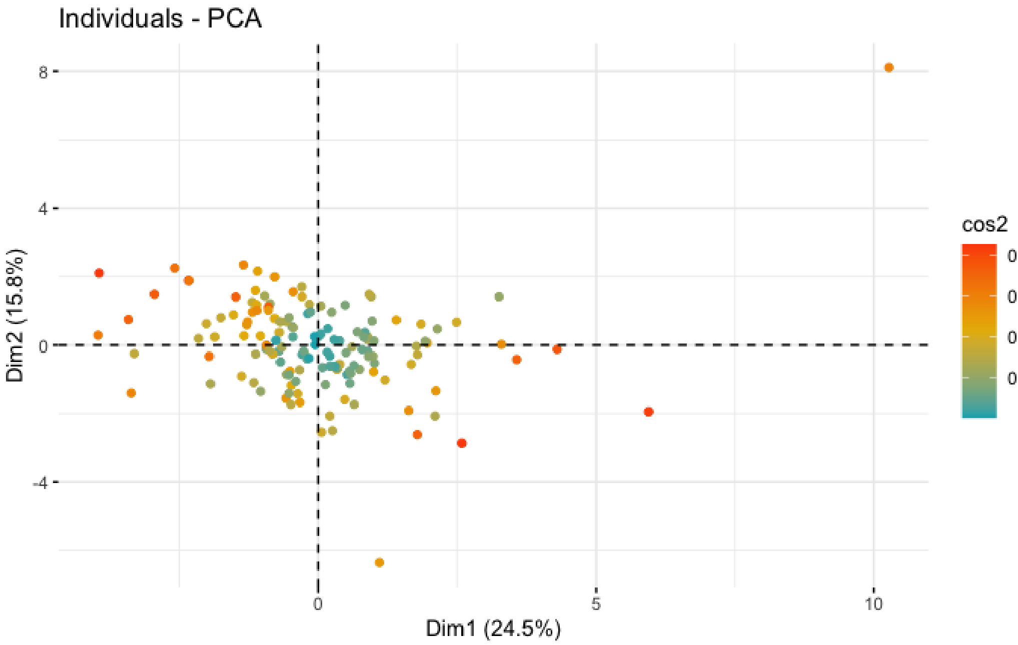


Figure B.326: Scatter plot of the kidney variables, n=145 (Disease control (n=114), Progressive disease (n=31)).

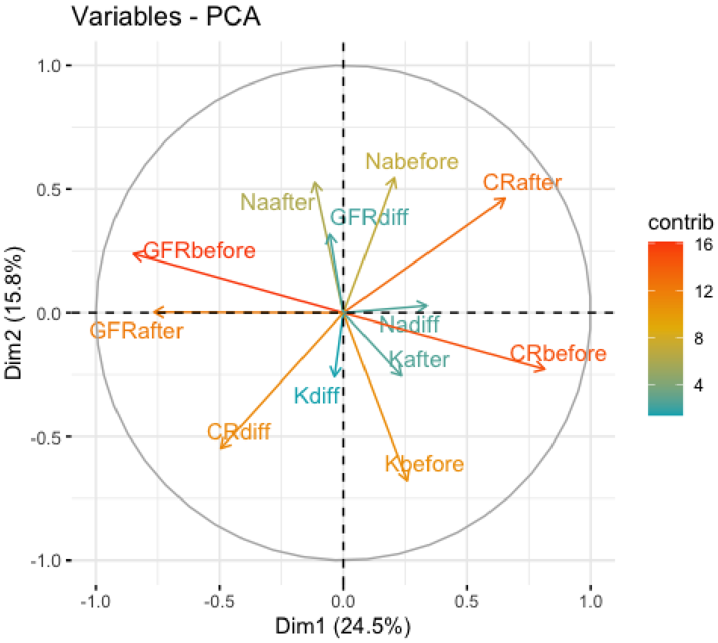


Figure B.327: Loading plot of the kidney variables, n=145 (Disease control (n=114), Progressive disease (n=31)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

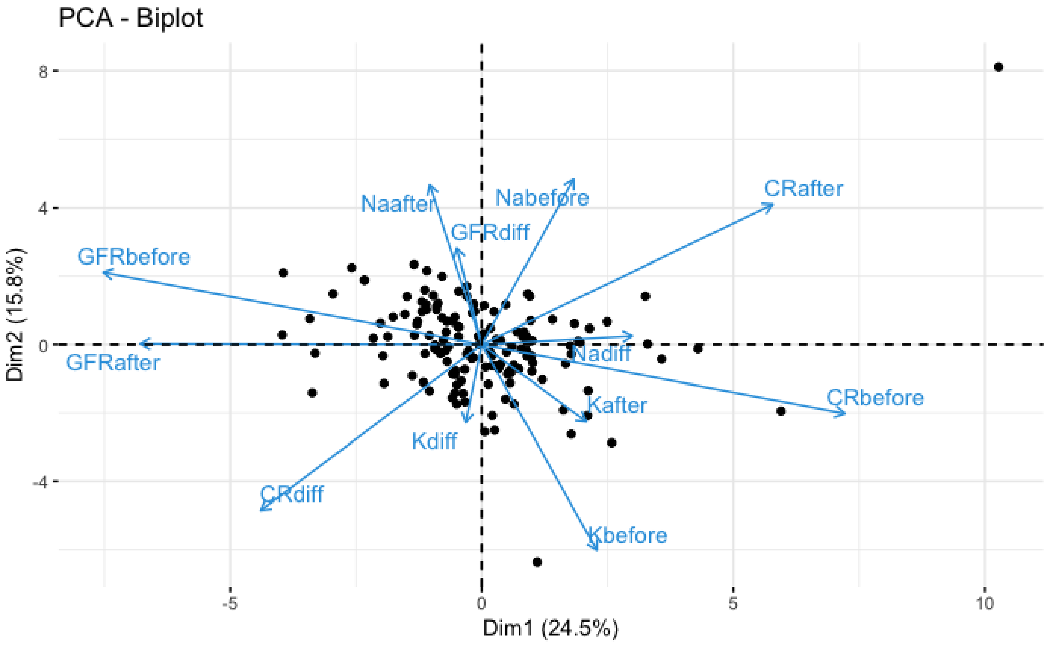


Figure B.328: Biplot plot of the kidney variables, n=145 (Disease control (n=114), Progressive disease (n=31)).

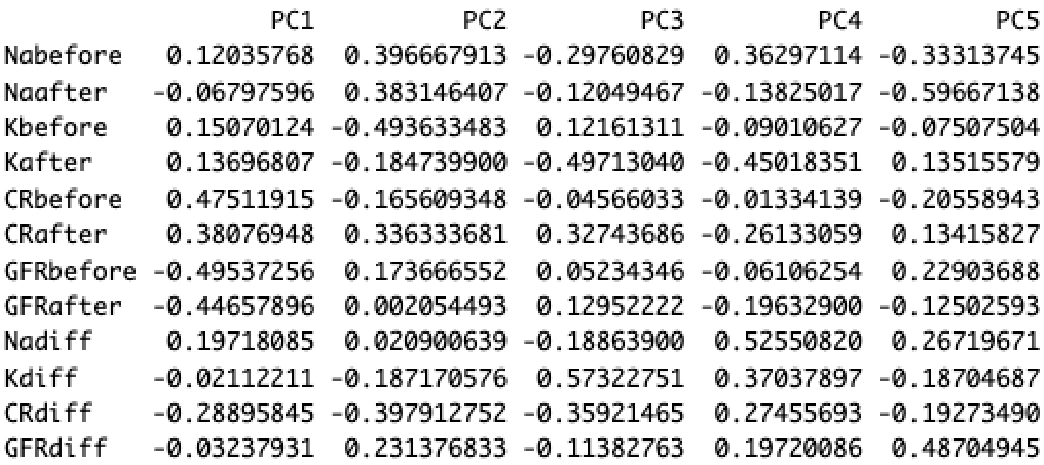

```
                  PC1          PC2         PC3         PC4         PC5
Nabefore   0.12035768  0.396667913 -0.29760829  0.36297114 -0.33313745
Naafter   -0.06797596  0.383146407 -0.12049467 -0.13825017 -0.59667138
Kbefore    0.15070124 -0.493633483  0.12161311 -0.09010627 -0.07507504
Kafter     0.13696807 -0.184739900 -0.49713040 -0.45018351  0.13515579
CRbefore   0.47511915 -0.165609348 -0.04566033 -0.01334139 -0.20558943
CRafter    0.38076948  0.336333681  0.32743686 -0.26133059  0.13415827
GFRbefore -0.49537256  0.173666552  0.05234346 -0.06106254  0.22903688
GFRafter  -0.44657896  0.002054493  0.12952222 -0.19632900 -0.12502593
Nadiff     0.19718085  0.020900639 -0.18863900  0.52550820  0.26719671
Kdiff     -0.02112211 -0.187170576  0.57322751  0.37037897 -0.18704687
CRdiff    -0.28895845 -0.397912752 -0.35921465  0.27455693 -0.19273490
GFRdiff   -0.03237931  0.231376833 -0.11382763  0.19720086  0.48704945
```

Figure B.329: Loading matrix of the kidney variables of the first 5 principal components, n=145 (Disease control (n=114), Progressive disease (n=31)).

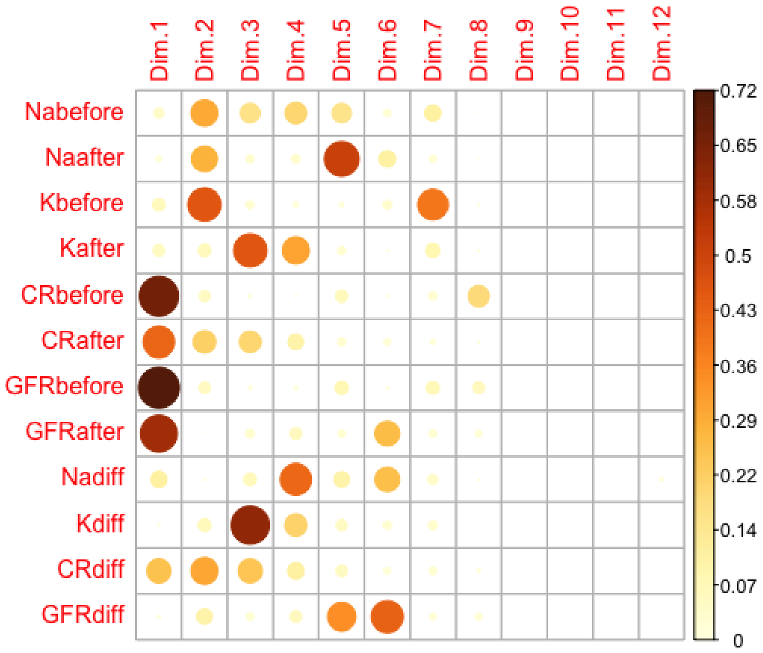


Figure B.330: Correlation of the original kidney variables with the principal components, n=145 (Disease control (n=114), Progressive disease (n=31)).

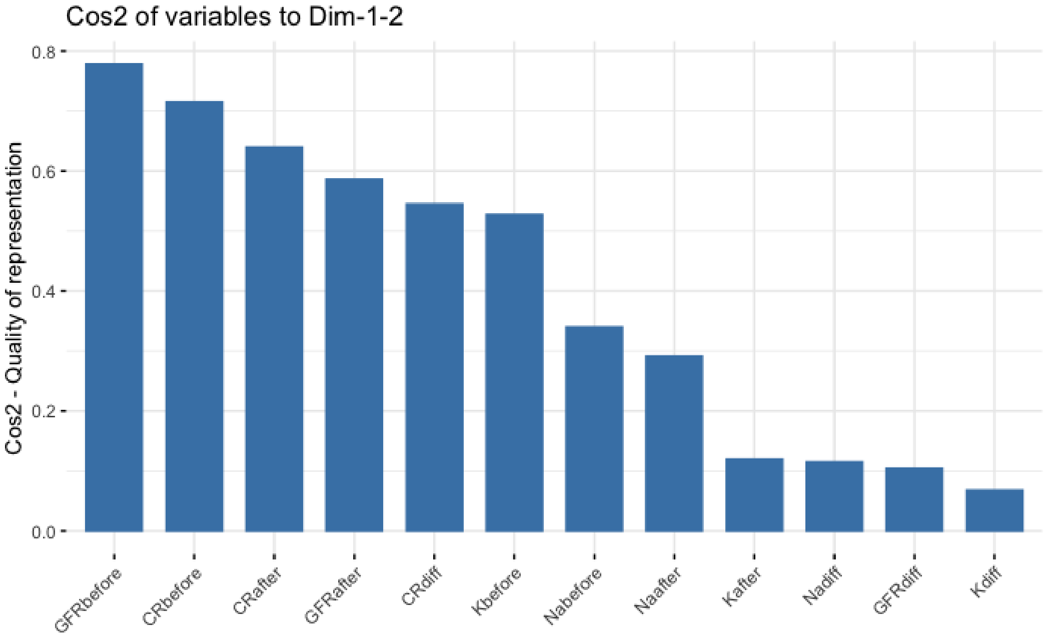


Figure B.331: Cos2 bar chart of the original kidney variables, n=145 (Disease control (n=114), Progressive disease (n=31)).

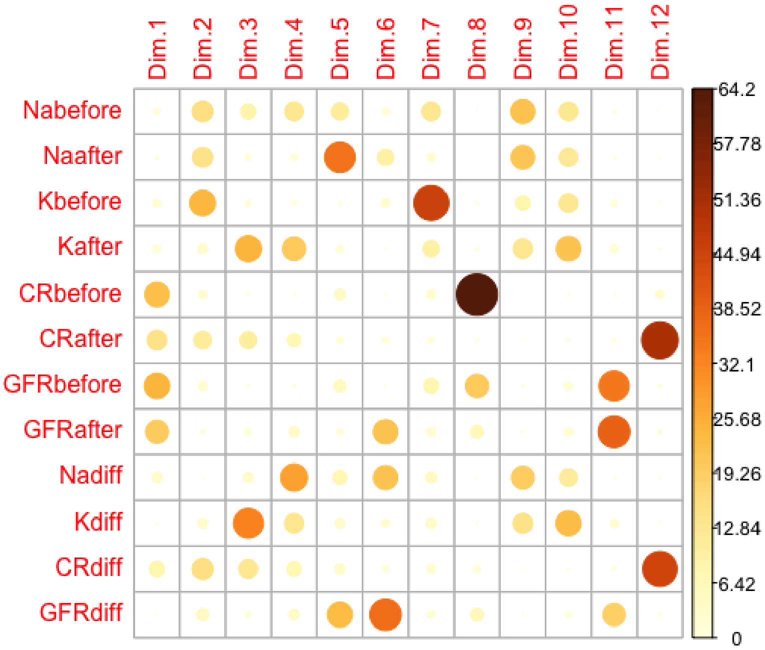


Figure B.332: Contribution of the original kidney variables to the principal components, n=145 (Disease control (n=114), Progressive disease (n=31)).

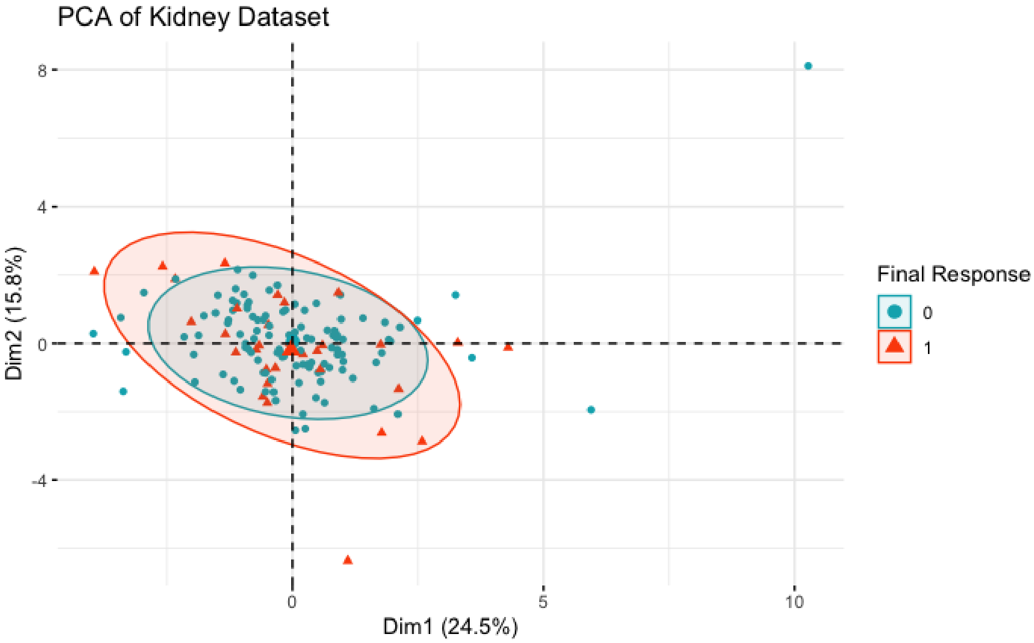


Figure B.333: PCA plot of the kidney group with 95% prediction ellipses, n=145 (Disease control (n=114), Progressive disease (n=31)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

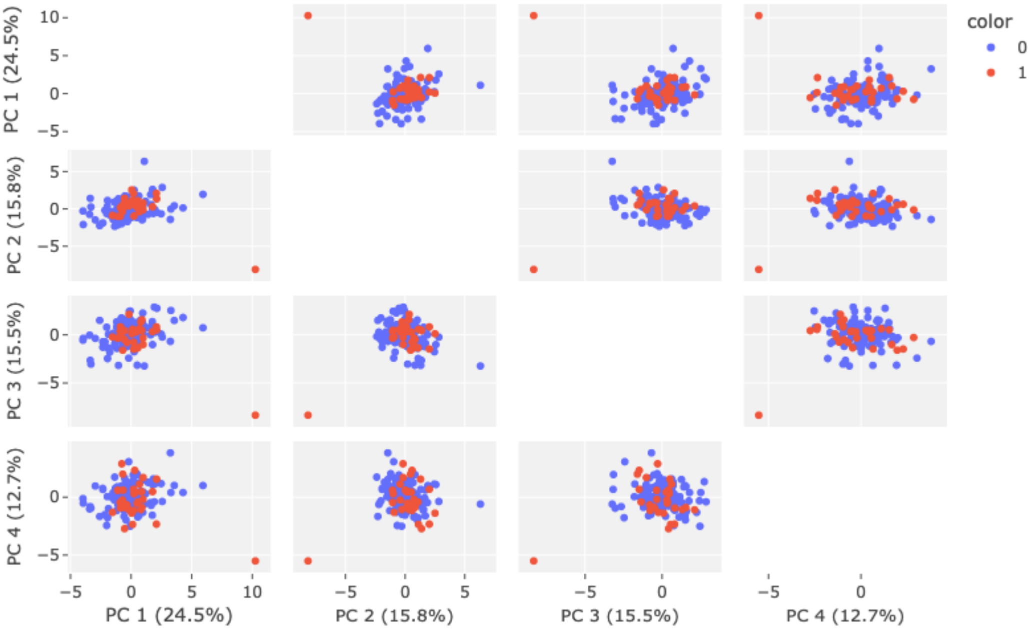


Figure B.334: Scatter matrix plot of the first four principal components. The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the kidney variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

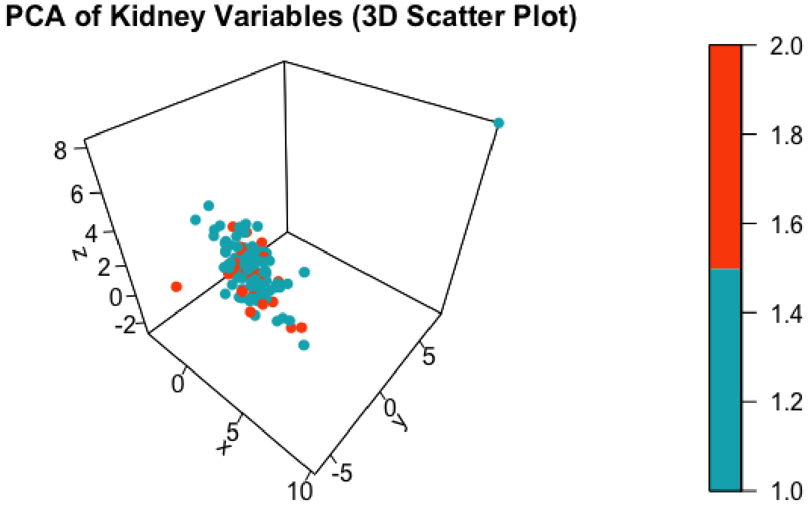


Figure B.335: 3D plot of the kidney variables projected onto the first three principal components.

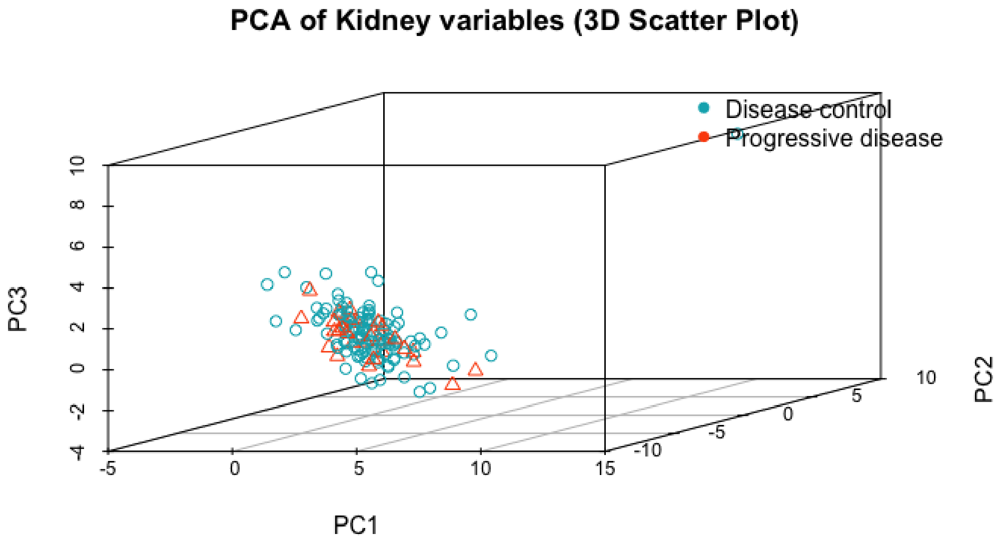


Figure B.336: 3D plot of the kidney variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease),.

B.4.3 Liver Function

The assessment of liver function is evaluated using the variables Aspartate Aminotransferase (ASAT), Alanine Aminotransferase (ALAT), Alkaline Phosphatase (AF), Gamma-Glutamyl Transferase (GGT), Bilirubin (BR), and the International Normalized Ratio (INR). These will be grouped together to represent the liver group. However, due to the limited data available for INR, it will not be used in the PCA conducted in this study. The liver dataset comprises of 166 observations of 15 explanatory variables. After the rescaling the variables, the correlation matrix presented in Figure B.337 is obtained. The correlation matrix indicates that ASAT values before and after the first chemotherapy treatment appear to be uncorrelated, while the ALAT and ASAT value before chemotherapy is almost entirely positively correlated as well as their differences. Moreover, large positive correlations are observed between AF and GGT, both before and after the treatment. From a medical standpoint, this result is expected, given that ASAT and ALAT are highly correlated as both enzymes play a critical role in the metabolism of amino acids, and their levels tend to increase or decrease together in response to liver damage or disease. Similarly, AF and GGT may also be highly correlated with each other, as they are both liver enzymes and are associated with liver function. However, the relationship between these variables may be influenced by other factors such as the underlying cause of liver dysfunction or disease, as well as other medical conditions. Furthermore, the BR variable demonstrates little correlation with any of the liver enzymes, although the before and after values exhibit a robust correlation. Interestingly, BRbefore shows a strong correlation with BRdiff (correlation of 0.954). The complete correlation matrix is presented in Figure B.338.

The screeplot in Figure B.339 indicates that the first principal component (PC1) explains approximately 39% of the total variation, while the second (PC2) explains 21% and the third (PC3) explains 12%. Together, the first three PCs account for only 72% of the total variation. To achieve a cut-off value of at least 80%, PC4 is necessary, as it explains 10% of the total variation on its own, and together with the first three PCs, it accounts for 82% of the total variation (further details can be found in Table B.99). This implies that a significant amount of the variability in the data remains unexplained by the first two identified PCs. The reasons for this unexplained variation could stem from various factors, such as noise in the data, non-linear relationships among the variables, or other sources of variation that are not accounted for by the linear combinations of the original variables. Nonetheless, it is still worthwhile to investigate the other PC plots to identify possible patterns in the first two dimensions defined by PC1 and PC2. The scatter plot of the scores shown in Figure B.340 reveals that many observations are concentrated around the origin and have a blue/green value, indicating that these observations are poorly represented in this principal component space. However, the cluster in the lower right corner appears to be better represented compared to the others. Furthermore, the outlier observations are well captured by the first two PCs. This suggests that the outlier observations may be driving a significant amount of the variability in the data, and they may be crucial to consider in further analysis and exploration, or may be eliminated to explore general trends better.

The loading plot depicted in Figure B.341 illustrates that the majority of loading vectors are pointing towards the left semi-circle, and mainly towards the first quadrant in the plane. Specifically, the variables ALAT and ASAT before values exhibit strong correlations with PC1, as well as ALAT and ASATdiff, while the variables ASAT and ALAT after values show correlation with PC2. Additionally, the vectors of the variables GGT and AF are close together in the loading plot, indicating a strong positive correlation between these variables. Although BRbefore and BRdiff are close together and close to PC1, they are only moderately represented in the loading plot, while BRafter is poorly represented, suggesting that PC1 and PC2 are not good at explaining any variation in this variable. In addition, due to the moderate representation in the first two dimensions, no conclusions about these variables can be drawn. Figure B.344 and Figure B.345 demonstrate that PC1 is primarily correlated with ASAT, ALAT before and diff values, and to a lesser extent, AFbefore, GGTbefore, BRbefore, AFdiff, GGTdiff, and BRdiff. This can mainly be explained by the fact that these variables exhibit positive correlation as well with each other, which was seen earlier in Figure B.337. On the other hand, the second dimension is mostly associated with AFafter and GGTafter, with smaller correlations to ASATafter and ALATafter. The third dimension does not exhibit any clear correlations, with the biggest one seeming to originate from AFdiff, while the fourth dimension (PC4) is primarily correlated with BRbefore, BRafter, and BRdiff. Thus, PC1 and PC2 are mainly related to liver enzymes, while PC4 is related to bilirubin.

Recall that the contribution of each original variable to each PC is determined by the loading values, which represent the correlation between the original variables and the PCs. They indicate the strength and direction of the relationship between each variable and each PC, with larger loading values indicating stronger correlations. However, it is possible for a variable to have a high correlation with a particular PC, but still have a relatively small contribution to that PC, if it is also highly correlated with other variables that have even stronger correlations with that PC. This is the case if the variables are highly correlated with each other, resulting in redundancy in the information they provide. In this case, the contribution of each variable to a particular PC may be diluted by the contributions of other highly correlated variables. For example in the liver data, PC1 is strongly correlated with several variables, including ASAT, ALAT before and diff values, as well as AFBefore, GGTbefore and a few others. However, the contribution plot in Figure B.346 shows that the contributions of the original variables to PC1 are relatively evenly distributed, with no single variable dominating. The reason may be because the variables are highly correlated with each other and together they capture a broad range of variability in the data that is not easily assigned to one single variable. In other words, when two or more variables are highly correlated, they are capturing similar information about the data. When we are performing PCA, we are trying to find a new set of variables (PCs) that capture as much of the variability in the data as possible. If two variables are highly correlated, then they are likely to both contribute to the same principal component and it may be difficult to attribute the variability captured by that component to one variable in particular. So, the variables may be redundant in the sense that they are capturing the same information, so it may be difficult to disentangle their individual contributions to the principal components. This is exactly seen back here: correlations between the variables and the principal components is high, but the contribution of each individual variable to those components is relatively low.

Furthermore, the biplot in Figure B.342 shows a summary of the scatter and loading plot provided earlier with the loadings mainly in the first and third quadrant while many observations are clustered around the origin with a large proportion in the fourth quadrant. The final response plotted in colour with the corresponding 95% prediction ellipses in Figure B.347 do not show a clear distinguishment between the final response outcomes. However, the only thing that is apparent is that the disease control group ellipse is smaller and more elongated along PC1, while the progressive disease group is more elongated along PC2. When the two ellipses overlap, it means that there is some degree of similarity between the two categories in terms of their distribution in the PCA space. However, the fact that one ellipse is more elongated and the other more circular suggests that there may be some differences in the variability or direction of the data long certain principal components. When an ellipse is more elongated, it means that the variance of the observations is larger in one direction compared to the other. This indicates that there is more variability along that principal component. On the other hand, when an ellipse is more circular, it means that the variance is similar in all directions, indicating less variability along each principal components. It is to be noted that the groups may be better separated on higher-order PCs. The plot in Figure B.348 shows the scores of the values plotted in the first four dimensions defined by the first four principal components. Nonetheless, no clear separation between the two final responses can be seen. Neither do the 3D-plots in Figure B.349 and Figure B.350 provide a more insights into the scores projected onto the first 3 principal components.

Liver Function	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	2.416	1.784	1.346	1.242	0.977
Proportion of Variance Explained	0.389	0.212	0.121	0.103	0.064
Cumulative Proportion	0.389	0.601	0.722	0.825	0.888

Table B.99: PCA summary values of the liver group. It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components.

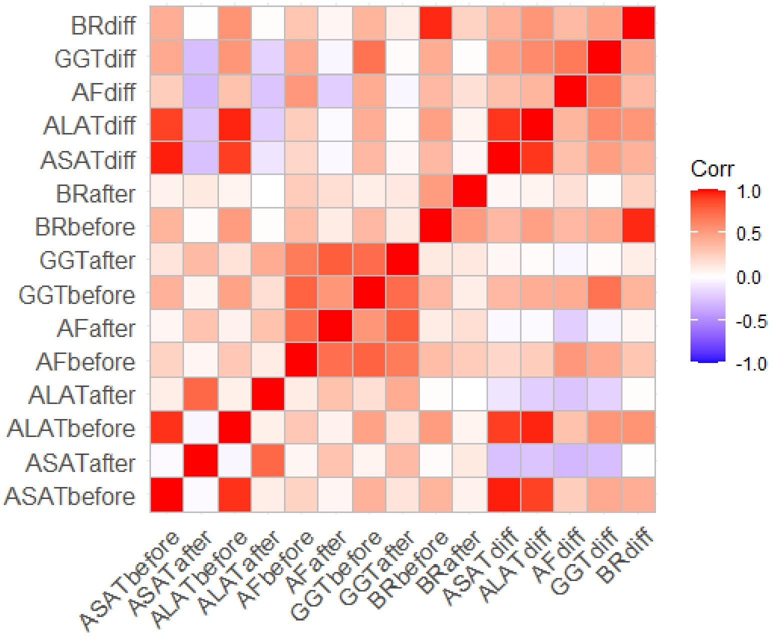


Figure B.337: Correlation Matrix of the various variables in the Liver group. Red = Positive correlation, Blue = negative correlation, White = No correlation.

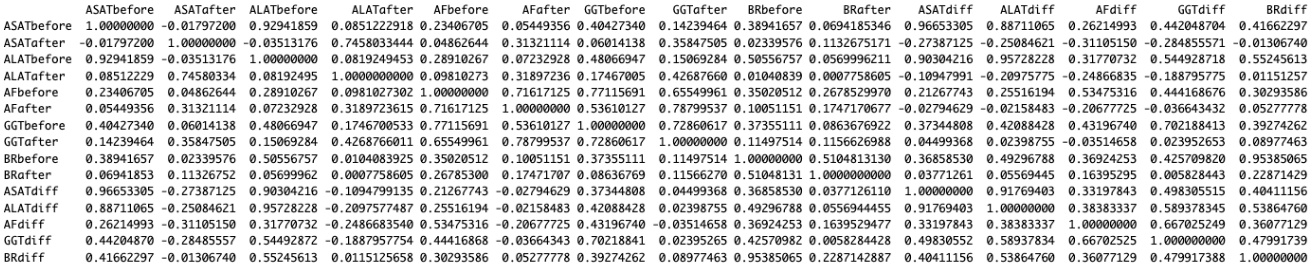

```
           ASATbefore   ASATafter  ALATbefore     ALATafter   AFbefore     AFafter  GGTbefore    GGTafter   BRbefore       BRafter    ASATdiff    ALATdiff      AFdiff      GGTdiff      BRdiff
ASATbefore 1.00000000 -0.01797200  0.92941859  0.0851222918 0.23406705  0.05449356 0.40427340  0.14239464 0.38941657  0.0694185346  0.96653305  0.88711065  0.26214993  0.442048704  0.41662297
ASATafter -0.01797200  1.00000000 -0.03513176  0.7458033444 0.04862644  0.31321114 0.06014138  0.35847505 0.02339576  0.1132675171 -0.27387125 -0.25084621 -0.31105150 -0.284855571 -0.01306740
ALATbefore 0.92941859 -0.03513176  1.00000000  0.0819249453 0.28910267  0.07232928 0.48066947  0.15069284 0.50556757  0.0569996211  0.90304216  0.95728228  0.31770732  0.544928718  0.55245613
ALATafter  0.08512229  0.74580334  0.08192495  1.0000000000 0.09810273  0.31897236 0.17467005  0.42687660 0.01040839  0.0007758605 -0.10947991 -0.20975775 -0.24866835 -0.188795775  0.01151257
AFbefore   0.23406705  0.04862644  0.28910267  0.0981027302 1.00000000  0.71617125 0.77115691  0.65549961 0.35020512  0.2678529970  0.21267743  0.25516194  0.53475316  0.444168676  0.30293586
AFafter    0.05449356  0.31321114  0.07232928  0.3189723615 0.71617125  1.00000000 0.53610127  0.78799537 0.10051151  0.1747170677 -0.02794629 -0.02158483 -0.20677725 -0.036643432  0.05277778
GGTbefore  0.40427340  0.06014138  0.48066947  0.1746700533 0.77115691  0.53610127 1.00000000  0.72860617 0.37355111  0.0863676922  0.37344808  0.42088428  0.43196740  0.702188413  0.39274262
GGTafter   0.14239464  0.35847505  0.15069284  0.4268766011 0.65549961  0.78799537 0.72860617  1.00000000 0.11497514  0.1156626988  0.04499368  0.02398755 -0.03514658  0.023952653  0.08977463
BRbefore   0.38941657  0.02339576  0.50556757  0.0104083925 0.35020512  0.10051151 0.37355111  0.11497514 1.00000000  0.5104813130  0.36858530  0.49296788  0.36924253  0.425709820  0.95385065
BRafter    0.06941853  0.11326752  0.05699962  0.0007758605 0.26785300  0.17471707 0.08636769  0.11566270 0.51048131  1.0000000000  0.03771261  0.05569445  0.16395295  0.005828443  0.22871429
ASATdiff   0.96653305 -0.27387125  0.90304216 -0.1094799135 0.21267743 -0.02794629 0.37344808  0.04499368 0.36858530  0.0377126110  1.00000000  0.91769403  0.33197843  0.498305515  0.40411156
ALATdiff   0.88711065 -0.25084621  0.95728228 -0.2097577487 0.25516194 -0.02158483 0.42088428  0.02398755 0.49296788  0.0556944455  0.91769403  1.00000000  0.38383337  0.589378345  0.53864760
AFdiff     0.26214993 -0.31105150  0.31770732 -0.2486683540 0.53475316 -0.20677725 0.43196740 -0.03514658 0.36924253  0.1639529477  0.33197843  0.38383337  1.00000000  0.667025249  0.36077129
GGTdiff    0.44204870 -0.28485557  0.54492872 -0.1887957754 0.44416868 -0.03664343 0.70218841  0.02395265 0.42570982  0.0058284428  0.49830552  0.58937834  0.66702525  1.000000000  0.47991739
BRdiff     0.41662297 -0.01306740  0.55245613  0.0115125658 0.30293586  0.05277778 0.39274262  0.08977463 0.95385065  0.2287142887  0.40411156  0.53864760  0.36077129  0.479917388  1.00000000
```

Figure B.338: Correlation Matrix of the various variables in the Liver group with numerical values.

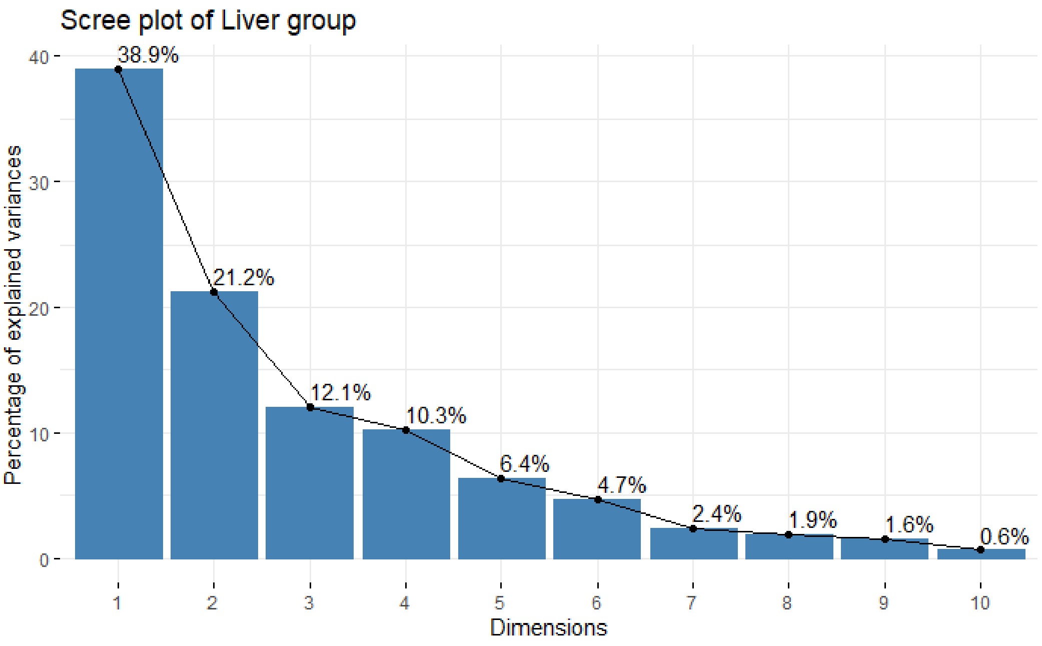


Figure B.339: Scree plot of the liver variables.

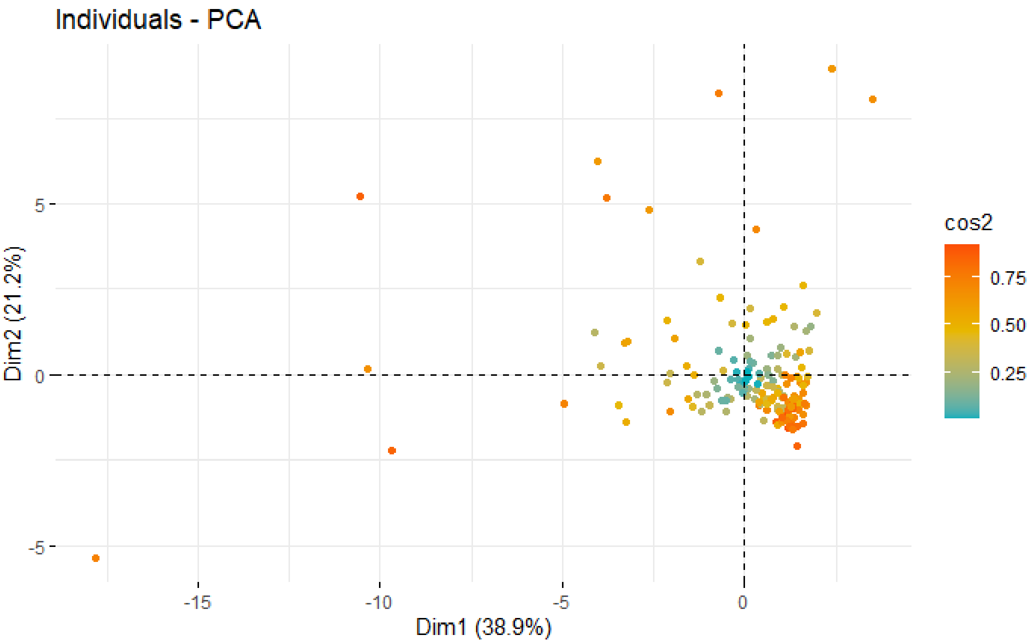


Figure B.340: Scatter plot of the liver variables.

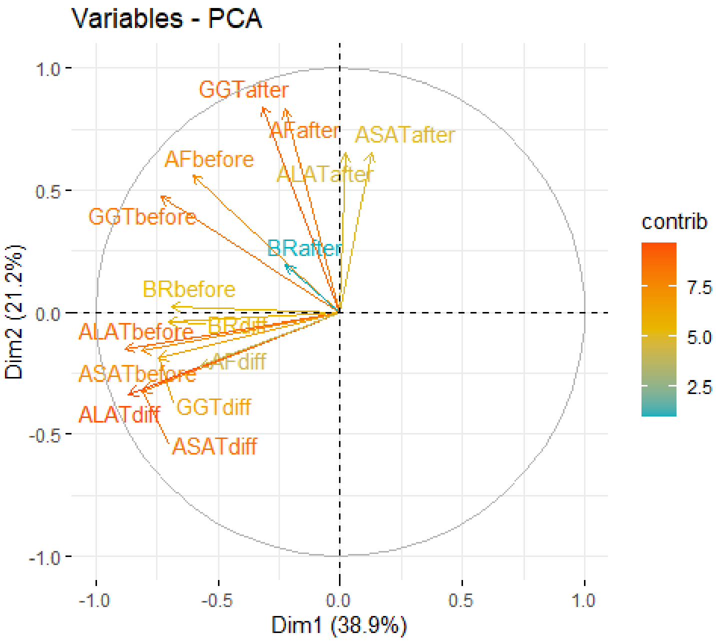


Figure B.341: Loading plot of the liver variables. The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

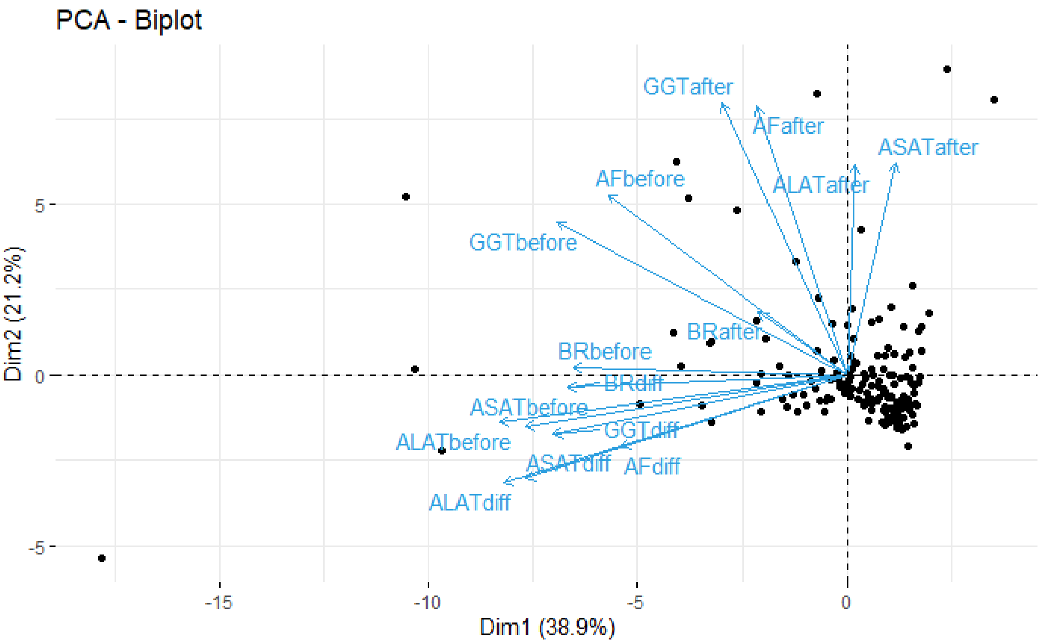
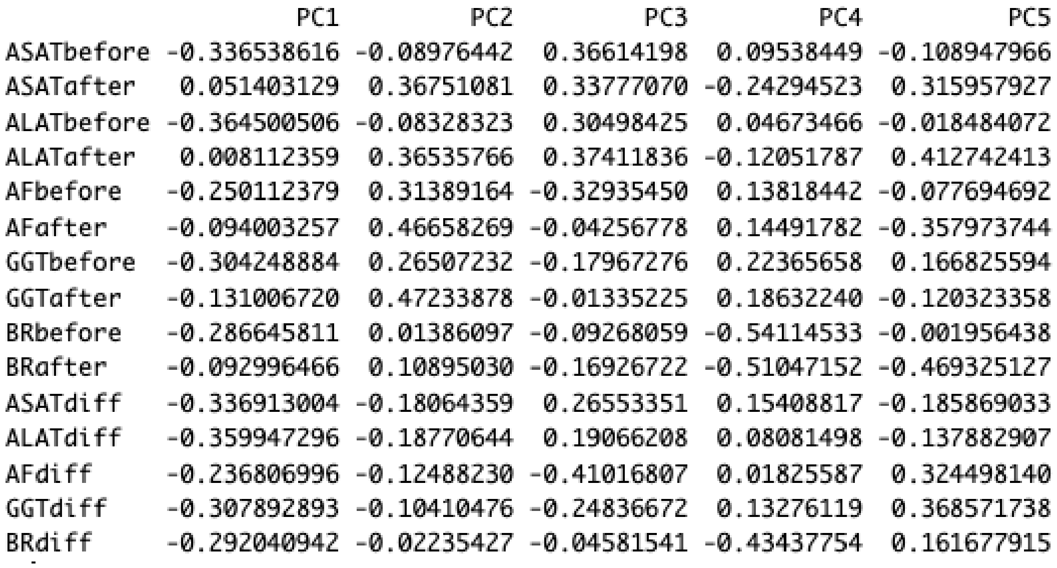


Figure B.342: Biplot plot of the liver variables.



```
                   PC1         PC2         PC3         PC4          PC5
ASATbefore -0.336538616 -0.08976442  0.36614198  0.09538449 -0.108947966
ASATafter   0.051403129  0.36751081  0.33777070 -0.24294523  0.315957927
ALATbefore -0.364500506 -0.08328323  0.30498425  0.04673466 -0.018484072
ALATafter   0.008112359  0.36535766  0.37411836 -0.12051787  0.412742413
AFbefore   -0.250112379  0.31389164 -0.32935450  0.13818442 -0.077694692
AFafter    -0.094003257  0.46658269 -0.04256778  0.14491782 -0.357973744
GGTbefore  -0.304248884  0.26507232 -0.17967276  0.22365658  0.166825594
GGTafter   -0.131006720  0.47233878 -0.01335225  0.18632240 -0.120323358
BRbefore   -0.286645811  0.01386097 -0.09268059 -0.54114533 -0.001956438
BRafter    -0.092996466  0.10895030 -0.16926722 -0.51047152 -0.469325127
ASATdiff   -0.336913004 -0.18064359  0.26553351  0.15408817 -0.185869033
ALATdiff   -0.359947296 -0.18770644  0.19066208  0.08081498 -0.137882907
AFdiff     -0.236806996 -0.12488230 -0.41016807  0.01825587  0.324498140
GGTdiff    -0.307892893 -0.10410476 -0.24836672  0.13276119  0.368571738
BRdiff     -0.292040942 -0.02235427 -0.04581541 -0.43437754  0.161677915
```

Figure B.343: Loading matrix of the liver variables of the first 5 principal components.

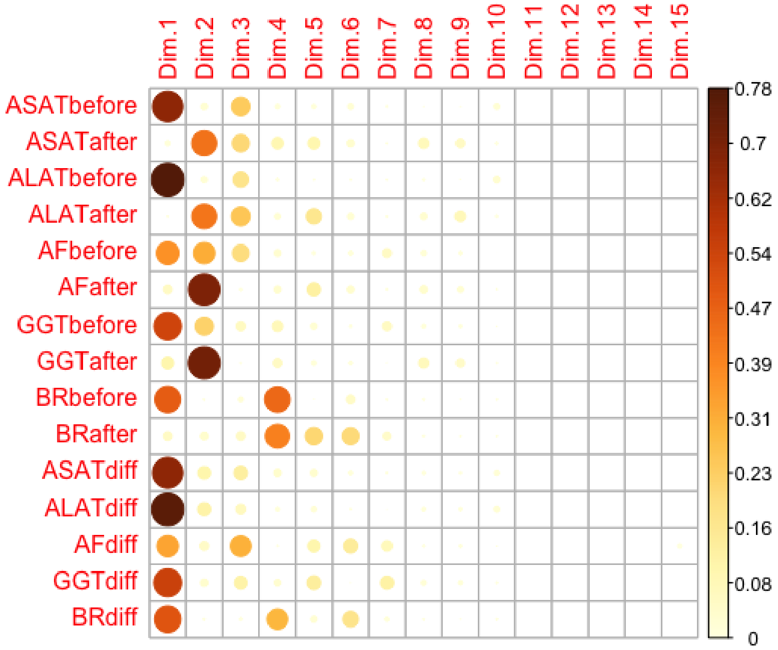


Figure B.344: Correlation of the original liver variables with the principal components.

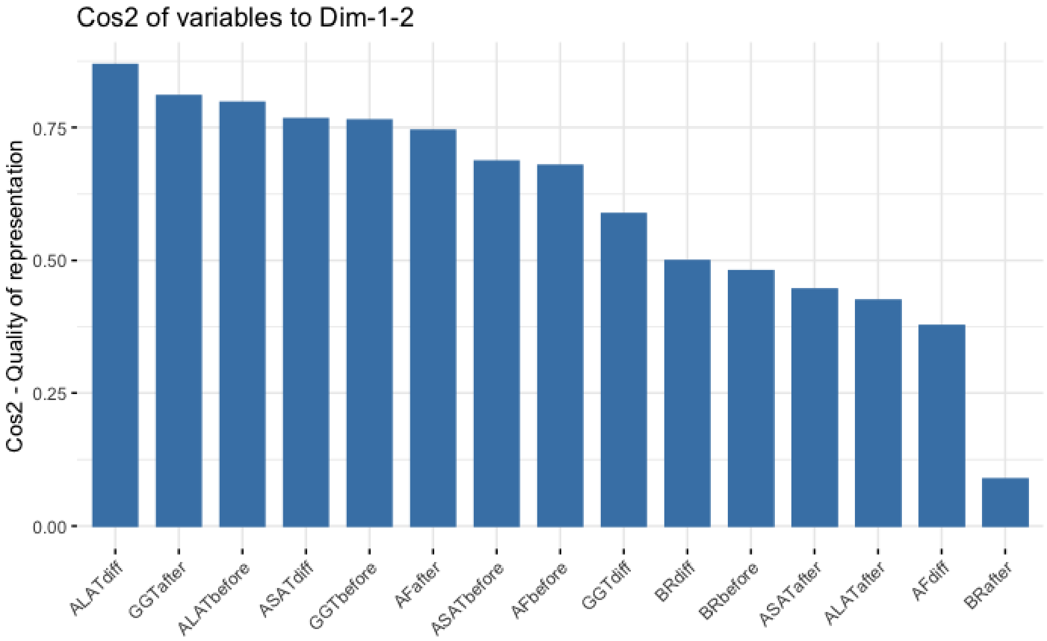


Figure B.345: Cos2 bar chart of the original liver variables.

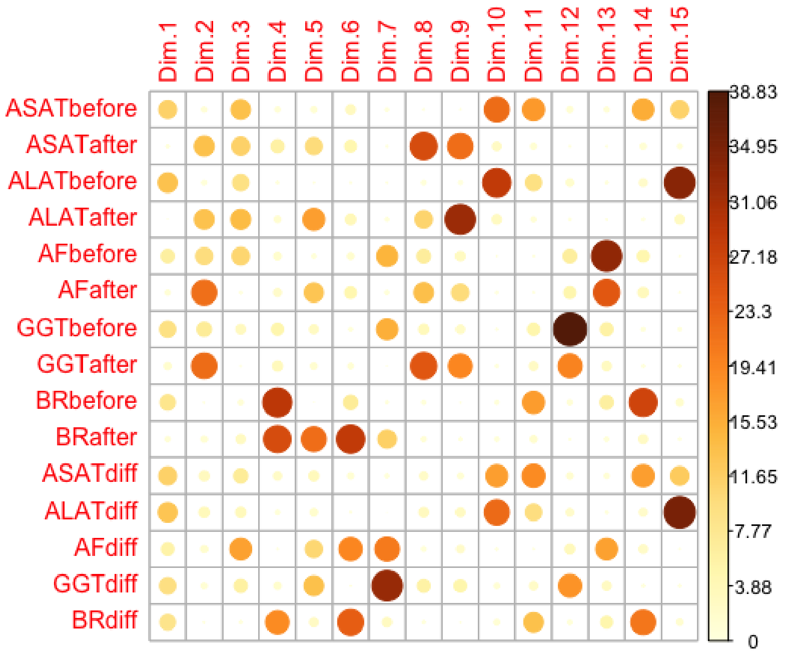


Figure B.346: Contribution of the original liver variables to the principal components.

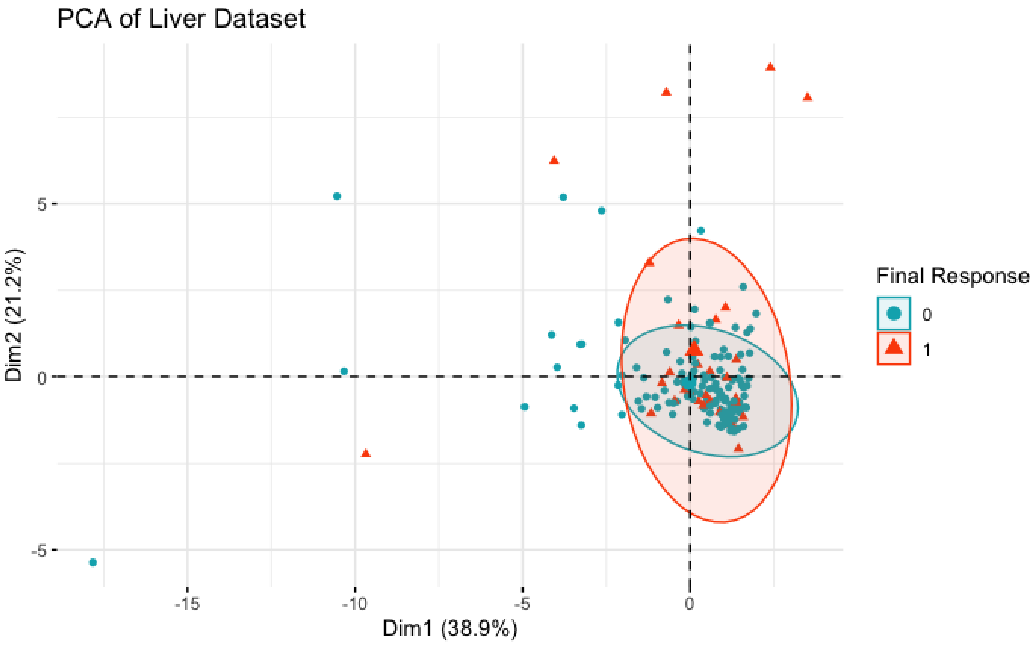


Figure B.347: PCA plot of the liver group with 95% prediction ellipses. The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

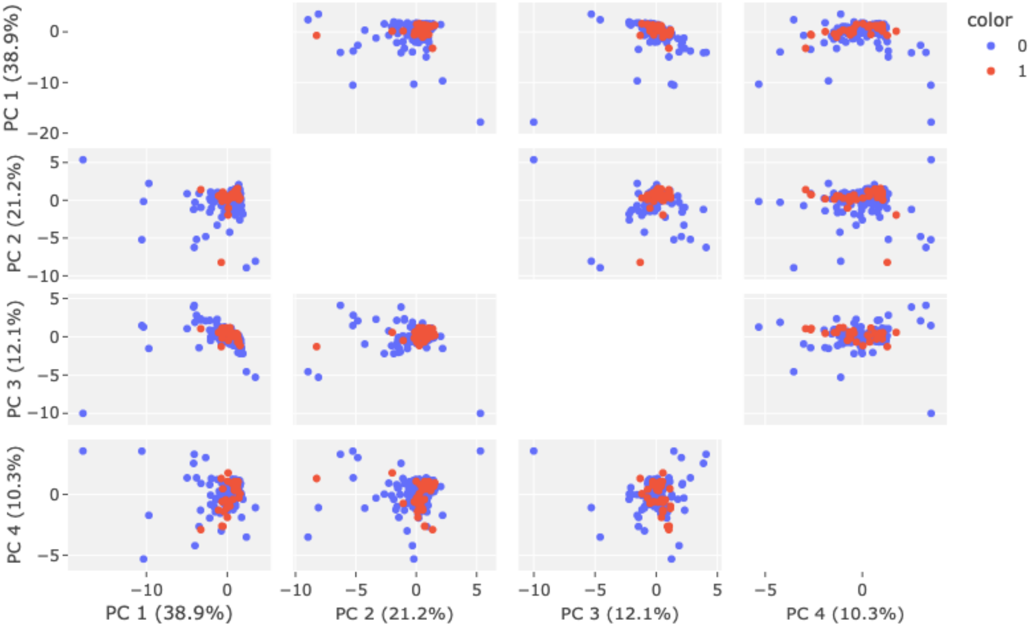


Figure B.348: Scatter matrix plot of the first four principal components, n=145 (Disease control (n=114), Progressive disease (n=31)). The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the liver variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

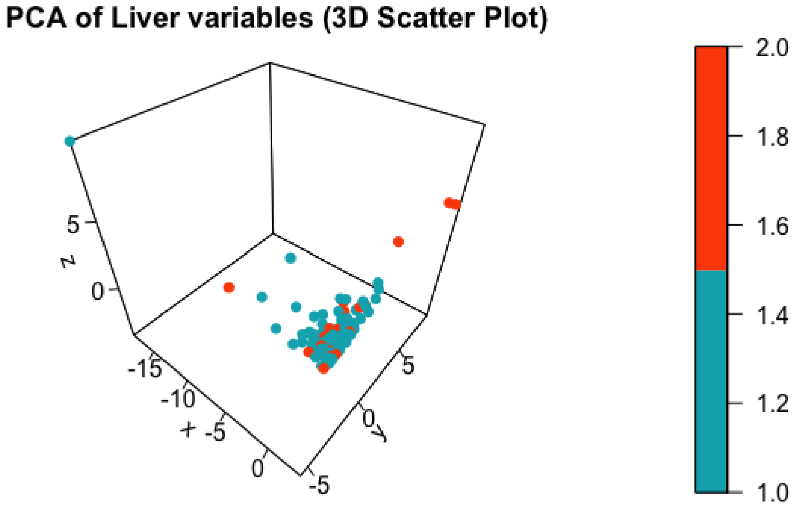


Figure B.349: 3D plot of the liver variables projected onto the first three principal components, n=145 (Disease control (n=114), Progressive disease (n=31)).

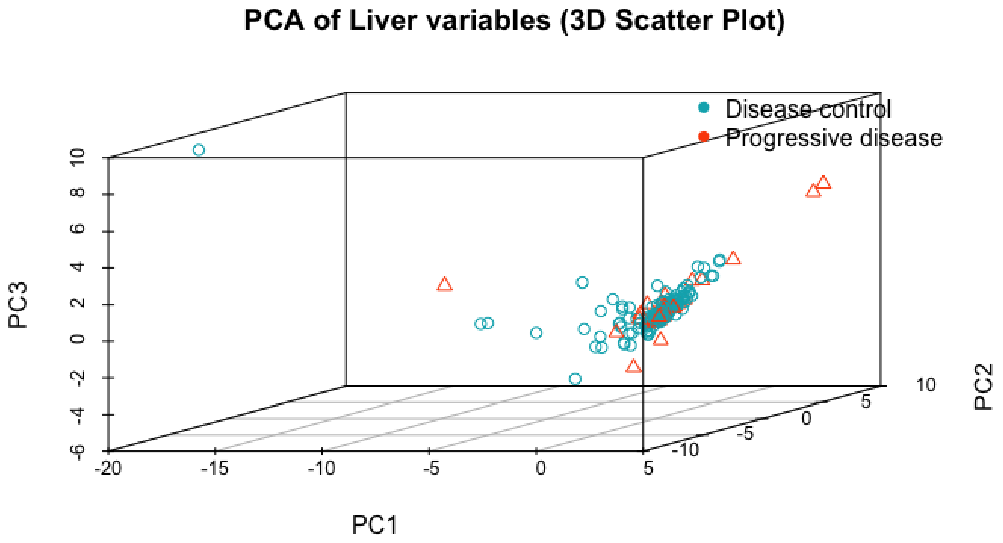


Figure B.350: 3D plot of the liver variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=145 (Disease control (n=114), Progressive disease (n=31)).

B.4.4 Nutrition status

The Nutrition Status will be analyzed using the variables: Albumin (Alb), Sodium (Na) and Potassium (K) before and after the first chemotherapy treatment. The dataset comprises a total of 102 observations with 9 explanatory variables. After rescaling the variables, the correlation matrix presented in Figure B.351 indicates a strong positive correlation between the before and after values of all variables. An interesting observation is that the Sodium and Potassium levels seem to be negatively correlated. In general, changes in sodium levels can lead to changes in potassium levels or vice versa, but they do not have to be negatively correlated. For example, in the cases of dehydration, both sodium and potassium levels may decrease due to fluid loss. However, the relationship can be influenced by many different factors. Furthermore, Albumin levels seem not to be strongly correlated to neither sodium nor potassium. PCA applied to the nutrition group reveals that the first three principal components explain a relatively low proportion of the variability in the data. Only by including the fifth PC, the cumulative proportion of variance explained surpasses 80% and even reaches 90%. These findings suggest that the data may have more complex structure, requiring additional variables to fully capture the range of variability. Given that the data comprises only 9 explanatory variables consisting of essentially 3 parameters (Alb, Na, and K), it may be preferable to consider the original variables rather than the PCs.

The scatter plot displayed in Figure B.354 indicates that the majority of the scores are poorly represented by only the first two dimensions, which is not surprising given that these dimensions capture less than half of the variance in the data. The loading plot depicted in Figure B.355 shows that only the variables “Na-after" and “K-after" are reasonably well-represented in the first two dimensions, with “Na-after" highly correlated with PC1 and “K-after" highly correlated with PC2. Conversely, the remaining variables are poorly projected in the first two dimensions, precluding any meaningful conclusions about them. The biplot in Figure B.356 and the prediction ellipse plot in Figure B.361 are also inconclusive and fail to reveal any discernible patterns, aside from the observation that the disease control ellipse appears circular while the progressive disease ellipse is elongated. Projecting the observations onto the first four principal components, as depicted in Figure B.362 does not show any potential separation between the two final response groups either. Nevertheless, the correlation plot in Figure B.358 indicates several intriguing observations. Specifically, PC1 is primarily correlated to the after first chemotherapy albumin and sodium values, while PC2 is mainly correlated by the before first chemotherapy sodium values and after first chemotherapy potassium values. PC3, on the other hand, is correlated to the difference in potassium levels, while PC4 is correlated to the before first chemotherapy albumin levels and PC5 is associated with the differences in albumin and sodium levels. Despite these notable correlations, the strength of these associations is not particularly strong (ranging from 0 to 0.68), as reflected in the low quality of representation in dimensions 1 and 2, illustrated in Figure B.359. Similarly, Figure B.360 demonstrates that the variables correlated with a particular dimension have high contributions, indicating that each dimension is mainly influenced by one or two original variables. Nevertheless, further exploration is necessary to determine whether working with the original variables rather than the principal components would be more advantageous in light of the fact that the variables are not highly correlated with each other. Consequently, reducing the dimensionality of the data would not be advantageous in this instance. Finally, Figure B.363 and Figure B.364 present the scores of the observations projected onto the first three principal components in 3D plots.

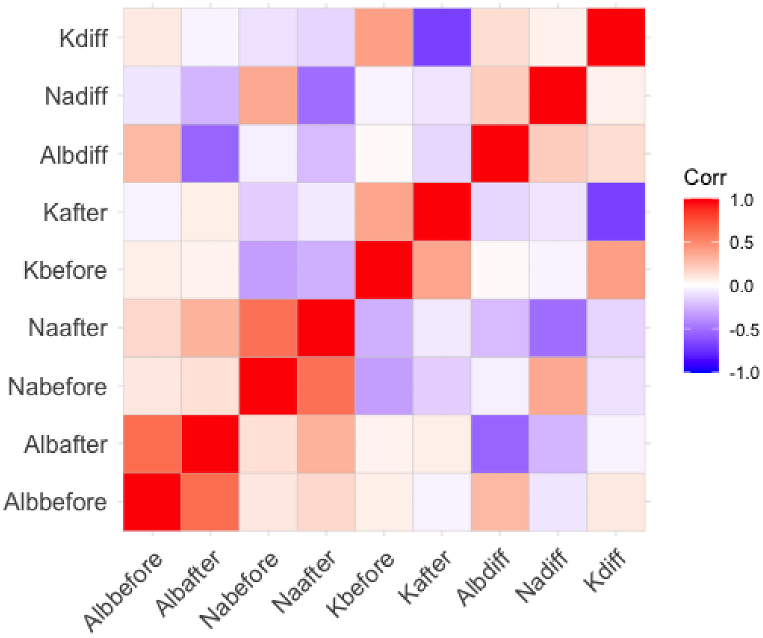
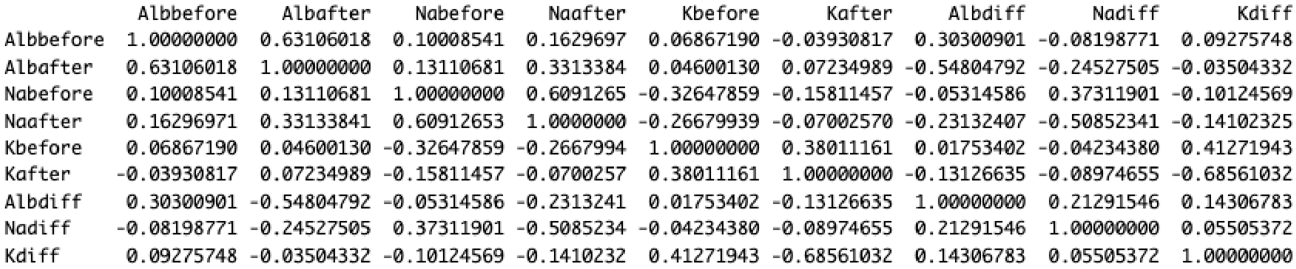


Figure B.351: Correlation Matrix of the various variables in the Nutrition group, n=102 (Disease control (n=80), Progressive disease (n=22)). Red = Positive correlation, Blue = negative correlation, White = No correlation.



```
            Albbefore    Albafter    Nabefore    Naafter     Kbefore      Kafter     Albdiff      Nadiff       Kdiff
Albbefore  1.00000000  0.63106018  0.10008541  0.1629697  0.06867190 -0.03930817  0.30300901 -0.08198771  0.09275748
Albafter   0.63106018  1.00000000  0.13110681  0.3313384  0.04600130  0.07234989 -0.54804792 -0.24527505 -0.03504332
Nabefore   0.10008541  0.13110681  1.00000000  0.6091265 -0.32647859 -0.15811457 -0.05314586  0.37311901 -0.10124569
Naafter    0.16296971  0.33133841  0.60912653  1.0000000 -0.26679939 -0.07002570 -0.23132407 -0.50852341 -0.14102325
Kbefore    0.06867190  0.04600130 -0.32647859 -0.2667994  1.00000000  0.38011161  0.01753402 -0.04234380  0.41271943
Kafter    -0.03930817  0.07234989 -0.15811457 -0.0700257  0.38011161  1.00000000 -0.13126635 -0.08974655 -0.68561032
Albdiff    0.30300901 -0.54804792 -0.05314586 -0.2313241  0.01753402 -0.13126635  1.00000000  0.21291546  0.14306783
Nadiff    -0.08198771 -0.24527505  0.37311901 -0.5085234 -0.04234380 -0.08974655  0.21291546  1.00000000  0.05505372
Kdiff      0.09275748 -0.03504332 -0.10124569 -0.1410232  0.41271943 -0.68561032  0.14306783  0.05505372  1.00000000
```

Figure B.352: Correlation Matrix of the various variables in the Nutrition group, n=102 (Disease control (n=80), Progressive disease (n=22)).

Nutrition	PC1	PC2	PC3	PC4	PC5
Standard Deviation	1.516	1.355	1.303	1.128	1.032
Proportion of Variance Explained	0.256	0.204	0.189	0.141	0.118
Cumulative Proportion	0.256	0.459	0.648	0.789	0.908

Table B.100: PCA summary values of the nutrition group, n=102 (Disease control (n=80), Progressive disease (n=22)). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components.

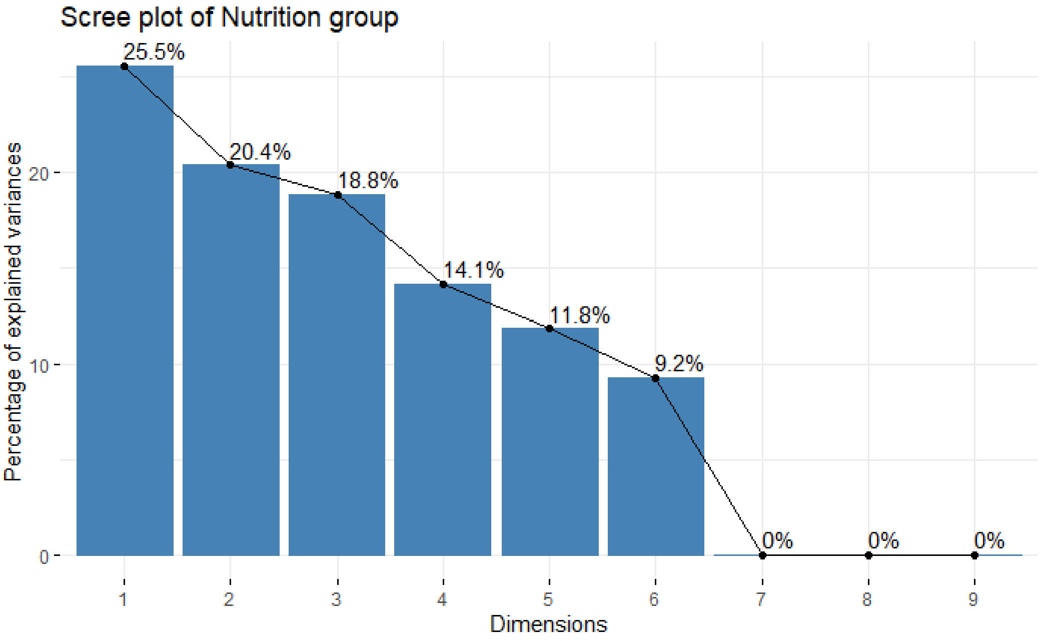


Figure B.353: Scree plot of the nutrition variables, n=102 (Disease control (n=80), Progressive disease (n=22)).

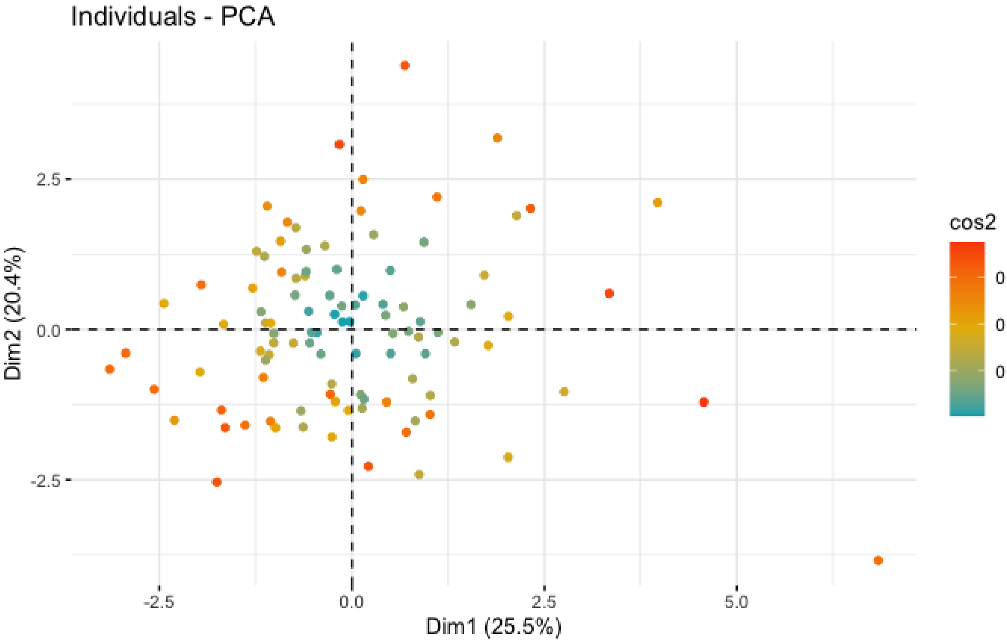


Figure B.354: Scatter plot of the nutrition variables, n=102 (Disease control (n=80), Progressive disease (n=22)).

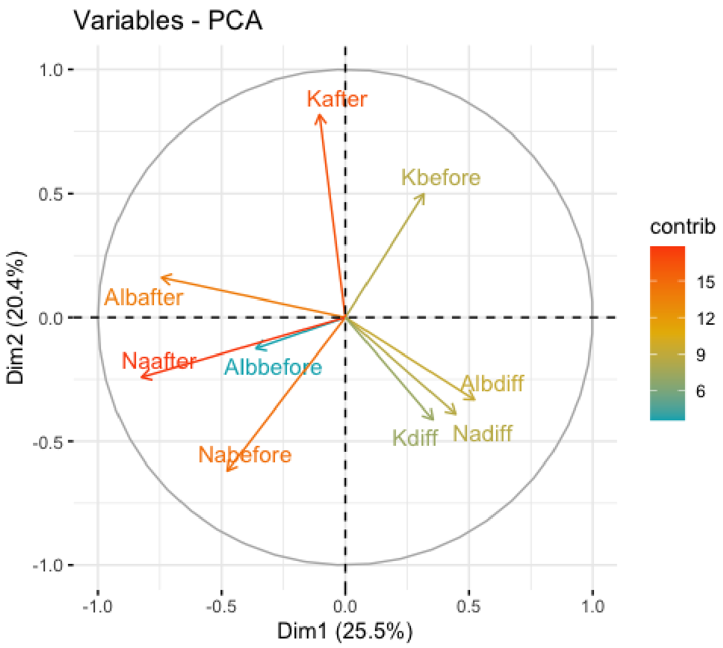


Figure B.355: Loading plot of the nutrition variables, n=102 (Disease control (n=80), Progressive disease (n=22)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

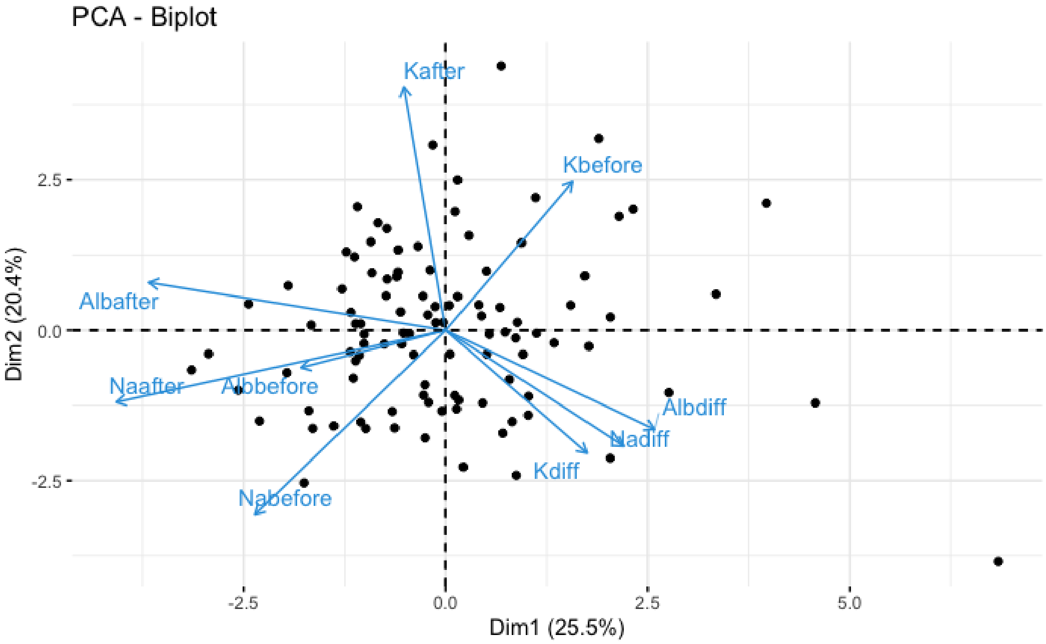
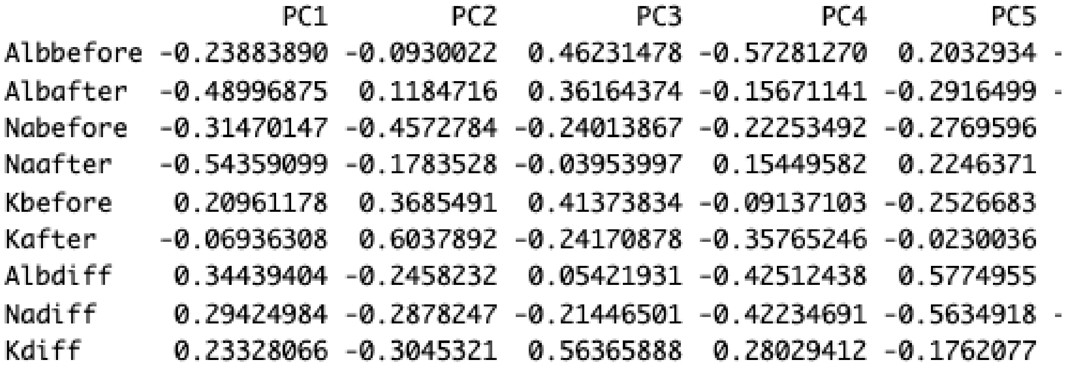


Figure B.356: Biplot plot of the nutrition variables, n=102 (Disease control (n=80), Progressive disease (n=22)).



```
                  PC1        PC2         PC3         PC4        PC5
Albbefore -0.23883890 -0.0930022  0.46231478 -0.57281270  0.2032934 -
Albafter  -0.48996875  0.1184716  0.36164374 -0.15671141 -0.2916499 -
Nabefore  -0.31470147 -0.4572784 -0.24013867 -0.22253492 -0.2769596
Naafter   -0.54359099 -0.1783528 -0.03953997  0.15449582  0.2246371
Kbefore    0.20961178  0.3685491  0.41373834 -0.09137103 -0.2526683
Kafter    -0.06936308  0.6037892 -0.24170878 -0.35765246 -0.0230036
Albdiff    0.34439404 -0.2458232  0.05421931 -0.42512438  0.5774955
Nadiff     0.29424984 -0.2878247 -0.21446501 -0.42234691 -0.5634918 -
Kdiff      0.23328066 -0.3045321  0.56365888  0.28029412 -0.1762077
```

Figure B.357: Loading matrix of the nutrition variables of the first 5 principal components, n=102 (Disease control (n=80), Progressive disease (n=22)).

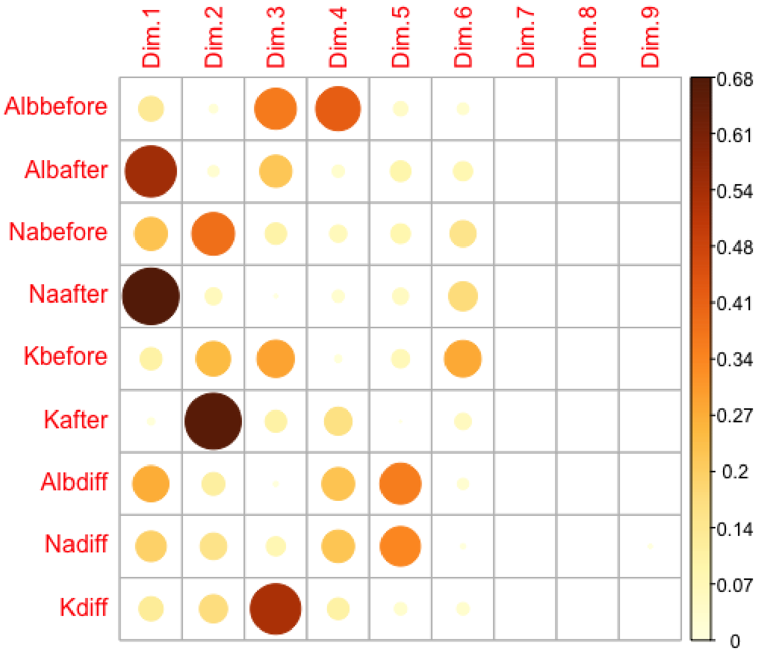


Figure B.358: Correlation of the original liver variables with the principal components, n=102 (Disease control (n=80), Progressive disease (n=22)).

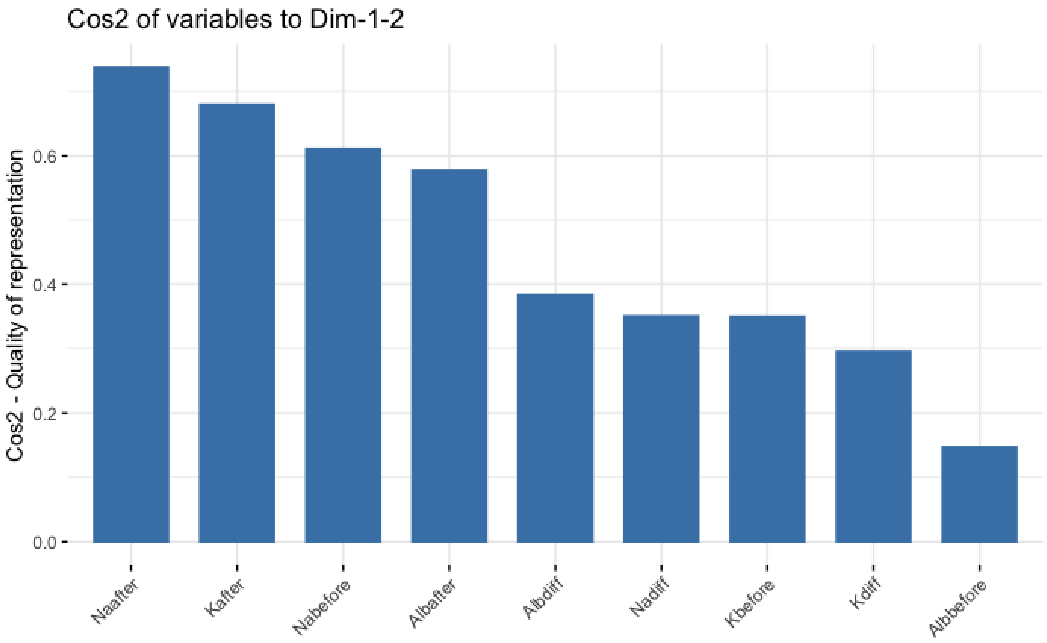


Figure B.359: Cos2 bar chart of the original nutrition variables, n=102 (Disease control (n=80), Progressive disease (n=22)).

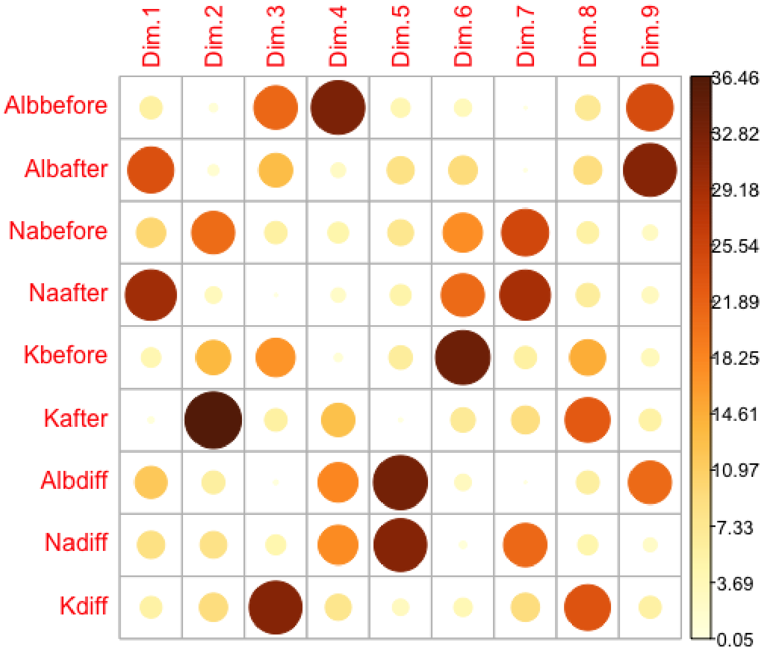


Figure B.360: Contribution of the original nutrition variables to the principal components, n=102 (Disease control (n=80), Progressive disease (n=22)).

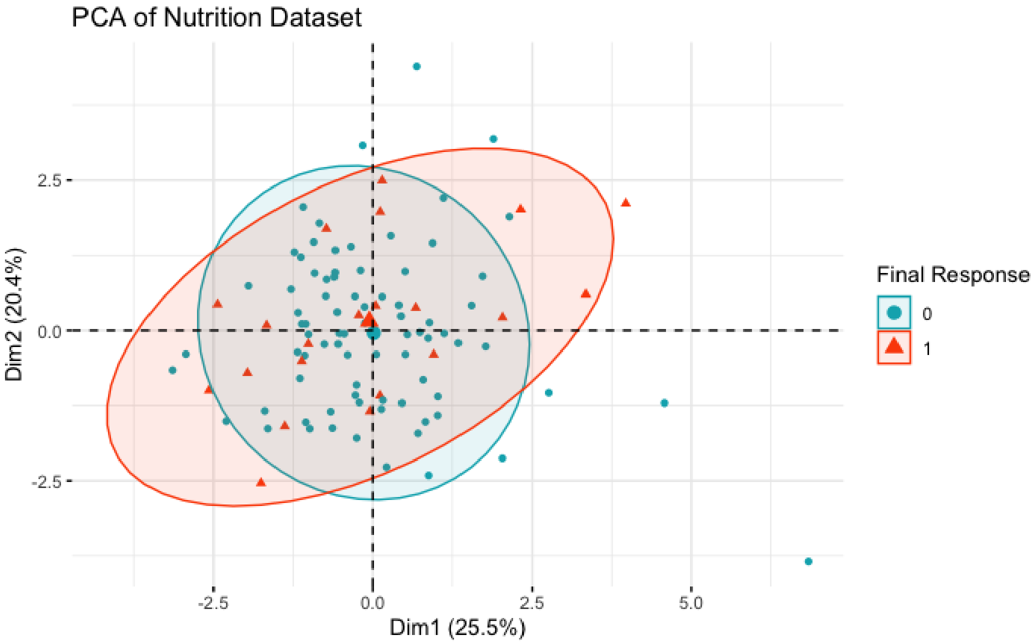


Figure B.361: PCA plot of the nutrition group with 95% prediction ellipses, n=102 (Disease control (n=80), Progressive disease (n=22)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

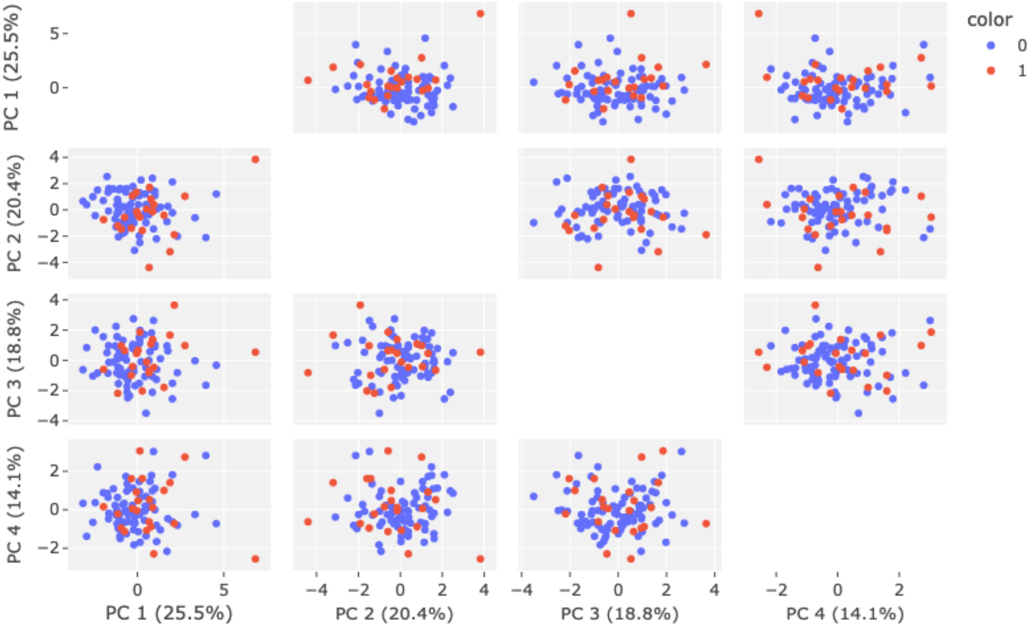


Figure B.362: Scatter matrix plot of the first four principal components, n=102 (Disease control (n=80), Progressive disease (n=22)). The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the nutrition variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

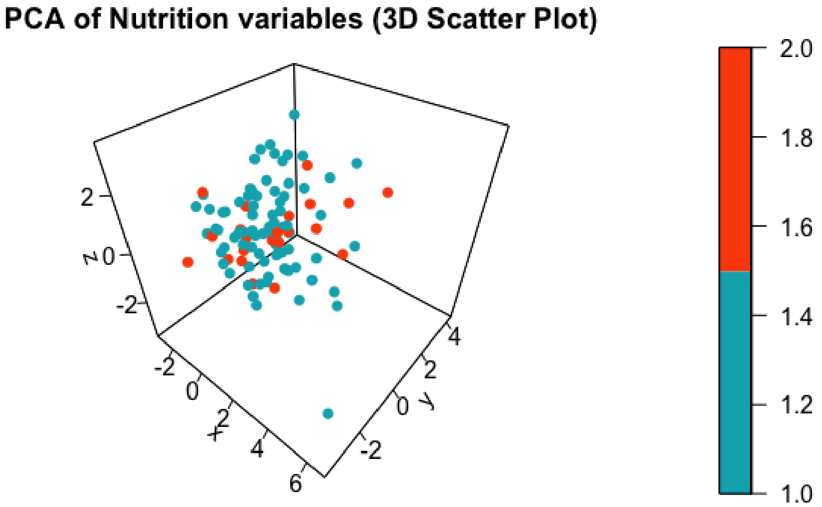


Figure B.363: 3D plot of the nutrition variables projected onto the first three principal components, n=102 (Disease control (n=80), Progressive disease (n=22)).

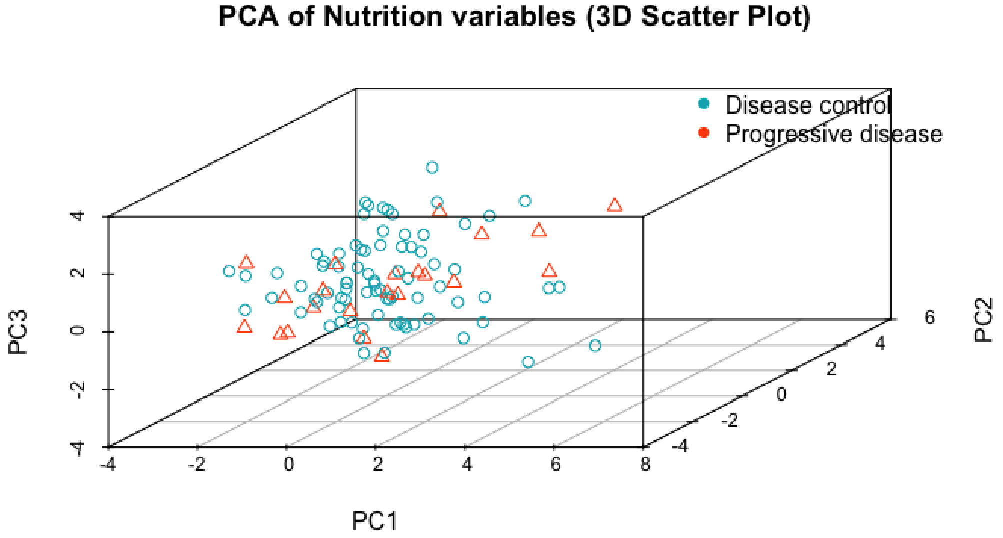


Figure B.364: 3D plot of the nutrition variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=102 (Disease control (n=80), Progressive disease (n=22)).

B.4.5 Inflammation

Inflammation will be assessed using the variables: C-reactive protein (CRP), Leukocytes (LK), Systemic Inflammation Index (SII), Neutrophil-to-Lymphocyte Ratio (NLR) and Platelet-to-Lymphocyte Ratio (PLR). The dataset contains a total of 90 observations with 15 explanatory variables, which have been standardized for the PCA analysis. The correlation matrix, as illustrated in Figure B.365, indicates that there exists positive correlation between the before and after first chemotherapy treatment values of most variables. Interestingly, the PLR values before chemotherapy display no correlation with the NLR values after chemotherapy, and the PLR values are negatively correlated with the LK values before and after chemotherapy. Furthermore, the LK values after chemotherapy exhibit no correlation with the CRP values after chemotherapy. The exact correlation coefficients can be found in Figure B.366.

PCA on the inflammation dataset yields weak explanatory principal components. The first PC explains only 32.6% of the total variation of the data, and at least 4 PCs are required to explain over 80% of the total variation. This suggests that there may be redundancy present due to highly correlated variables or non-linear relationships, similar to the case of liver, kidney, and nutrition groups analyzed earlier. The scatter plot in Figure B.368 shows that most observations are centered around the origin, and many are poorly represented by only using the first two PCs. On the other hand, the loading plot in Figure B.369 shows that several variables are well represented in the first two dimensions, specifically NLRbefore, SIIbefore, PLRbefore, SIIafter, NLRafter, SIIdiff, and NLRdiff, as their vectors are close to the correlation circle and have a red colour. In particular, the before values of NLR, PLR, and SII are highly correlated, as are the SII and NLR after values, and the SIIdiff and NLRdiff. This is expected since SII is calculated using both neutrophils and platelets as shown in Equation (B.4). Additionally, most vectors are in the first two quadrants showing positive correlation with PC2 mainly, and either negative or positive correlation with PC1. Moreover, the biplot in Figure B.370 offers a summary of the aforementioned findings by illustrating how the observations are related to each other and to the variables. The distance between observations in the biplot indicates the degree of similarity or dissimilarity in terms of their variable values. Additionally, the position of an observation relative to the arrows reflects the proportion of each variable present in the observation and how much of each principal component it represents. However, no clear trends or groupings can be observed.

The correlation between the original inflammation variables and the principal components can be observed in Figure B.372. The results support previous observations that PC1 is significantly correlated with SIIdiff, NLRdiff, and PLRdiff, which is logical since these three variables are also strongly correlated with each other. Furthermore, PC2 is strongly correlated with the SII, NLR, and PLR before and after the initial chemotherapy treatment values, which aligns with the same reasoning as before. As illustrated in Figure B.373, PC3 is primarily determined by LKafter and LKdiff, while PC4 is mainly influenced by LKbefore values. Additionally, due to the high correlation present, the contribution of each variable is limited in the first few dimensions, as shown in Figure B.374. In other words, there are redundant variables in this group that are capturing the same information, causing high correlations between the variables and the PCs, but low contributions of each individual variable to those components. Finally, the plot in Figure B.375 shows no clear separation in the final response group using only PC1 and PC2. Projecting the scores onto the first four principal components does not yield a clear separation between the two groups either. Neither do the scores in the first three dimensions, using PC1, PC2, and PC3, which are visualized in Figure B.377 and Figure B.378.

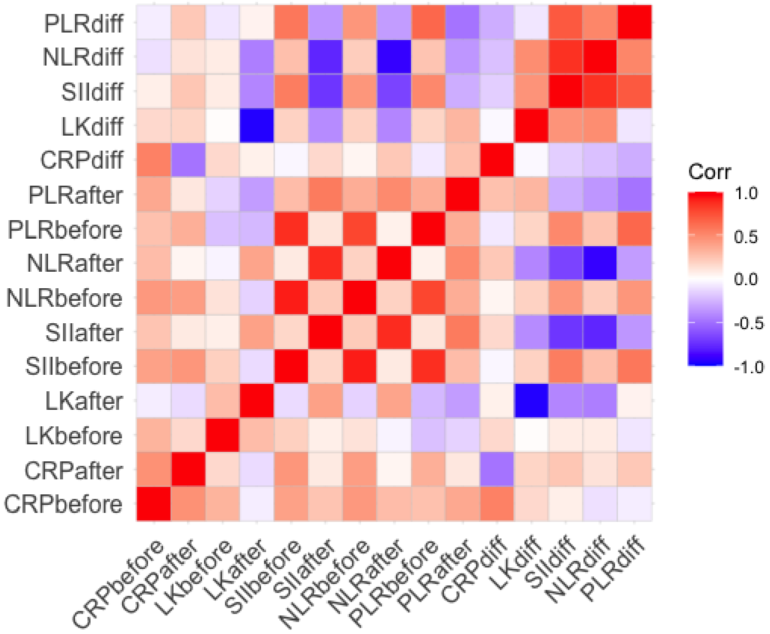
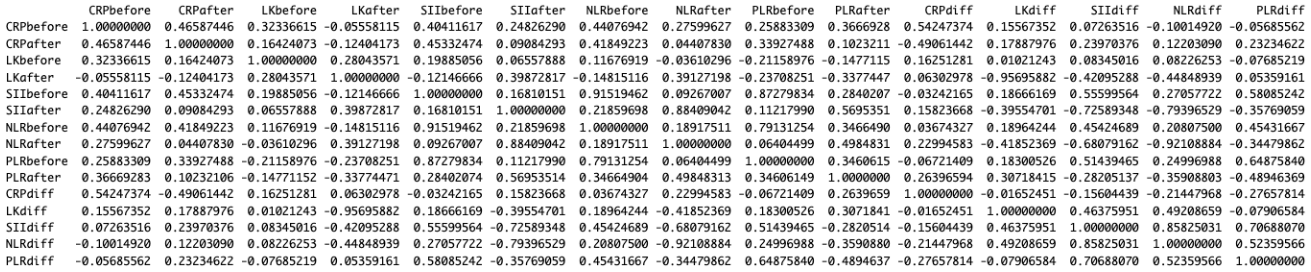


Figure B.365: Correlation Matrix of the various variables in the Inflammation group, n=90 (Disease control (n=71), Progressive disease (n=19)). Red = Positive correlation, Blue = negative correlation, White = No correlation.



```
            CRPbefore    CRPafter    LKbefore     LKafter   SIIbefore    SIIafter   NLRbefore    NLRafter   PLRbefore   PLRafter     CRPdiff      LKdiff     SIIdiff     NLRdiff     PLRdiff
CRPbefore  1.00000000  0.46587446  0.32336615 -0.05558115  0.40411617  0.24826290  0.44076942  0.27599627  0.25883309  0.3666928  0.54247374  0.15567352  0.07263516 -0.10014920 -0.05685562
CRPafter   0.46587446  1.00000000  0.16424073 -0.12404173  0.45332474  0.09084293  0.41849223  0.04407830  0.33927488  0.1023211 -0.49061442  0.17887976  0.23970376  0.12203090  0.23234622
LKbefore   0.32336615  0.16424073  1.00000000  0.28043571  0.19885056  0.06557888  0.11676919 -0.03610296 -0.21158976 -0.1477115  0.16251281  0.01021243  0.08345016  0.08226253 -0.07685219
LKafter   -0.05558115 -0.12404173  0.28043571  1.00000000 -0.12146666  0.39872817 -0.14815116  0.39127198 -0.23708251 -0.3377447  0.06302978 -0.95695882 -0.42095288 -0.44848939  0.05359161
SIIbefore  0.40411617  0.45332474  0.19885056 -0.12146666  1.00000000  0.16810151  0.91519462  0.09267007  0.87279834  0.2840207 -0.03242165  0.18666169  0.55599564  0.27057722  0.58085242
SIIafter   0.24826290  0.09084293  0.06557888  0.39872817  0.16810151  1.00000000  0.21859698  0.88409042  0.11217990  0.5695351  0.15823668 -0.39554701 -0.72589348 -0.79396529 -0.35769059
NLRbefore  0.44076942  0.41849223  0.11676919 -0.14815116  0.91519462  0.21859698  1.00000000  0.18917511  0.79131254  0.3466490  0.03674327  0.18964244  0.45424689  0.20807500  0.45431667
NLRafter   0.27599627  0.04407830 -0.03610296  0.39127198  0.09267007  0.88409042  0.18917511  1.00000000  0.06404499  0.4984831  0.22994583 -0.41852369 -0.68079162 -0.92108884 -0.34479862
PLRbefore  0.25883309  0.33927488 -0.21158976 -0.23708251  0.87279834  0.11217990  0.79131254  0.06404499  1.00000000  0.3460615 -0.06721409  0.18300526  0.51439465  0.24996988  0.64875840
PLRafter   0.36669283  0.10232106 -0.14771152 -0.33774471  0.28402074  0.56953514  0.34664904  0.49848313  0.34606149  1.0000000  0.26396594  0.30718415 -0.28205137 -0.35908803 -0.48946369
CRPdiff    0.54247374 -0.49061442  0.16251281  0.06302978 -0.03242165  0.15823668  0.03674327  0.22994583 -0.06721409  0.2639659  1.00000000 -0.01652451 -0.15604439 -0.21447968 -0.27657814
LKdiff     0.15567352  0.17887976  0.01021243 -0.95695882  0.18666169 -0.39554701  0.18964244 -0.41852369  0.18300526  0.3071841 -0.01652451  1.00000000  0.46375951  0.49208659 -0.07906584
SIIdiff    0.07263516  0.23970376  0.08345016 -0.42095288  0.55599564 -0.72589348  0.45424689 -0.68079162  0.51439465 -0.2820514 -0.15604439  0.46375951  1.00000000  0.85825031  0.70688070
NLRdiff   -0.10014920  0.12203090  0.08226253 -0.44848939  0.27057722 -0.79396529  0.20807500 -0.92108884  0.24996988 -0.3590880 -0.21447968  0.49208659  0.85825031  1.00000000  0.52359566
PLRdiff   -0.05685562  0.23234622 -0.07685219  0.05359161  0.58085242 -0.35769059  0.45431667 -0.34479862  0.64875840 -0.4894637 -0.27657814 -0.07906584  0.70688070  0.52359566  1.00000000
```

Figure B.366: Correlation Matrix of the various variables in the Inflammation group, n=90 (Disease control (n=71), Progressive disease (n=19)).

Inflammation	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	2.213	1.936	1.494	1.259	1.123
Proportion of Variance Explained	0.326	0.250	0.149	0.106	0.084
Cumulative Proportion	0.326	0.576	0.725	0.831	0.915

Table B.101: PCA summary values of the inflammation group, n=90 (Disease control (n=71), Progressive disease (n=19)). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components.

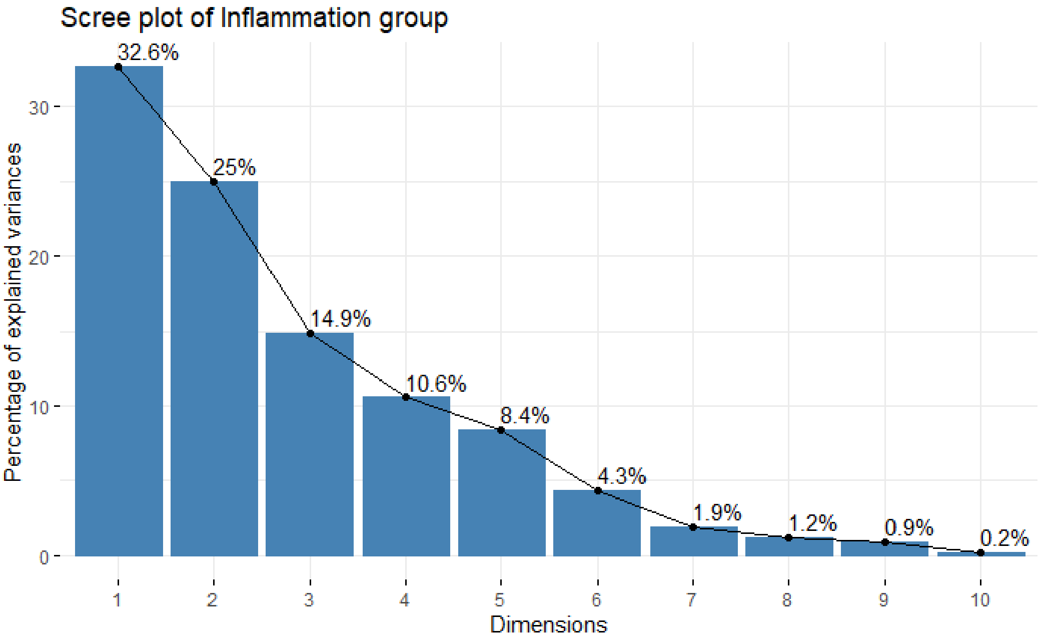


Figure B.367: Scree plot of the inflammation variables, n=90 (Disease control (n=71), Progressive disease (n=19)).

B.4.6 Patient Characteristics

Inclusion of patient characteristics, such as weight, height, age and BMI, can provide valuable information in a dataset. Hence, these four variables are included into the patient characteristics group. It should be noted that PCA is only applicable to continuous variables, thus categorical variables such as gender and smoking status are not included in the analysis. The patient characteristics dataset comprises 189 observations of 4 explanatory variables. The correlation plot depicted in Figure B.379 suggests a positive correlation between height and weight, which is biologically plausible as taller individuals tend to weigh more. Similarly, weight and BMI exhibit a positive correlation, which can be explained by the relationship

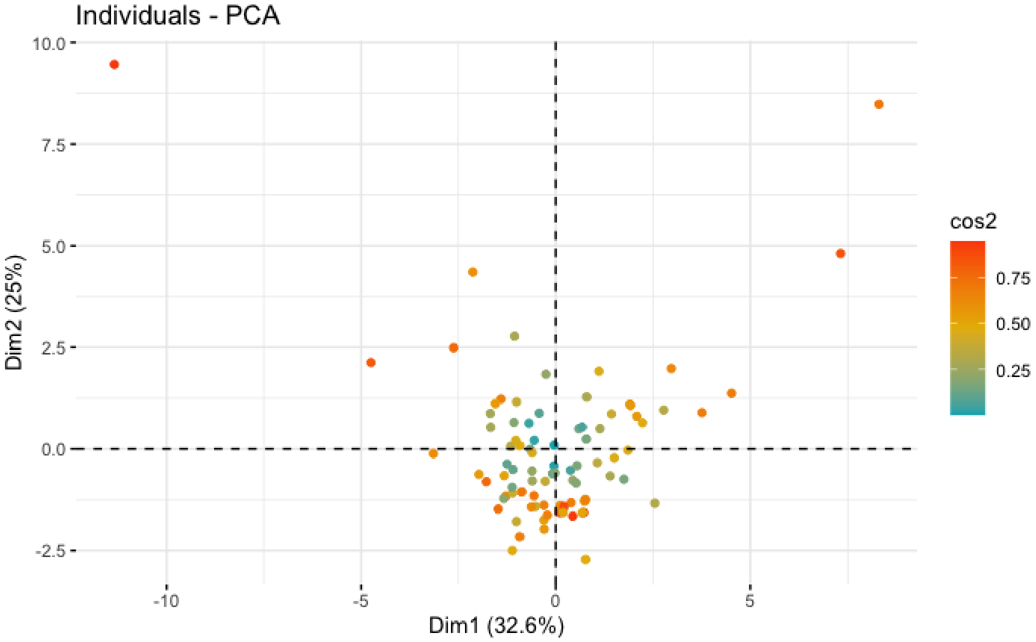


Figure B.368: Scatter plot of the inflammation variables, n=90 (Disease control (n=71), Progressive disease (n=19)).

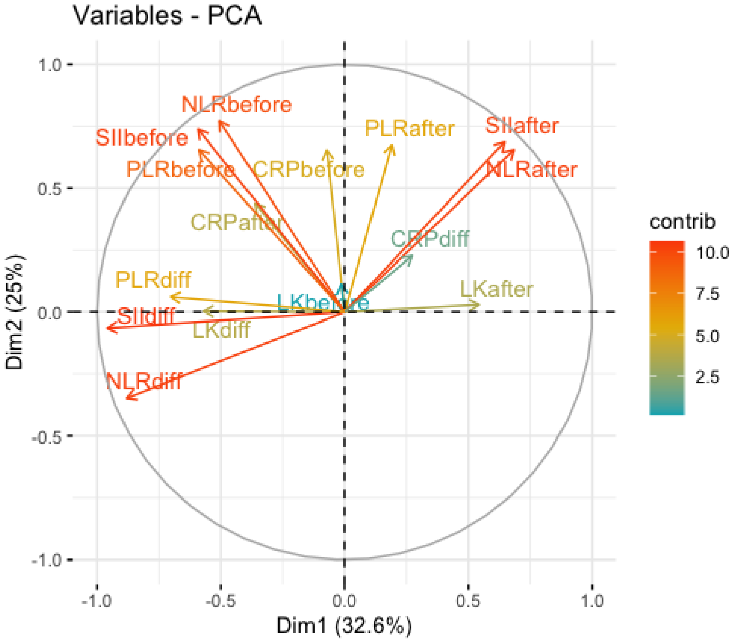


Figure B.369: Loading plot of the inflammation variables, n=90 (Disease control (n=71), Progressive disease (n=19)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

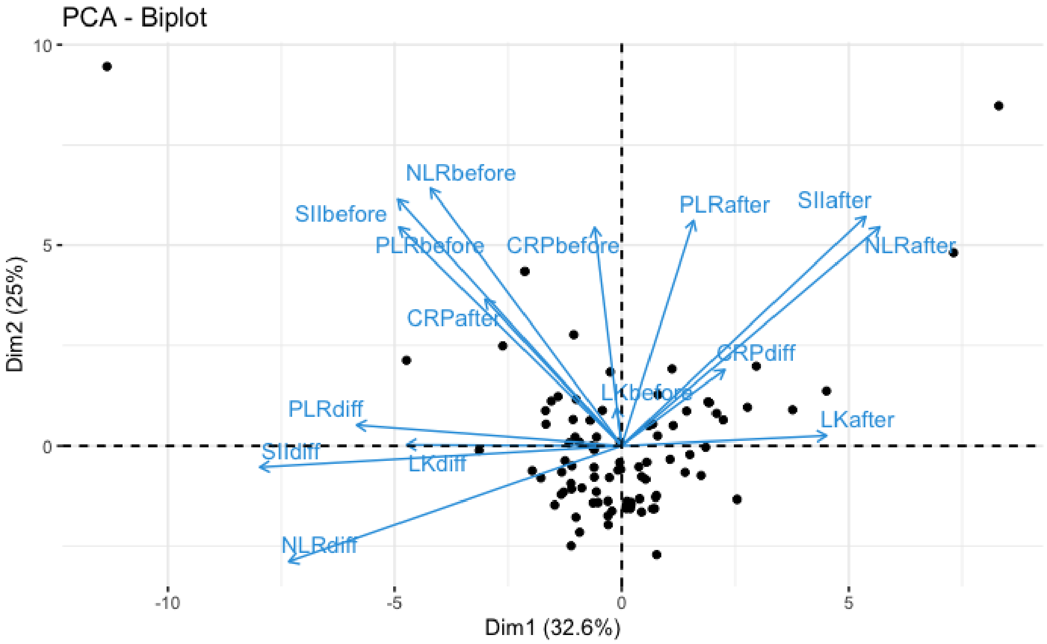
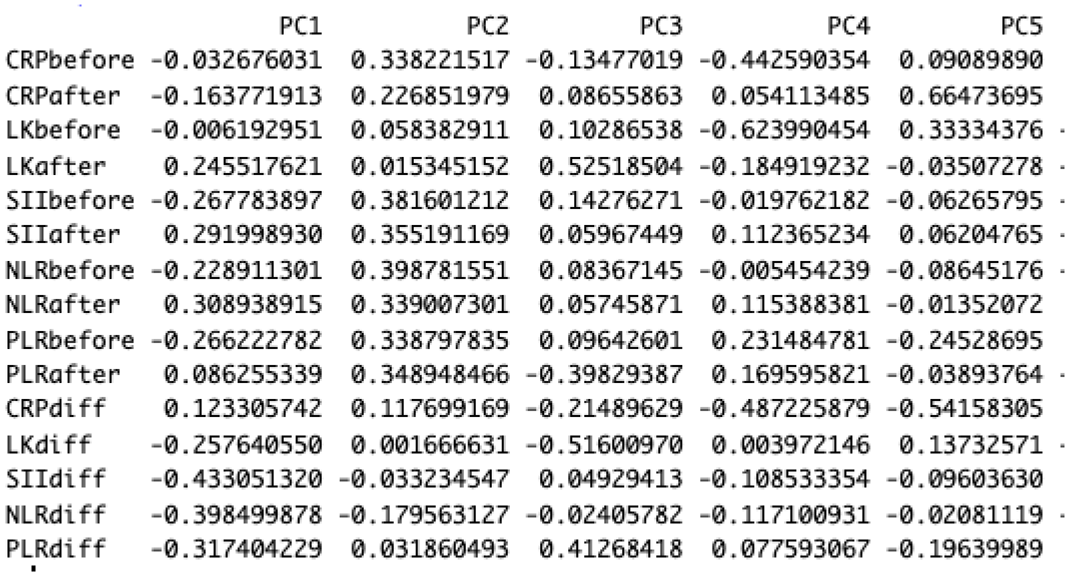


Figure B.370: Biplot plot of the inflammation variables, n=90 (Disease control (n=71), Progressive disease (n=19)).



```
                   PC1          PC2         PC3          PC4         PC5
CRPbefore -0.032676031  0.338221517 -0.13477019 -0.442590354  0.09089890
CRPafter  -0.163771913  0.226851979  0.08655863  0.054113485  0.66473695
LKbefore  -0.006192951  0.058382911  0.10286538 -0.623990454  0.33334376
LKafter    0.245517621  0.015345152  0.52518504 -0.184919232 -0.03507278
SIIbefore -0.267783897  0.381601212  0.14276271 -0.019762182 -0.06265795
SIIafter   0.291998930  0.355191169  0.05967449  0.112365234  0.06204765
NLRbefore -0.228911301  0.398781551  0.08367145 -0.005454239 -0.08645176
NLRafter   0.308938915  0.339007301  0.05745871  0.115388381 -0.01352072
PLRbefore -0.266222782  0.338797835  0.09642601  0.231484781 -0.24528695
PLRafter   0.086255339  0.348948466 -0.39829387  0.169595821 -0.03893764
CRPdiff    0.123305742  0.117699169 -0.21489629 -0.487225879 -0.54158305
LKdiff    -0.257640550  0.001666631 -0.51600970  0.003972146  0.13732571
SIIdiff   -0.433051320 -0.033234547  0.04929413 -0.108533354 -0.09603630
NLRdiff   -0.398499878 -0.179563127 -0.02405782 -0.117100931 -0.02081119
PLRdiff   -0.317404229  0.031860493  0.41268418  0.077593067 -0.19639989
```

Figure B.371: Loading matrix of the inflammation variables of the first 5 principal components, n=90 (Disease control (n=71), Progressive disease (n=19)).

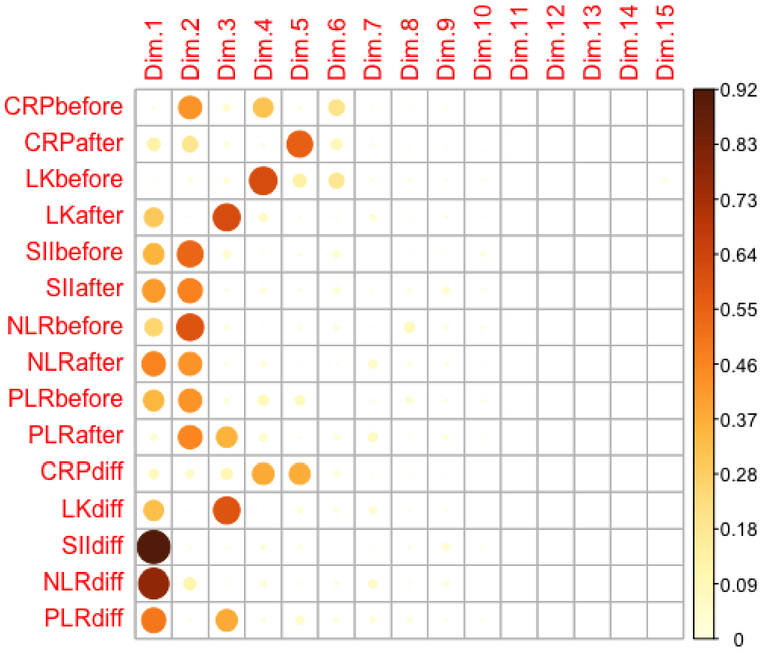


Figure B.372: Correlation of the original inflammation variables with the principal components, n=90 (Disease control (n=71), Progressive disease (n=19)).

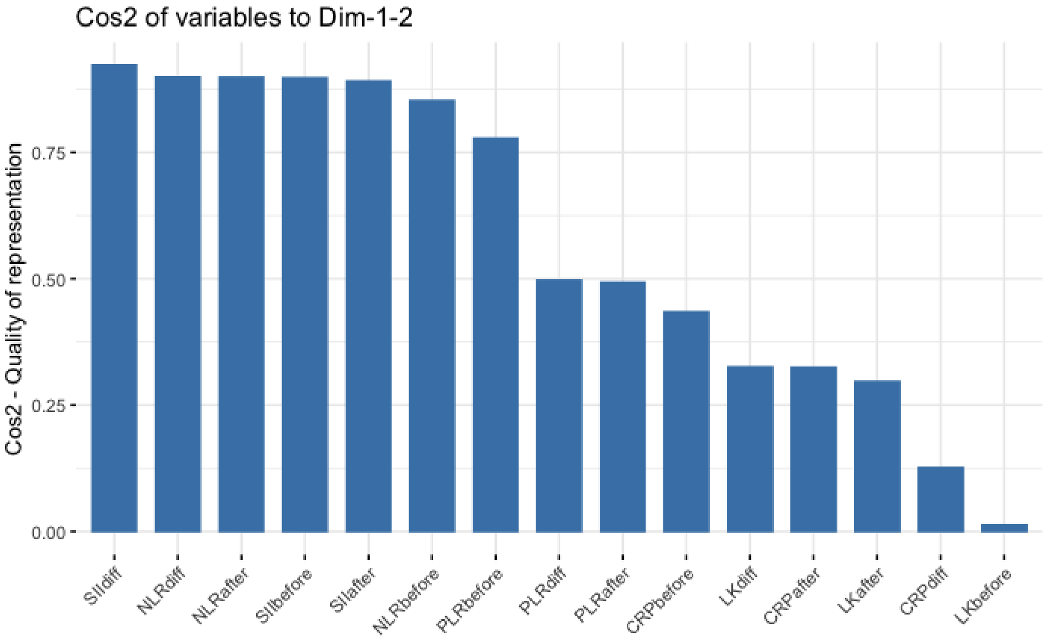


Figure B.373: Cos2 bar chart of the original inflammation variables, n=90 (Disease control (n=71), Progressive disease (n=19)).

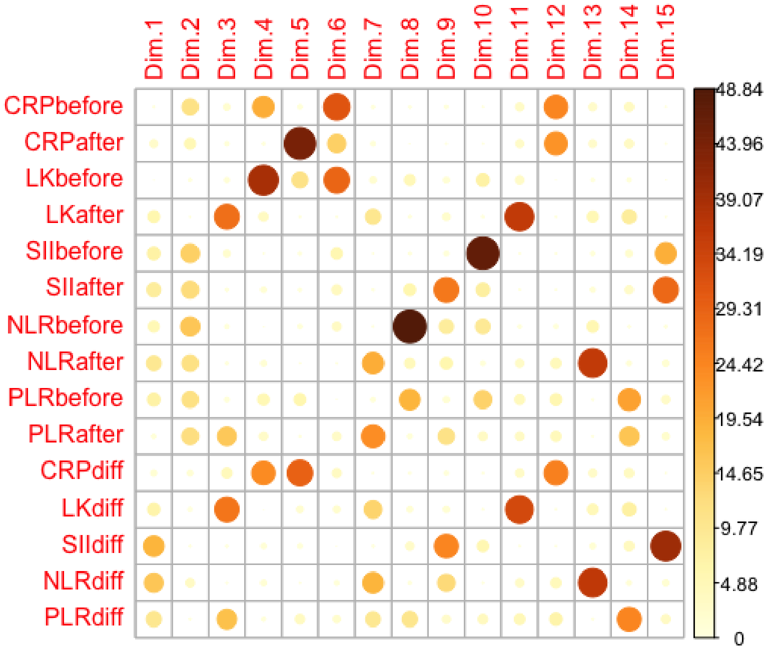


Figure B.374: Contribution of the original inflammation variables to the principal components, n=90 (Disease control (n=71), Progressive disease (n=19)).

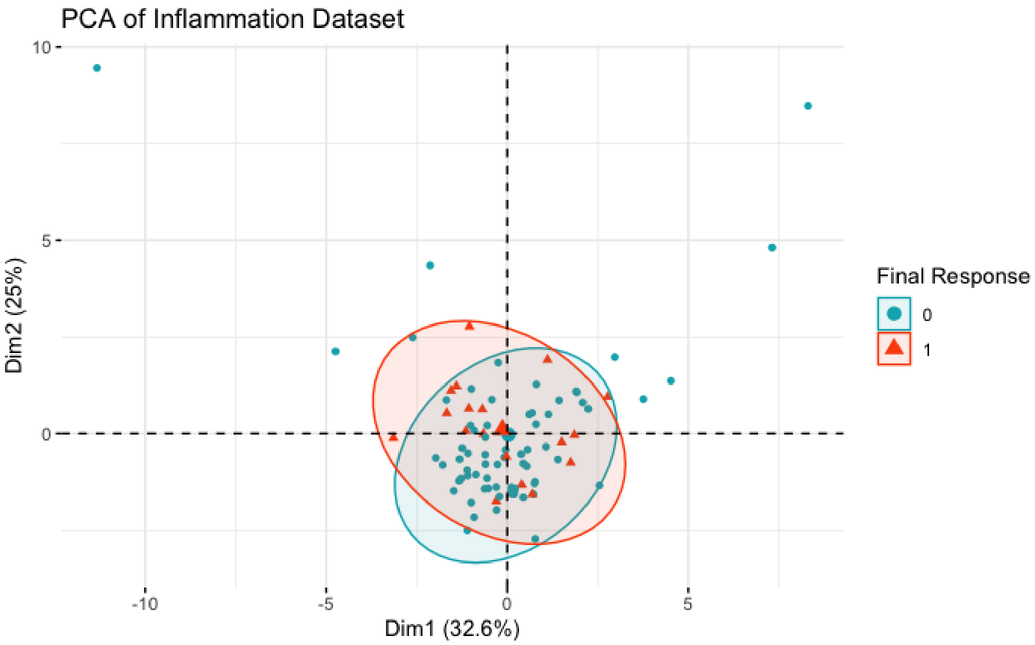


Figure B.375: PCA plot of the inflammation group with 95% prediction ellipses, n=90 (Disease control (n=71), Progressive disease (n=19)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

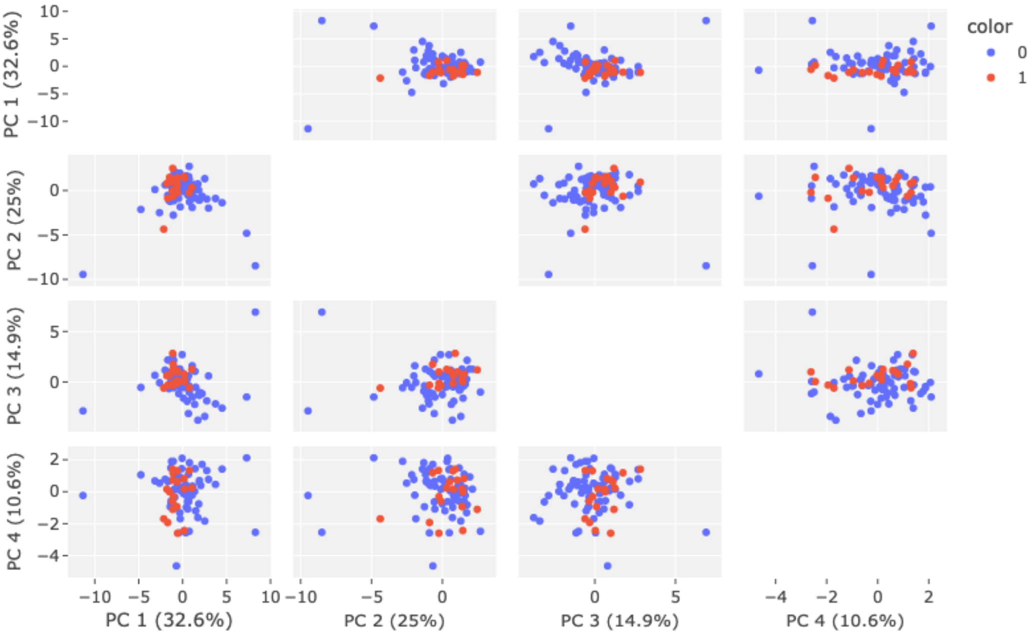


Figure B.376: Scatter matrix plot of the first four principal components, n=90 (Disease control (n=71), Progressive disease (n=19)). The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the inflammation variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

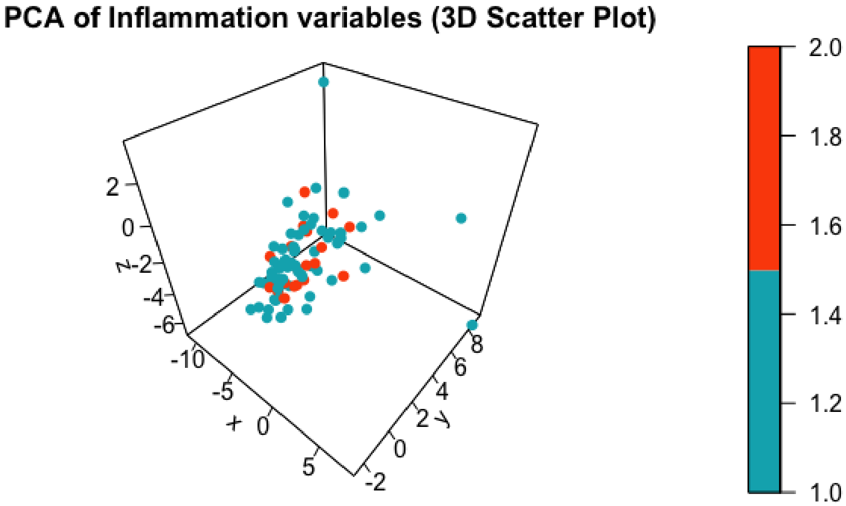


Figure B.377: 3D plot of the inflammation variables projected onto the first three principal components, n=90 (Disease control (n=71), Progressive disease (n=19)).

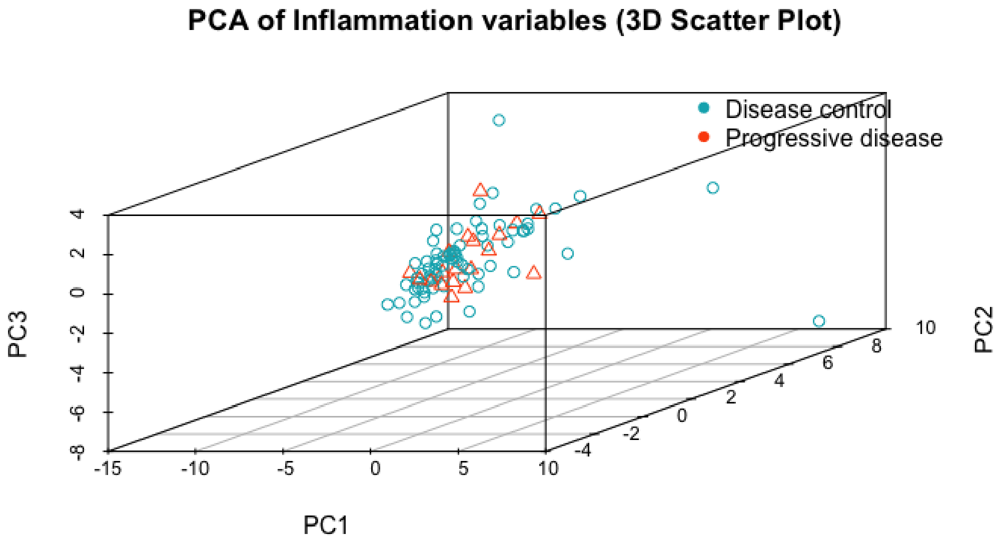


Figure B.378: 3D plot of the inflammation variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=90 (Disease control (n=71), Progressive disease (n=19)).

given in equation (B.2). Interestingly, age does not show a significant correlation with any other variable and displays a slightly negative correlation with height and weight. The exact correlation coefficients are provided in Figure B.380.

PCA is then applied to this small dataset, and the standard deviation, proportion of variance explained, as well as cumulative proportion are presented in Table B.102. Using 3 PCs, almost all the variation in the data can be accounted for. This observation is not surprising since height and weight are used to compute BMI and there are only four variables. Therefore, it would suffice to only consider BMI and age as variables. The screeplot shown in Figure B.381 clearly indicates that PC1 explains over half of the total variation, while PC2 and PC3 account for the other half. The correlation matrix between the original variables and PCs, illustrated in Figure B.386, indicates that PC1 is extremely positively correlated with weight and relatively strongly correlated with BMI. On the contrary, PC2 displays no significant correlations, whereas PC3 exhibits a strong correlation with age, which is consistent with Figure B.387. Finally, the contributions of each original variable to each principal component, displayed in Figure B.388, reveal that PC1 is mainly influenced by weight, PC2 has contributions from height, age and BMI, PC3 is mostly influenced by age, and PC4 is defined by weight.

To further analyze the patient characteristics dataset, several visualizations were created. The scatter plot in Figure B.382 displays the scores of the observations, but does not reveal clear groupings or trends. Additionally, observations close to the origin are poorly represented in the first two dimensions, similar to previous scatter plots. The loading plot in Figure B.383 illustrates that BMI, weight, and height are well-represented by the first two principal components, whereas age is poorly represented. This is not surprising, as earlier results showed that age contributed most to PC3, resulting in loss of information when projected onto PC1 and PC2. Moreover, the biplot in Figure B.384 shows no strong correlations between the observations (black dots) and the original variables. The scores appear scattered in a cloud around the origin. The loading matrix in Figure B.385 confirms the earlier findings. Finally, the plot in Figure B.389 shows no clear deviation in the two groups based on the final response. Neither does projecting the data onto the first four dimensions provide any valuable insights. Nor do the 3D-plots in Figure B.391 and Figure B.392 provide any more visual insights of the scores projected on the space defined by the first three principal components.

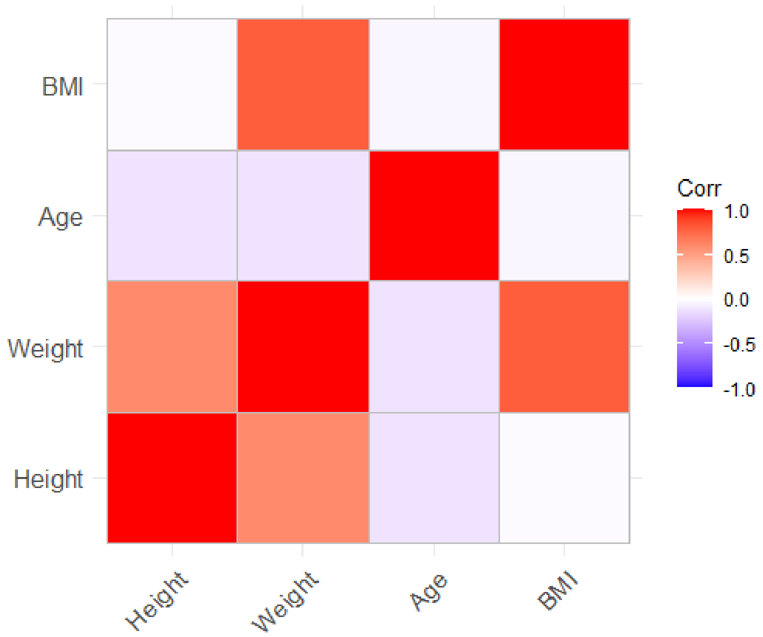
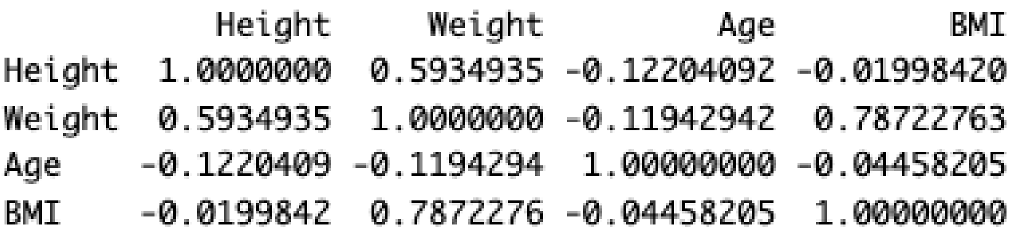


Figure B.379: Correlation Matrix of the various variables in the Patient Characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)). Red = Positive correlation, Blue = negative correlation, White = No correlation.



```
            Height     Weight         Age         BMI
Height   1.0000000  0.5934935 -0.12204092 -0.01998420
Weight   0.5934935  1.0000000 -0.11942942  0.78722763
Age     -0.1220409 -0.1194294  1.00000000 -0.04458205
BMI     -0.0199842  0.7872276 -0.04458205  1.00000000
```

Figure B.380: Correlation Matrix of the various variables in the Patient Characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)).

Patient Characteristics	**PC1**	**PC2**	**PC3**	**PC4**
Standard Deviation	1.415	1.035	0.961	0.066
Proportion of Variance Explained	0.501	0.268	0.231	0.001
Cumulative Proportion	0.501	0.768	0.999	1.000

Table B.102: PCA summary values of the patient characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first four principal components. Note that there are only 4 PCs as the data only has 4 variables.

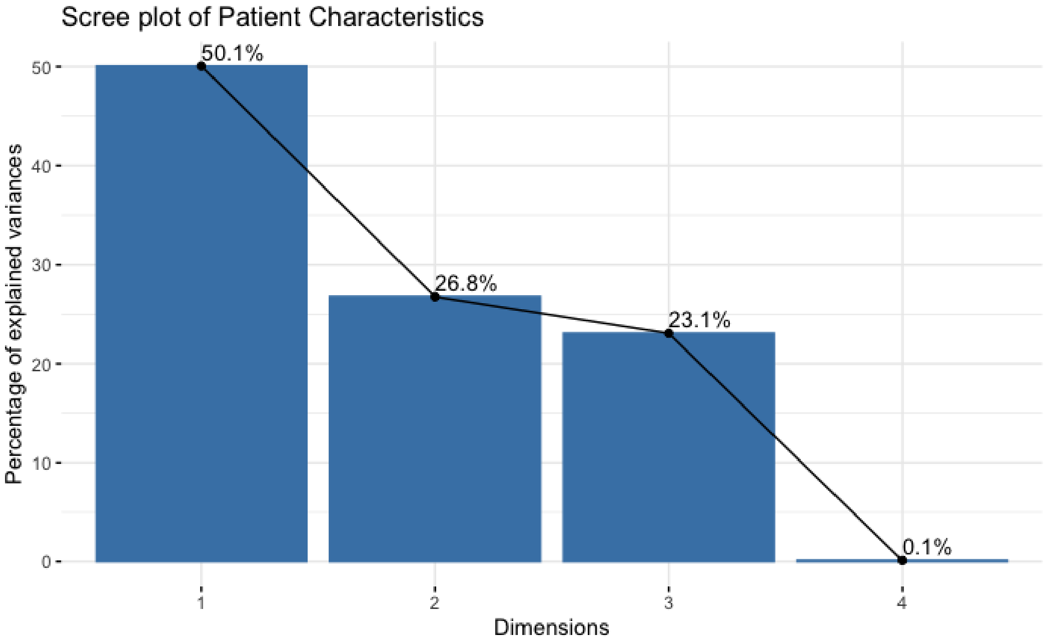


Figure B.381: Scree plot of the Patient Characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)).

B.4.7 Tumor Markers CEA and CA19-9

The tumor markers CEA and CA19-9 are essential in assessing the status and treatment of PDAC. Consequently, it is of interest to examine which values of these tumor markers are most critical in determining the final response to chemotherapy. The values of CEA and CA19-9 at diagnosis, before chemotherapy, after chemotherapy, and the difference between the values before and after the first chemotherapy treatment are the variables under consideration and together form the tumor markers dataset. The dataset consists of 96 observations with 8 explanatory variables. The correlation matrix presented in Figure B.393 shows that all the values at diagnosis, before and after are strongly positively correlated for each of the individual tumor markers. However, the difference of each tumor marker seems to exhibit no correlation or a slightly negative correlation with any of the diagnosis/before/after values. Detailed correlation coefficients are available in Figure B.394. After conducting PCA on the variables, it is observed that the first two principal components explain roughly 83% of the total variation in the data. Adding a third dimension raises the total to almost 95%, as depicted in Table B.103 and Figure B.395. This implies that the tumor marker data is organized in a way that can be efficiently represented using only a few dimensions, with the first two components capturing most of the variation. The correlation matrix between the original tumor marker variables and principal components, as presented in Figure B.400, demonstrates that the first dimension (PC1) is highly correlated with the CEA and CA19-9 values at diagnosis, before, and after the first chemotherapy treatment. The second dimension is highly correlated with CEAdiff, and the third dimension is correlated with CA19-9 diff. Nonetheless, due to the high correlation between variables, it is difficult to disentangle their individual contributions to the principal components. This results in the correlations between the original variables and principal components being high, while the contribution of each individual variable to those components is relatively low. It is evident from Figure B.402 that PC2 is mainly determined by the CEAdiff, and PC3 is mostly influenced by the CA19-9 diff.

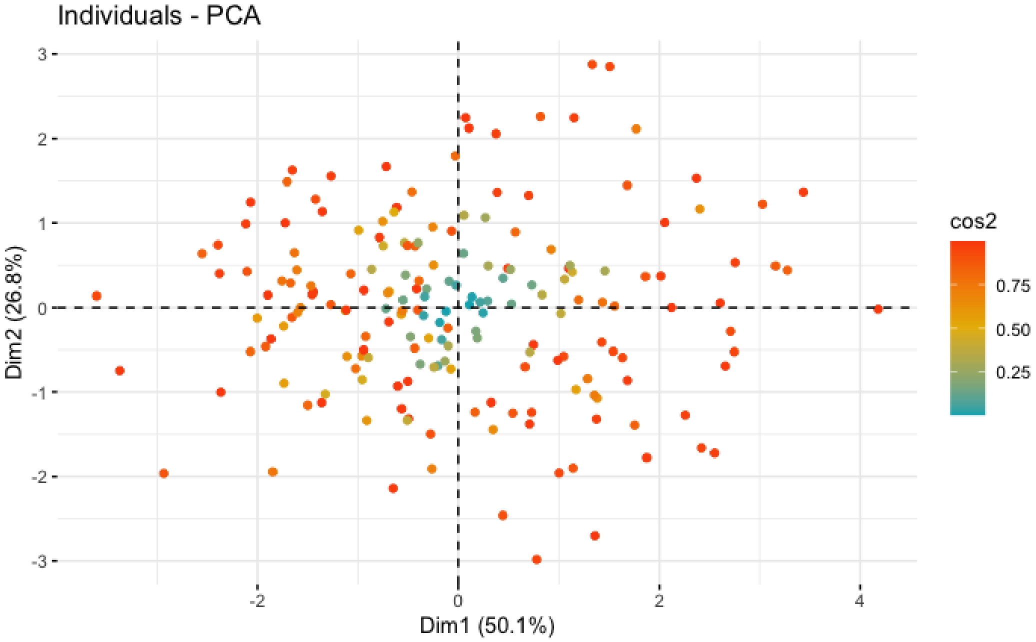


Figure B.382: Scatter plot of the Patient characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)).

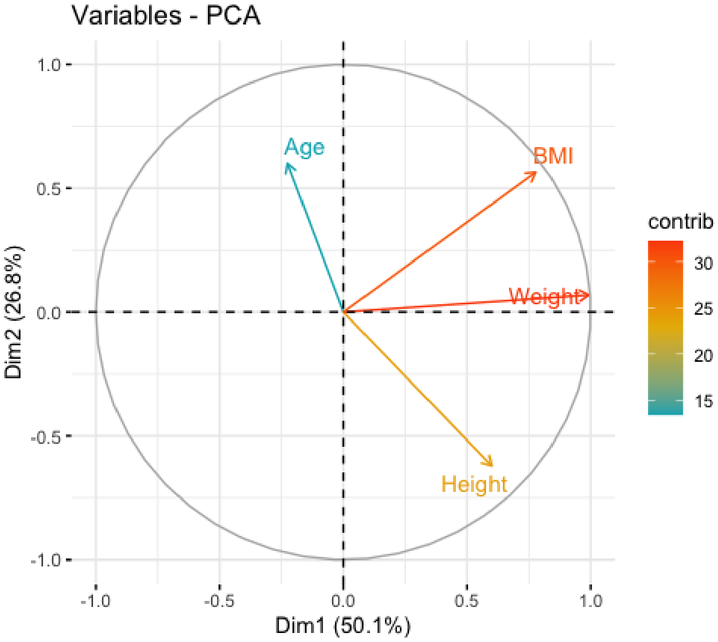


Figure B.383: Loading plot of the patient characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

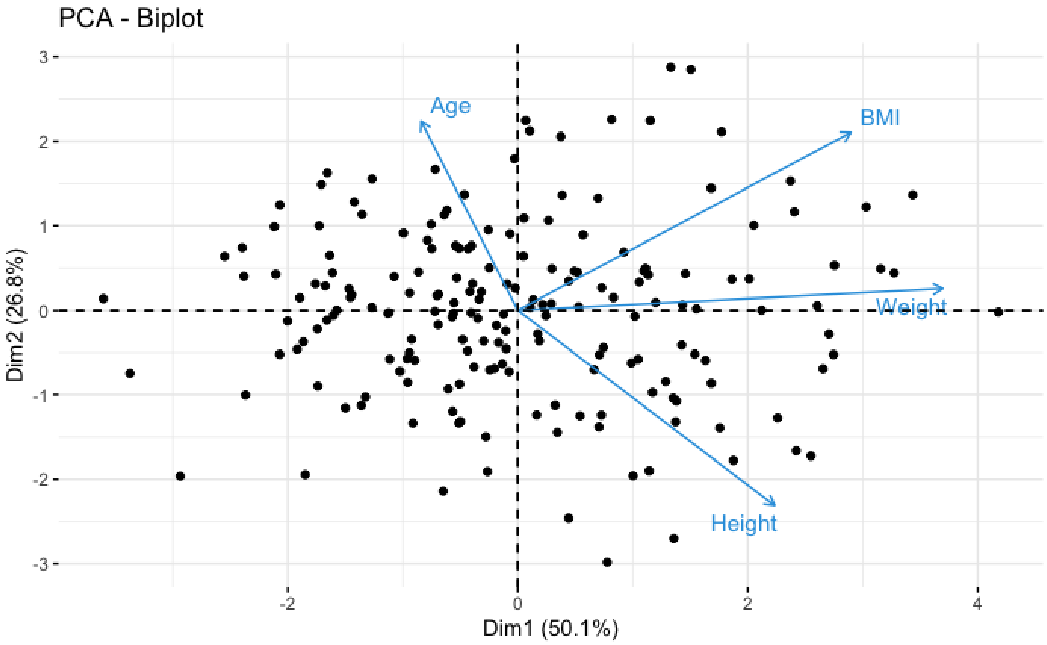
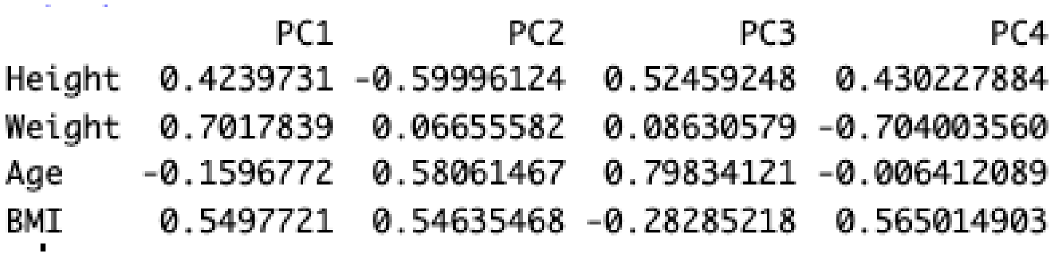


Figure B.384: Biplot plot of the patient characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)).



```
             PC1         PC2         PC3          PC4
Height  0.4239731 -0.59996124  0.52459248  0.430227884
Weight  0.7017839  0.06655582  0.08630579 -0.704003560
Age    -0.1596772  0.58061467  0.79834121 -0.006412089
BMI     0.5497721  0.54635468 -0.28285218  0.565014903
```

Figure B.385: Loading matrix of the patient characteristic variables of the first 4 principal components, n=189 (Disease control (n=151), Progressive disease (n=38)).

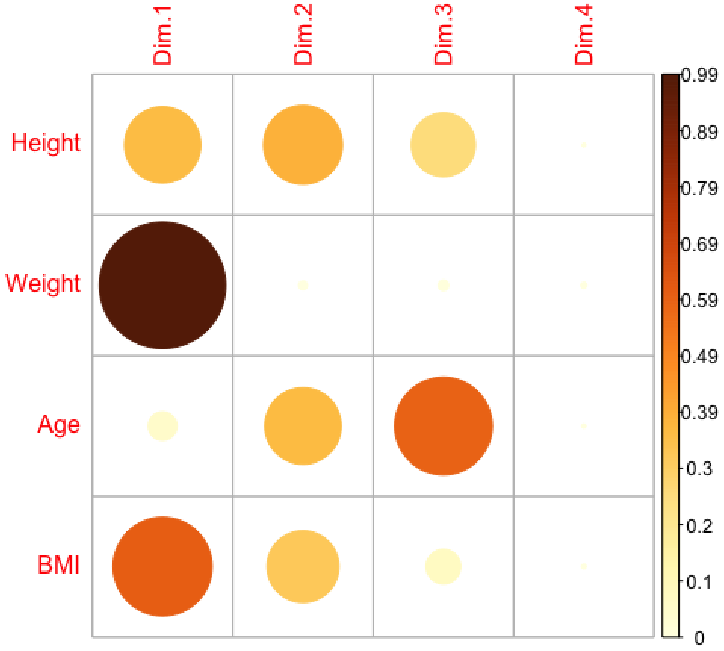


Figure B.386: Correlation of the original patient characteristic variables with the principal components, n=189 (Disease control (n=151), Progressive disease (n=38)).

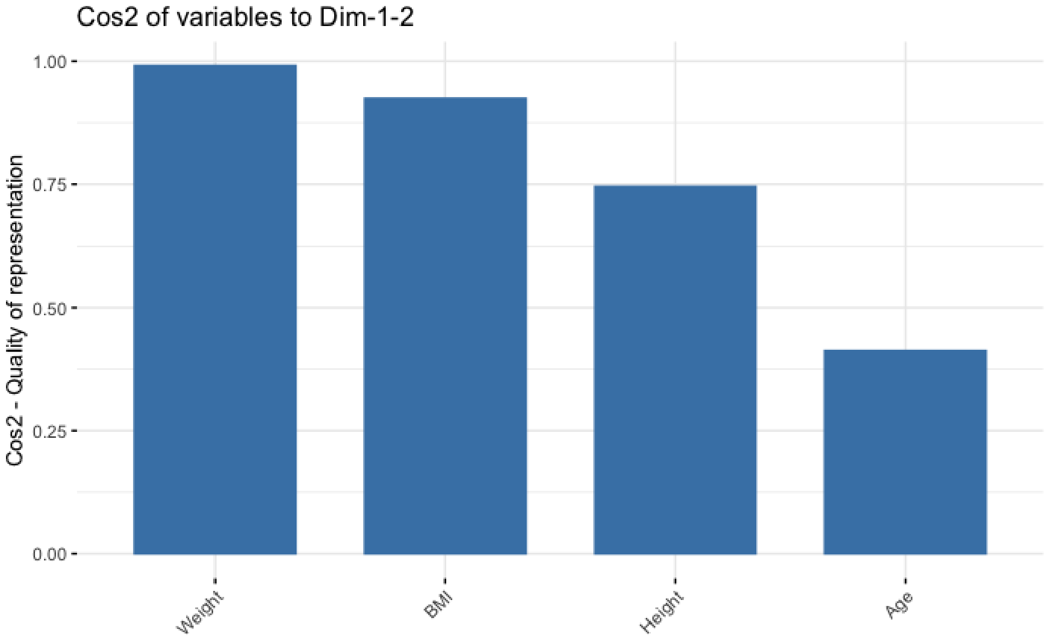


Figure B.387: Cos2 bar chart of the original patient characteristics, n=189 (Disease control (n=151), Progressive disease (n=38)).

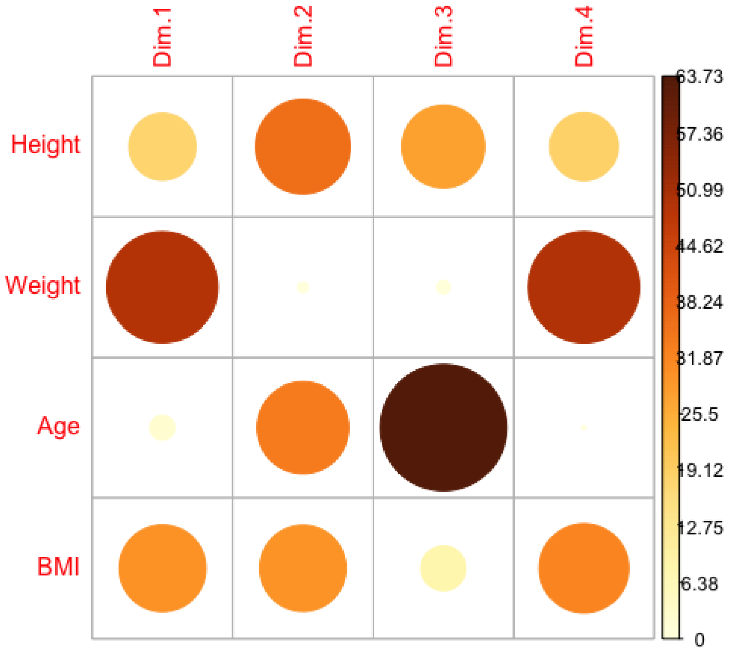


Figure B.388: Contribution of the original patient characteristics to the principal components, n=189 (Disease control (n=151), Progressive disease (n=38)).

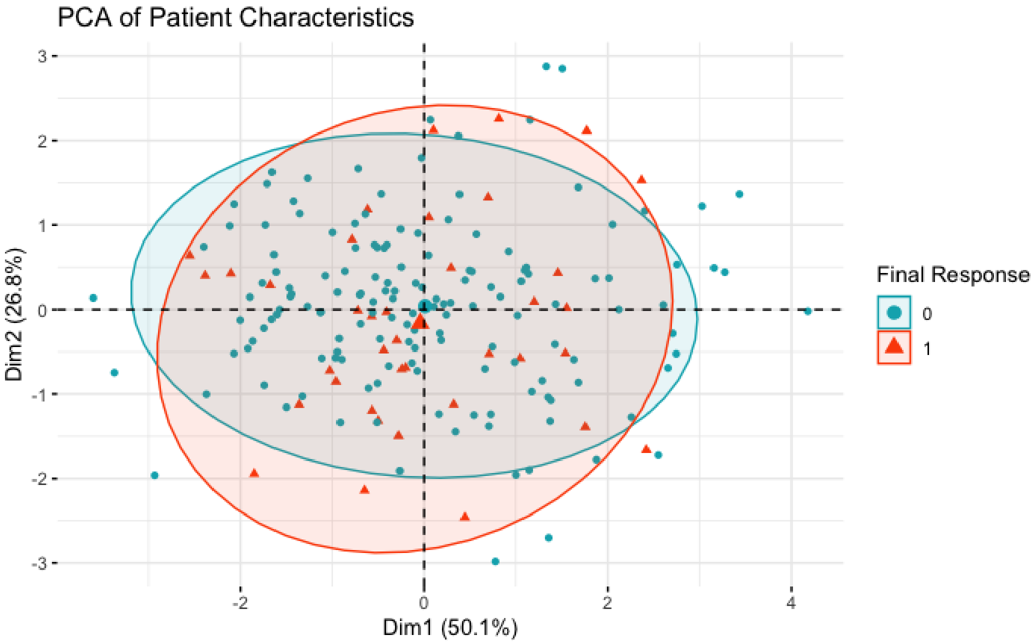


Figure B.389: PCA plot of the patient characteristics with 95% prediction ellipses, n=189 (Disease control (n=151), Progressive disease (n=38)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

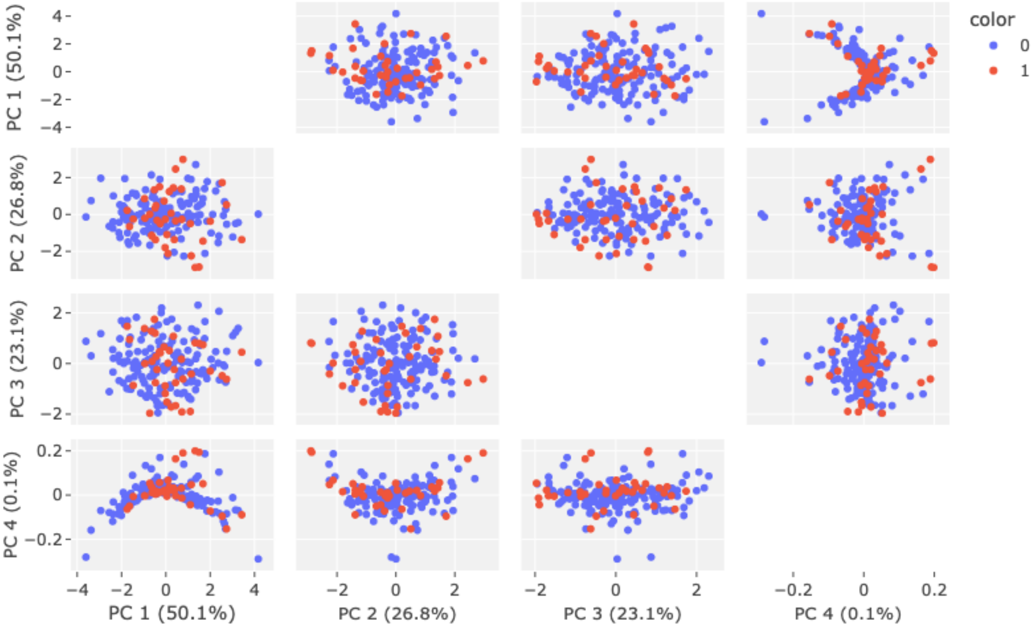


Figure B.390: Scatter matrix plot of the first four principal components, n=189 (Disease control (n=151), Progressive disease (n=38)). The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the patient characteristics with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

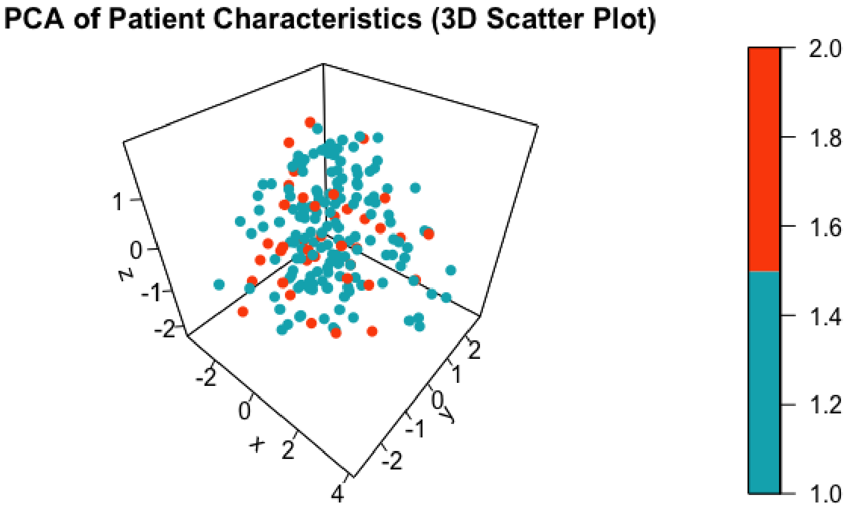


Figure B.391: 3D plot of the patient characteristics projected onto the first three principal components, n=189 (Disease control (n=151), Progressive disease (n=38)).

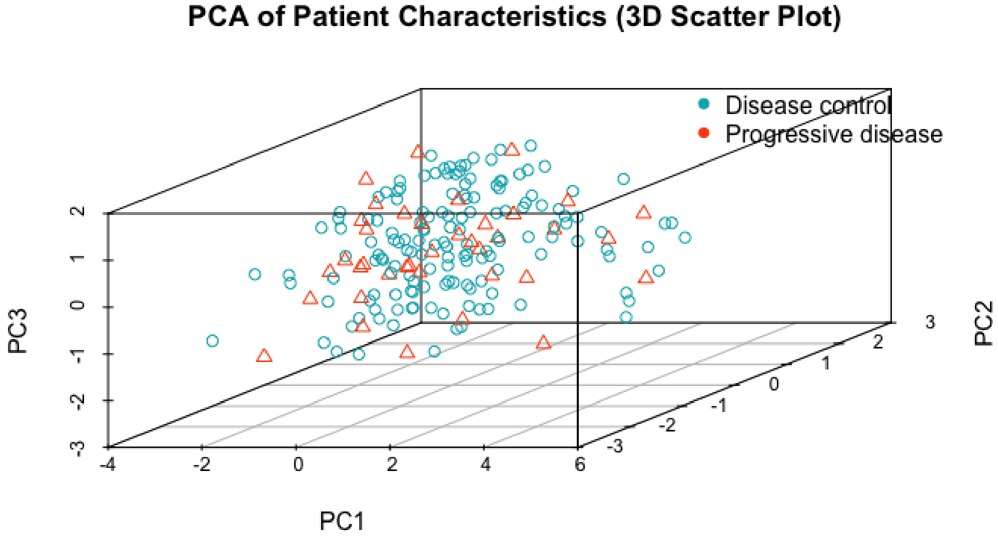


Figure B.392: 3D plot of the patient characteristics with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=189 (Disease control (n=151), Progressive disease (n=38)).

To continue, the scatter plot in Figure B.396 exhibits a concentration of scores around the origin with the scores located more towards the left being better represented in the plot. The presence of outliers is responsible for the large scale of the plot. As observed in the loading plot in Figure B.397, the PC1 dimension displays a highly positive correlation with the tumor marker values at diagnosis, before and after, while PC2 is negatively correlated with CEAdiff. As CA19-9 is primarily captured by PC3, it is poorly represented in the plot. The biplot in Figure B.398 fails to reveal any clear patterns between the loading vectors of the original variables and the scores of the observations. The 95% prediction ellipse plots in Figure B.403 and Figure B.404 show that many of the disease control patients are captured by a negative PC1 value and elongated along PC2. The ellipse of the disease control group is comparatively smaller and confounded within the ellipse of the progressive disease group. Since both ellipses display greater elongation along PC2, it indicates that the variance is significantly greater in the PC2 dimension than in the others. Nevertheless, a clear division between the two groups is absent. Likewise, the scores in a 2D plot in the first four principal components, as shown in Figure B.390 does not provide any more insights into a separation between the final response groups. Lastly, the three-dimensional plots in Figure B.406 and Figure B.407 provide a visual representation of the scores in the first three principal components with no clear distinguishment either.

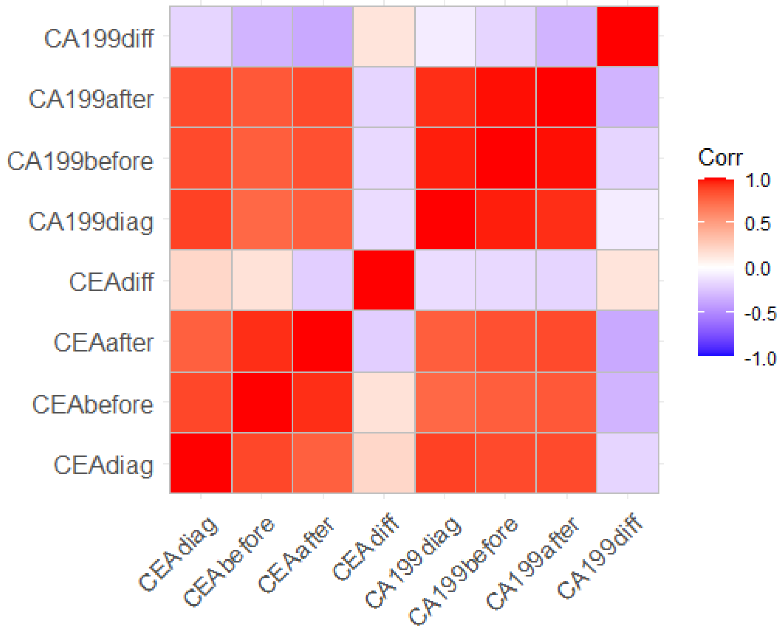
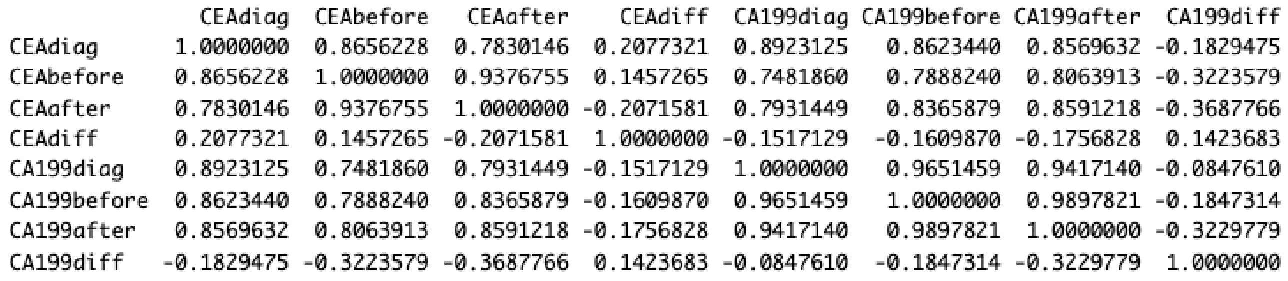


Figure B.393: Correlation Matrix of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)). Red = Positive correlation, Blue = negative correlation, White = No correlation.



```
             CEAdiag  CEAbefore   CEAafter    CEAdiff  CA199diag CA199before CA199after  CA199diff
CEAdiag    1.0000000  0.8656228  0.7830146  0.2077321  0.8923125   0.8623440  0.8569632 -0.1829475
CEAbefore  0.8656228  1.0000000  0.9376755  0.1457265  0.7481860   0.7888240  0.8063913 -0.3223579
CEAafter   0.7830146  0.9376755  1.0000000 -0.2071581  0.7931449   0.8365879  0.8591218 -0.3687766
CEAdiff    0.2077321  0.1457265 -0.2071581  1.0000000 -0.1517129  -0.1609870 -0.1756828  0.1423683
CA199diag  0.8923125  0.7481860  0.7931449 -0.1517129  1.0000000   0.9651459  0.9417140 -0.0847610
CA199before 0.8623440  0.7888240  0.8365879 -0.1609870  0.9651459   1.0000000  0.9897821 -0.1847314
CA199after 0.8569632  0.8063913  0.8591218 -0.1756828  0.9417140   0.9897821  1.0000000 -0.3229779
CA199diff -0.1829475 -0.3223579 -0.3687766  0.1423683 -0.0847610  -0.1847314 -0.3229779  1.0000000
```

Figure B.394: Correlation Matrix of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)).

Tumor Markers	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	2.324	1.105	0.982	0.563	0.289
Proportion of Variance Explained	0.675	0.153	0.121	0.040	0.010
Cumulative Proportion	0.675	0.828	0.948	0.988	0.998

Table B.103: PCA summary values of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)). It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components. .

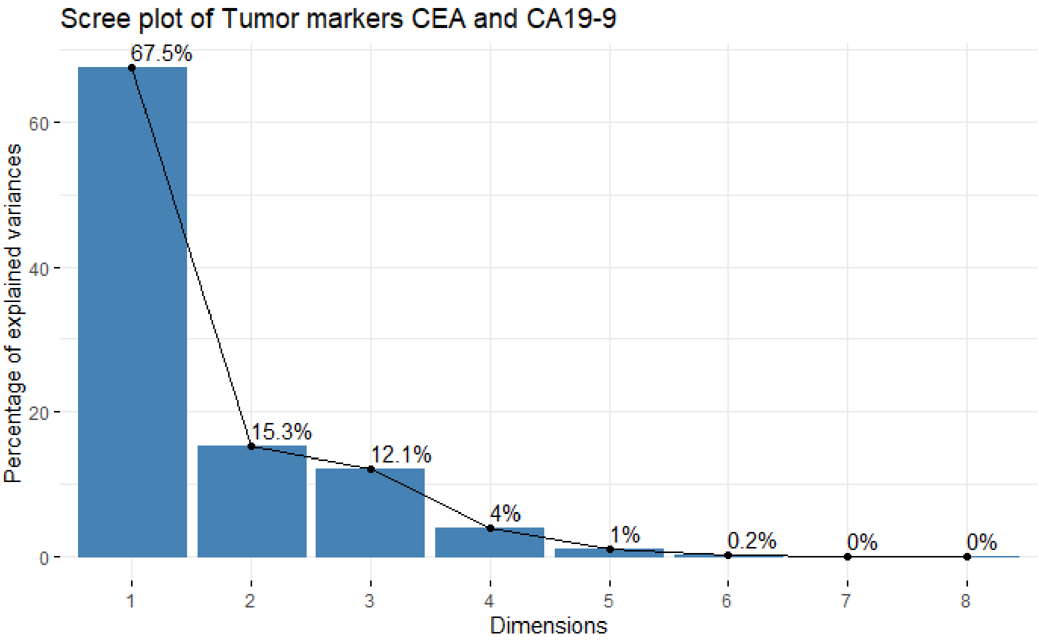


Figure B.395: Scree plot of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)).

B.4.8 All blood and tumor markers

Given that the grouping of blood markers has been based on expert opinion, it may be valuable to evaluate all blood and tumor markers together using PCA. This would allow us to identify which variables are selected by PCA, determine the significance of the original variables, and establish the number of principal components necessary to capture a substantial portion of the total variation in the data. The considered dataset comprises 59 observations and 34 explanatory variables, with missing variables removed. The correlation matrix of all variables and tumor markers is presented in Figure B.408. Interestingly, few variables demonstrate positive correlations (aside from before and after values of the same variables), with a clear negative correlation observed between the before values of GFR and CR as well. After performing PCA on all before and after values of blood variables, along with CEA and CA19-9 tumor markers, we observe from the loading matrix in Figure B.413 that the proportion of variance explained by the first 10 principal components accounts for nearly 80% of the total variation in the data. This suggests that reducing the data from a 34-dimensional space to a 10 or 11-dimensional space is adequate for retaining about 80% of the total variation. The screeplot in Figure B.409 indicates that the first two dimensions explain a small portion of the total variation, which can be attributed to the presence of noise or outliers, complex underlying relationships between variables, and high variable correlations.

Therefore, the scatterplot, loading plot, biplot, and prediction ellipse plot in Figure B.410, Figure B.411, Figure B.412, Figure B.417, respectively, are intended to provide a visual representation of the first two principal component spaces and should be viewed primarily as a means of facilitating intuition. It should be noted that these plots are limited in their ability to convey meaningful information, as the first two principal components only explain 27.5% of the total variation. In addition, the correlation between each original variable and the principal components is illustrated in Figure B.414,

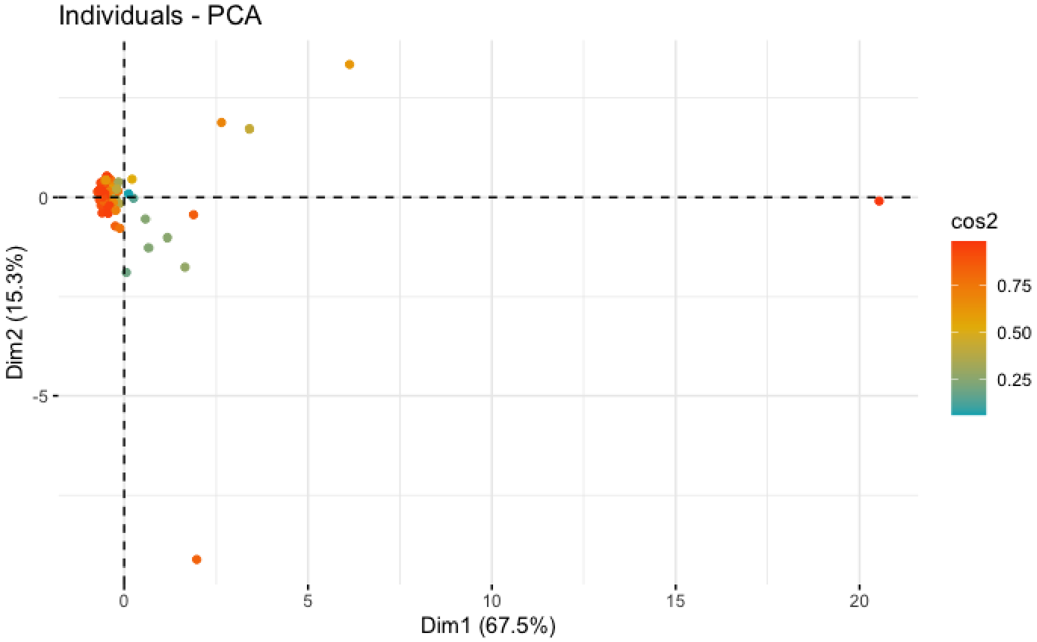


Figure B.396: Scatter plot of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)).

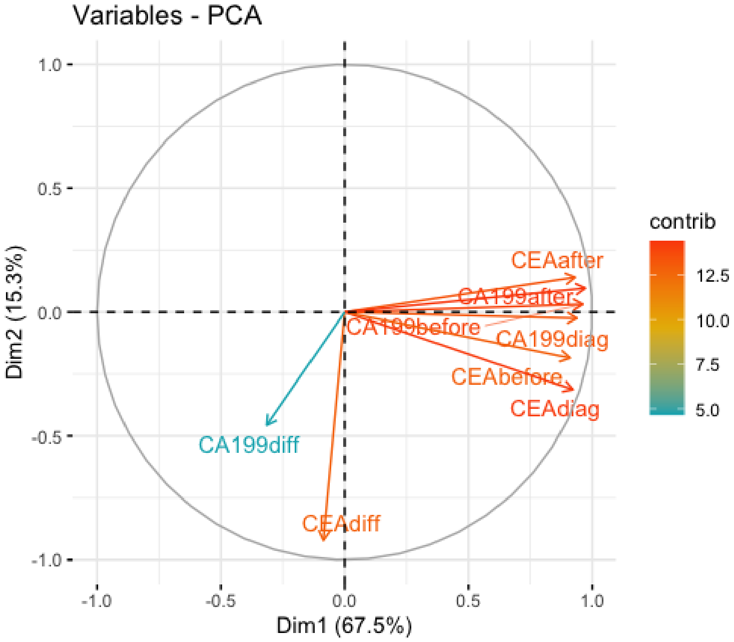


Figure B.397: Loading plot of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

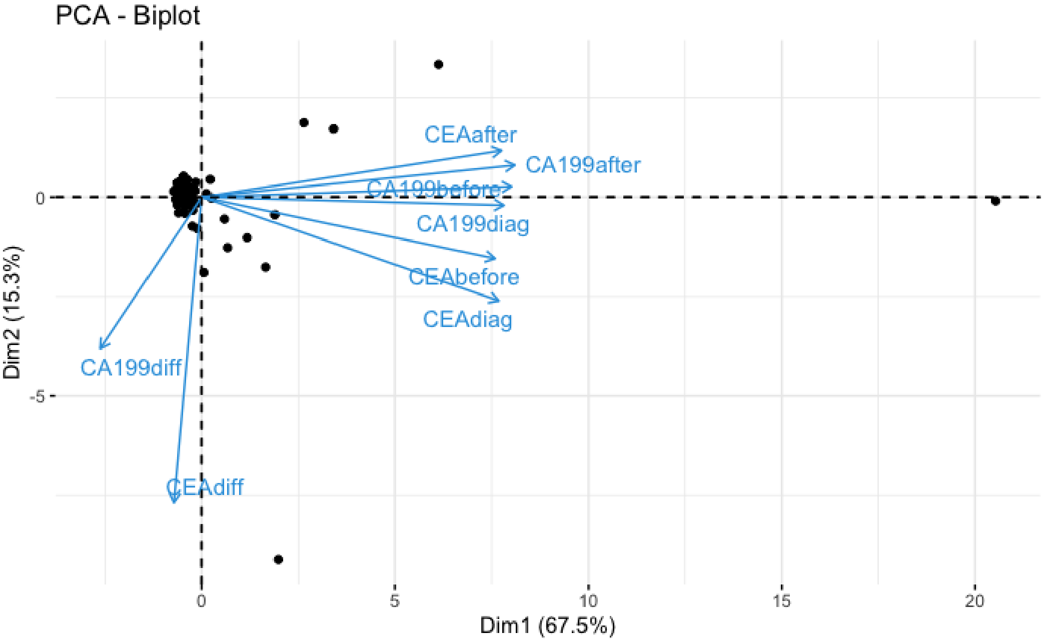
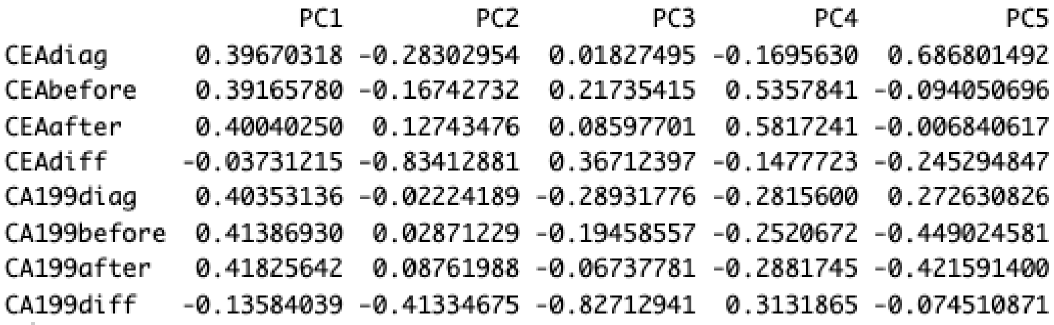


Figure B.398: Biplot plot of the tumor markers CEA and CA19-9, n=96 (Disease control (n=76), Progressive disease (n=20)).



```
                   PC1         PC2         PC3        PC4          PC5
CEAdiag     0.39670318 -0.28302954  0.01827495 -0.1695630  0.686801492
CEAbefore   0.39165780 -0.16742732  0.21735415  0.5357841 -0.094050696
CEAafter    0.40040250  0.12743476  0.08597701  0.5817241 -0.006840617
CEAdiff    -0.03731215 -0.83412881  0.36712397 -0.1477723 -0.245294847
CA199diag   0.40353136 -0.02224189 -0.28931776 -0.2815600  0.272630826
CA199before 0.41386930  0.02871229 -0.19458557 -0.2520672 -0.449024581
CA199after  0.41825642  0.08761988 -0.06737781 -0.2881745 -0.421591400
CA199diff  -0.13584039 -0.41334675 -0.82712941  0.3131865 -0.074510871
```

Figure B.399: Loading matrix of the tumor variables of the first 5 principal components, n=96 (Disease control (n=76), Progressive disease (n=20)).

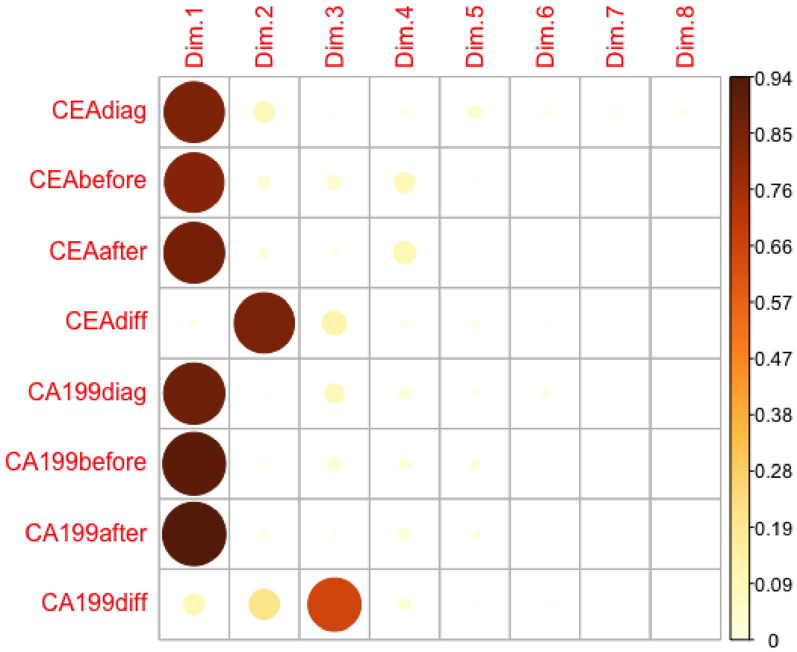


Figure B.400: Correlation of the original tumor marker variables with the principal components, n=96 (Disease control (n=76), Progressive disease (n=20)).

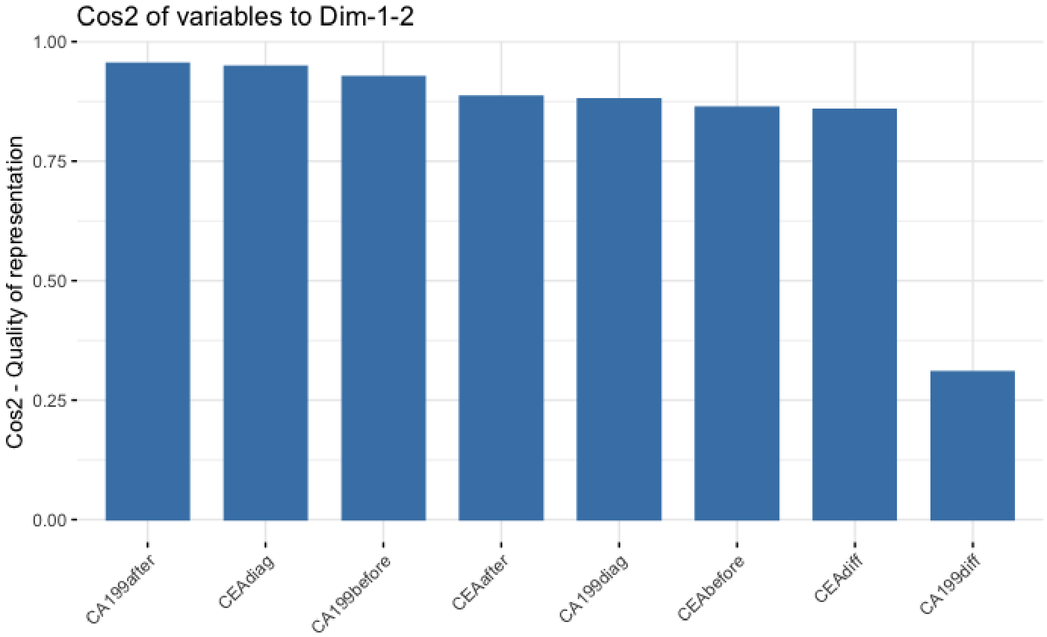


Figure B.401: Cos2 bar chart of the original tumor marker variables, n=96 (Disease control (n=76), Progressive disease (n=20)).

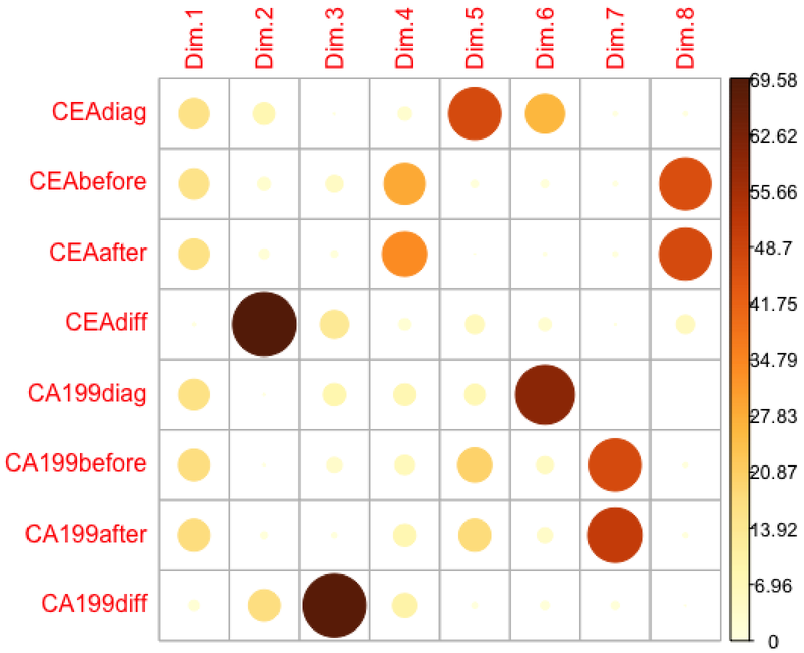


Figure B.402: Contribution of the original tumor marker variables to the principal components, n=96 (Disease control (n=76), Progressive disease (n=20)).

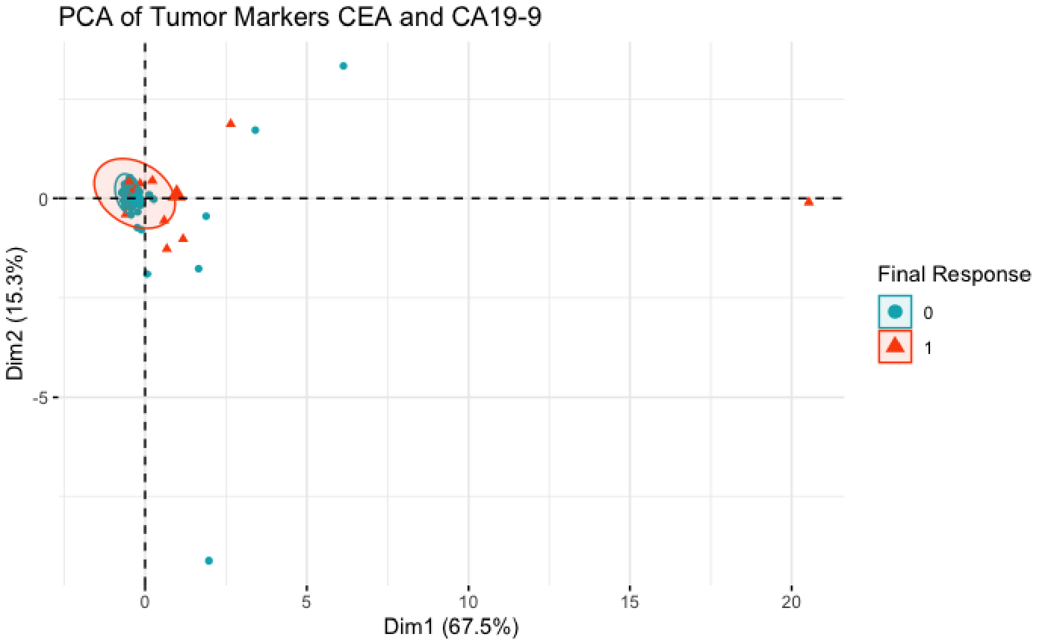


Figure B.403: PCA plot of the tumor markers CEA and CA19-9 with 95% prediction ellipses, n=96 (Disease control (n=76), Progressive disease (n=20)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

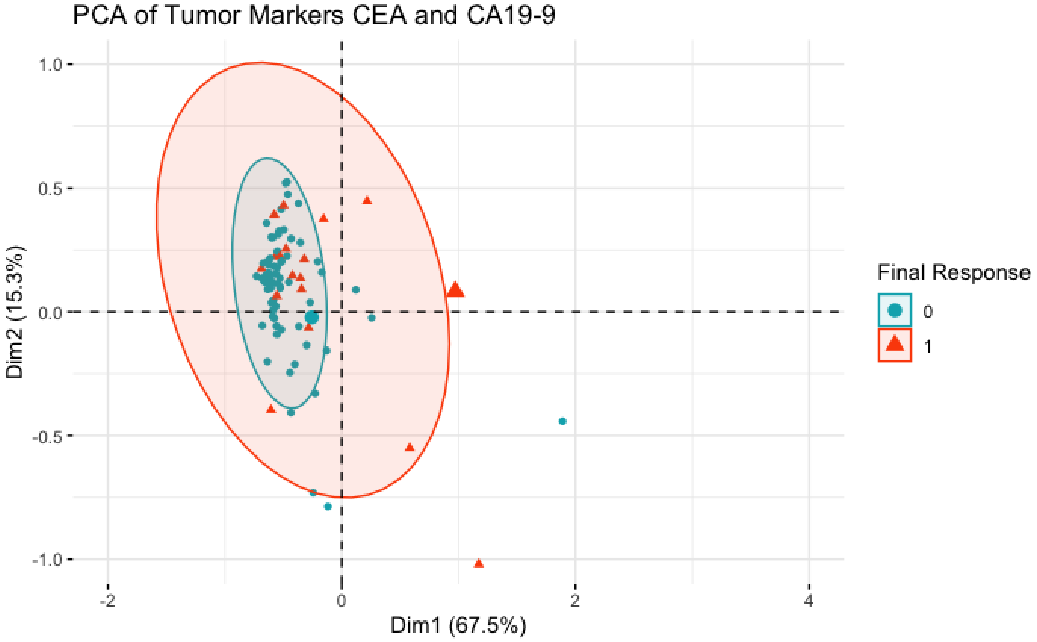


Figure B.404: PCA plot of the tumor markers CEA and CA19-9 with 95% prediction ellipses zoomed in, n=96 (Disease control (n=76), Progressive disease (n=20)).

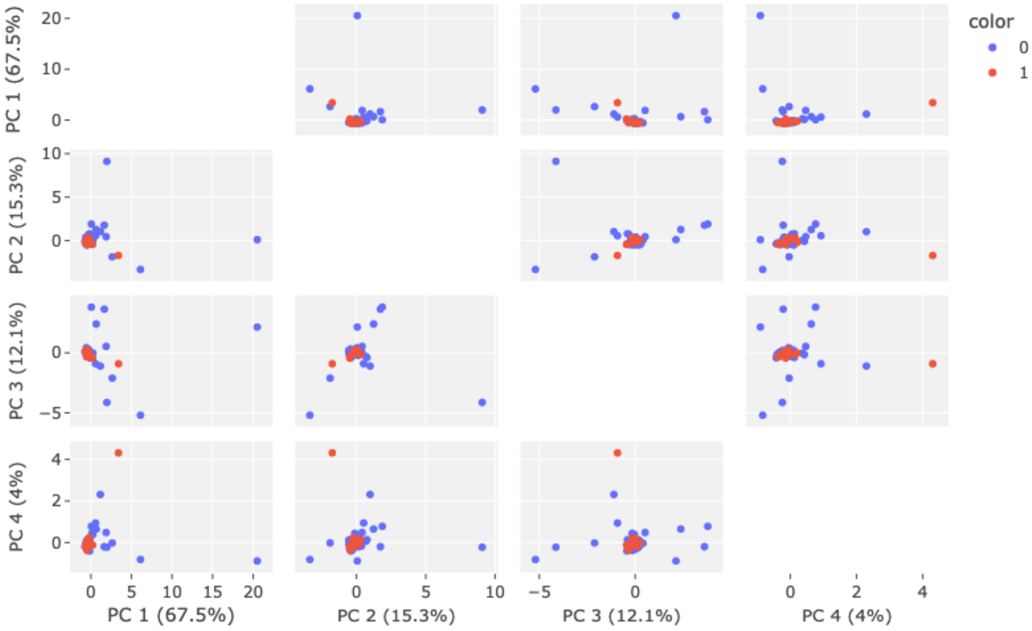


Figure B.405: Scatter matrix plot of the first four principal components. The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing the tumor markers with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease), n=96 (Disease control (n=76), Progressive disease (n=20)).

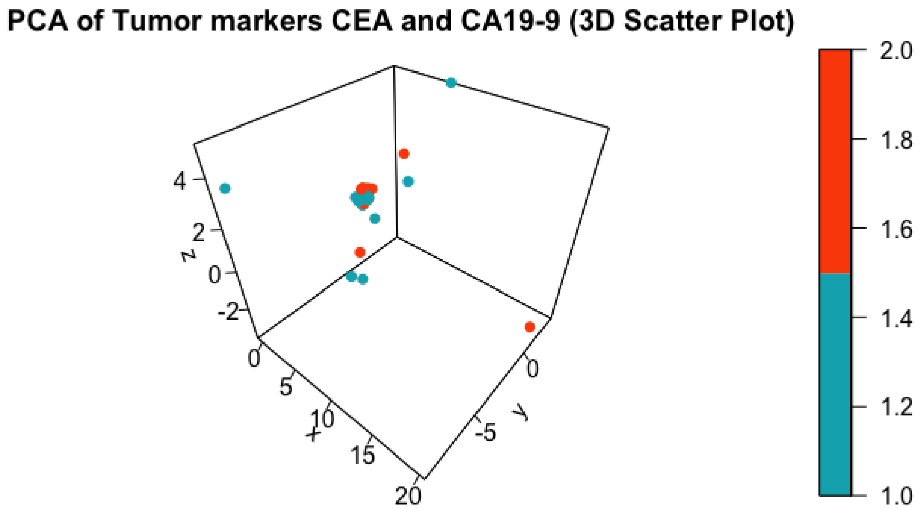


Figure B.406: 3D plot of the tumor marker variables projected onto the first three principal components, n=96 (Disease control (n=76), Progressive disease (n=20)).

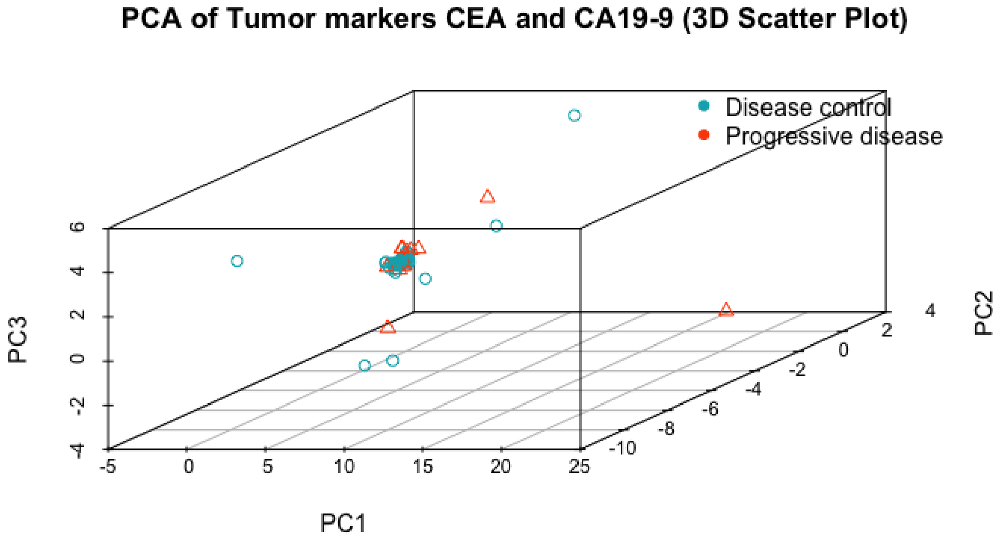


Figure B.407: 3D plot of the tumor marker variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=96 (Disease control (n=76), Progressive disease (n=20)).

while the variables that are best represented in PC1 and PC2 are displayed in Figure B.415. Additionally, Figure B.416 depicts the contributions of each original variable to the principal components. Analysis of Figure B.415 indicates that CEA, GGT, AF, ALAT, ASAT, GFR, and CA19-9 are the variables that are best represented and contribute the most to PC1 and PC2. Finally, the Figure B.417 shows that the disease control group has a smaller prediction region compared to the progressive disease group. However, as this is only a projection onto PC1 and PC2, definitive conclusions cannot be drawn. Neither does plotting the scores in the first four principal dimensions as visualized in Figure B.418 provide any clear differences between the two groups. Lastly, the 3D-plots in Figure B.419 and Figure B.420 offer a visualization of the first three dimensions.

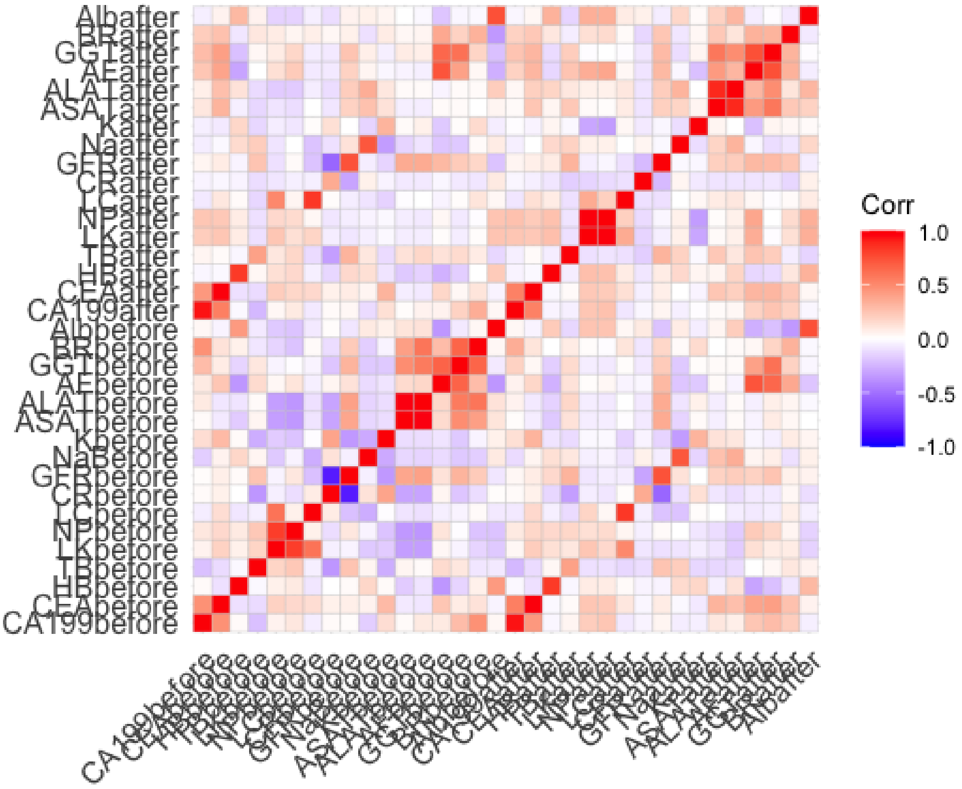


Figure B.408: Correlation Matrix of all the blood and tumor markers, n=59 (Disease control (n=44), Progressive disease (n=15)).

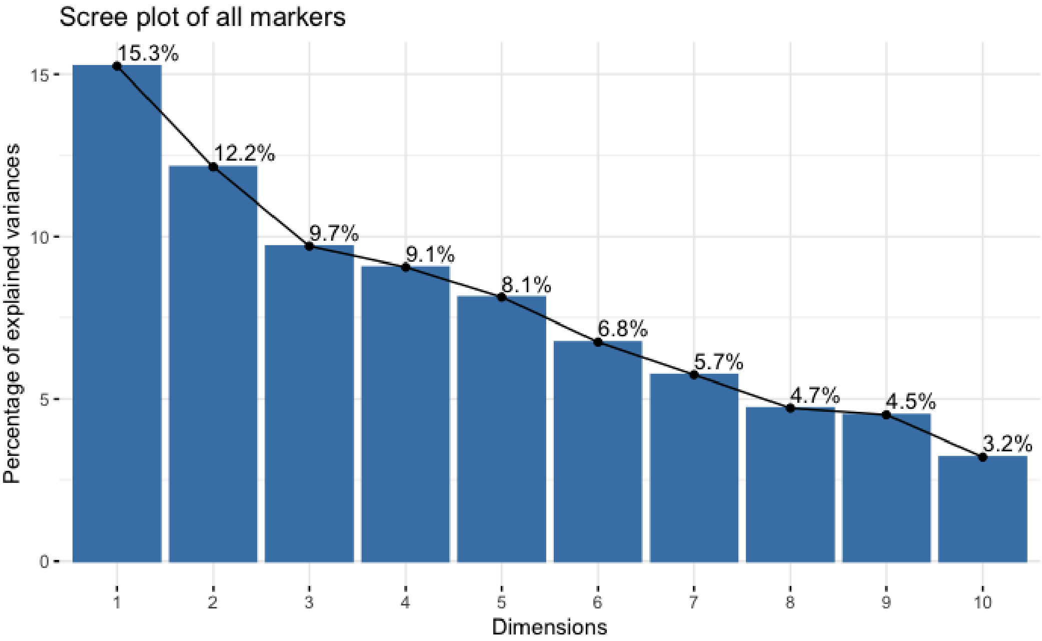


Figure B.409: Scree plot of all the blood and the tumor markers CEA and CA19-9, n=59 (Disease control (n=44), Progressive disease (n=15)).

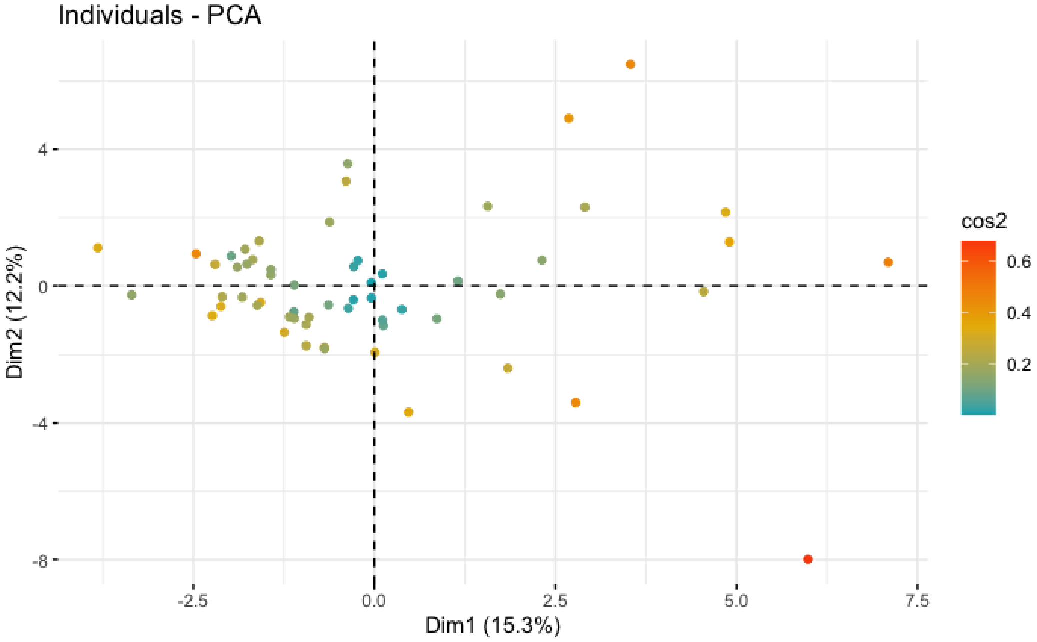


Figure B.410: Scatter plot of all the blood and the tumor markers CEA and CA19-9, n=59 (Disease control (n=44), Progressive disease (n=15)).

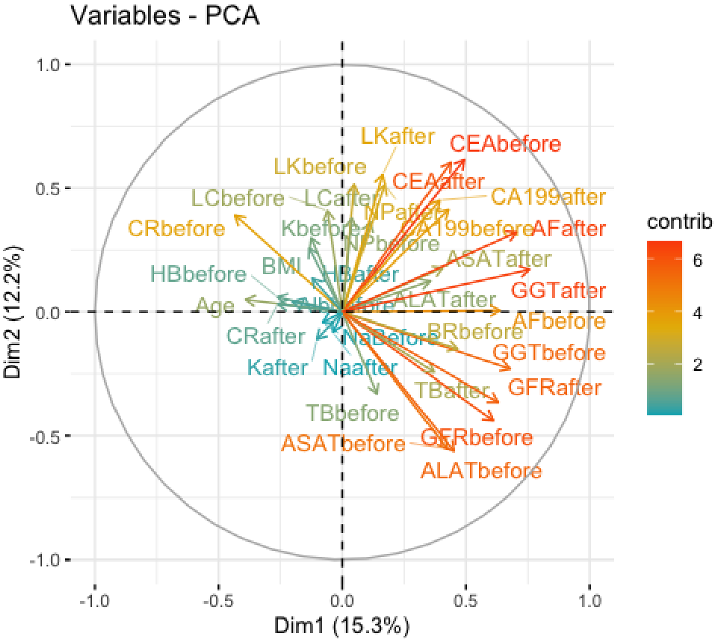


Figure B.411: Loading plot of all the blood and the tumor markers CEA and CA19-9, n=59 (Disease control (n=44), Progressive disease (n=15)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

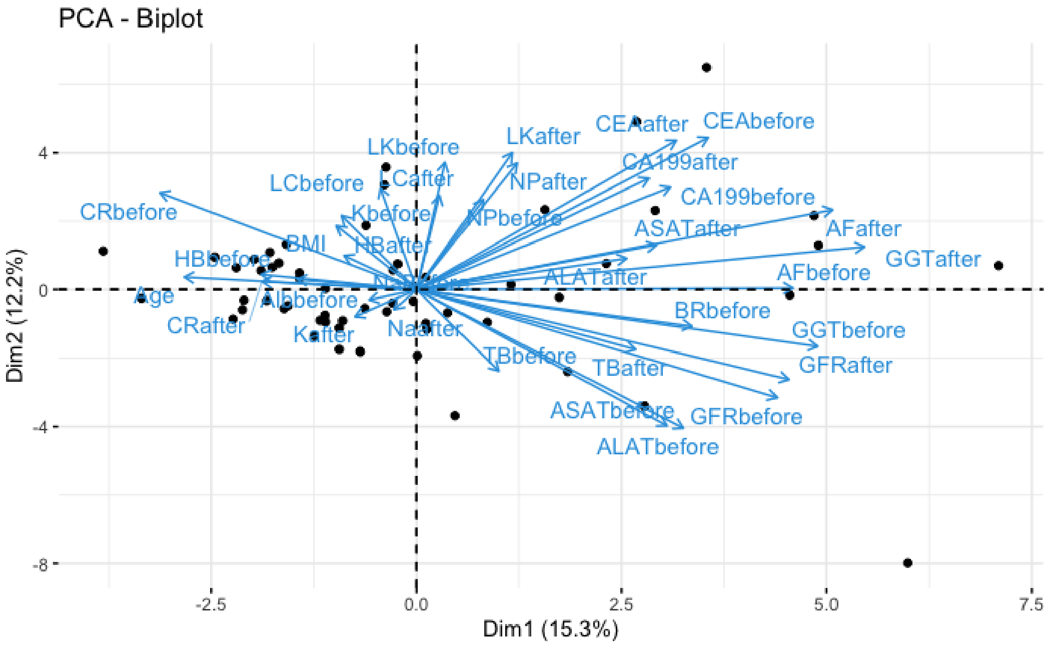
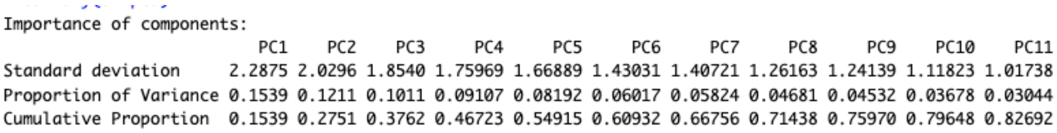


Figure B.412: Biplot plot of all the blood and the tumor markers CEA and CA19-9, n=59 (Disease control (n=44), Progressive disease (n=15)).



```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.2875 2.0296 1.8540 1.75969 1.66889 1.43031 1.40721 1.26163 1.24139 1.11823 1.01738
Proportion of Variance 0.1539 0.1211 0.1011 0.09107 0.08192 0.06017 0.05824 0.04681 0.04532 0.03678 0.03044
Cumulative Proportion  0.1539 0.2751 0.3762 0.46723 0.54915 0.60932 0.66756 0.71438 0.75970 0.79648 0.82692
```

Figure B.413: Table of PCA summary values all the blood and the tumor variables of the first 11 principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

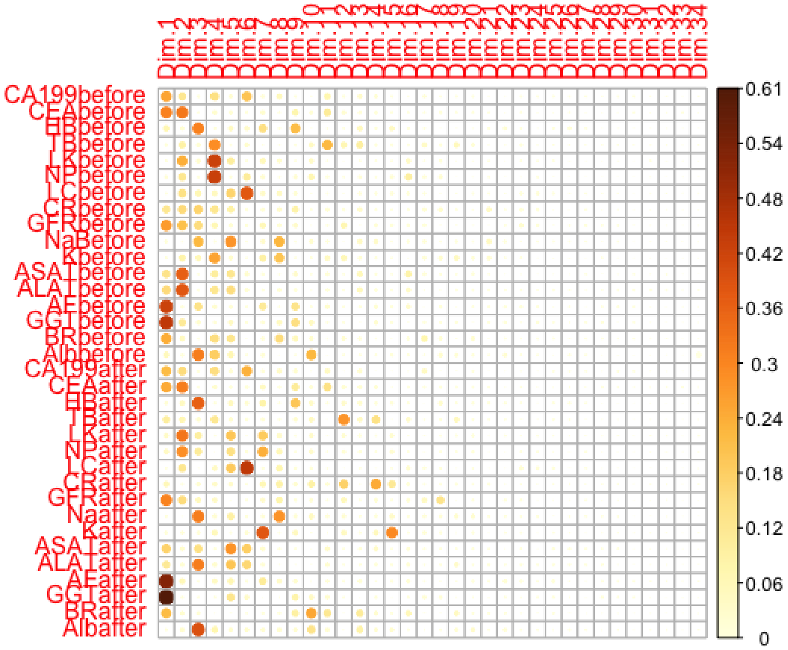


Figure B.414: Correlation plot of the original blood and tumor marker variables with the principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

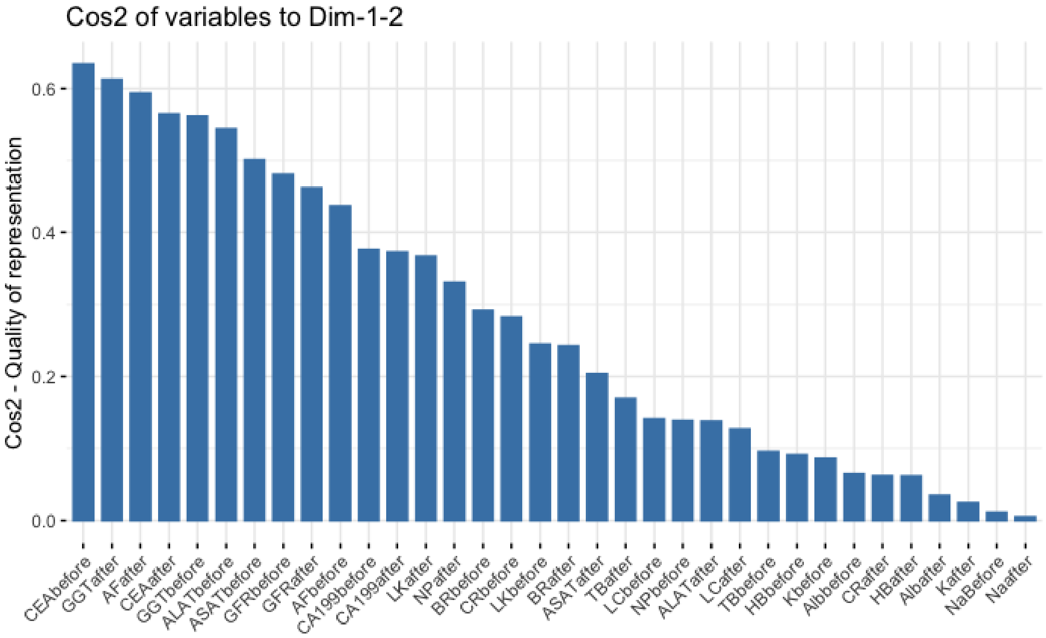


Figure B.415: Cos2 bar chart of the original blood and tumor marker variables, n=59 (Disease control (n=44), Progressive disease (n=15)).

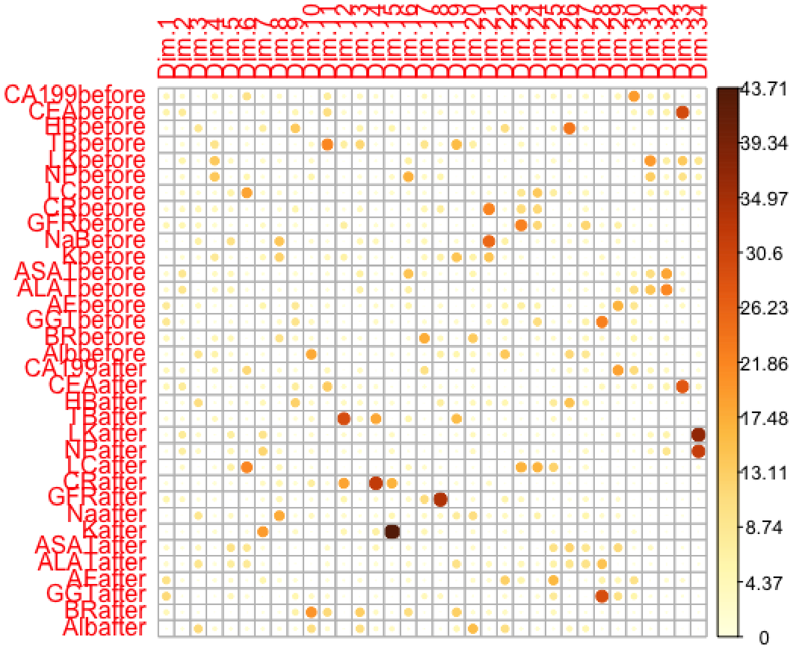


Figure B.416: Contribution of the original blood and tumor marker variables to the principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

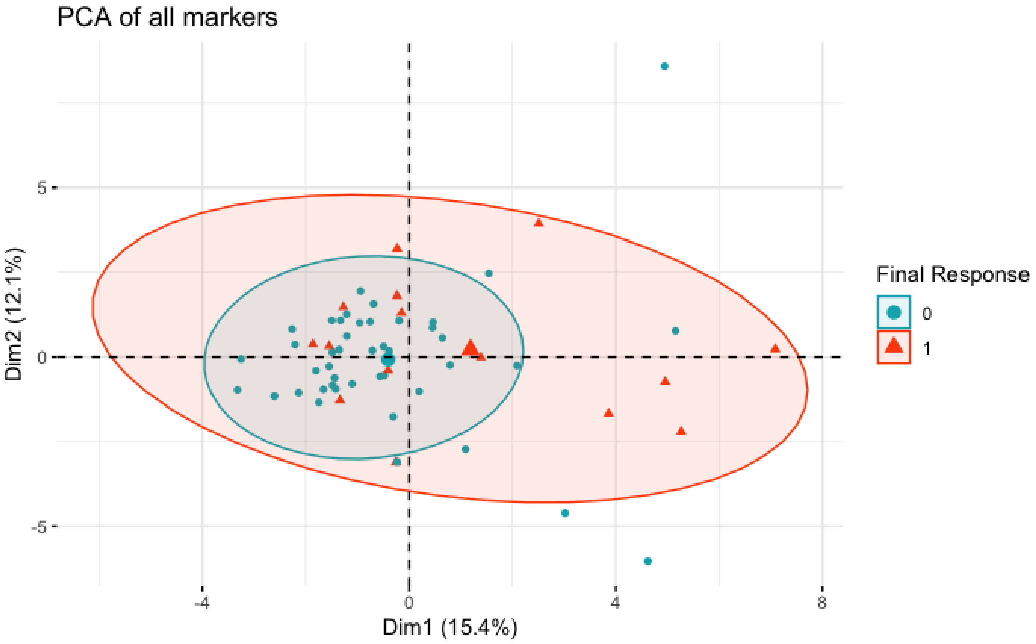


Figure B.417: PCA plot of all the blood markers and the tumor markers CEA and CA19-9 with 95% prediction ellipses. The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease), n=59 (Disease control (n=44), Progressive disease (n=15)).

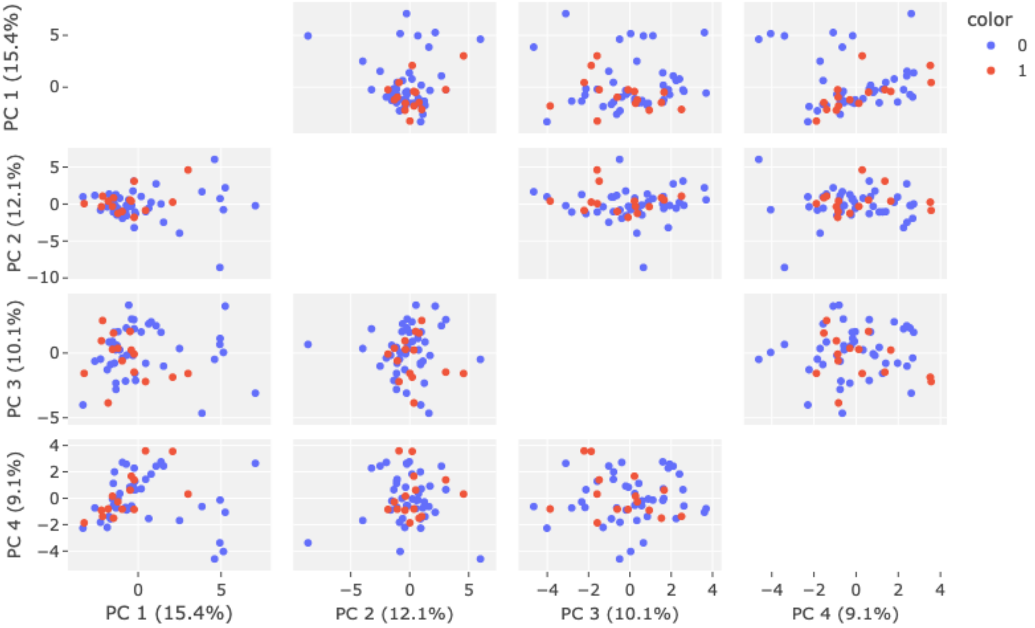


Figure B.418: Scatter matrix plot of the first four principal components. The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing all the measured variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease), n=59 (Disease control (n=44), Progressive disease (n=15)).

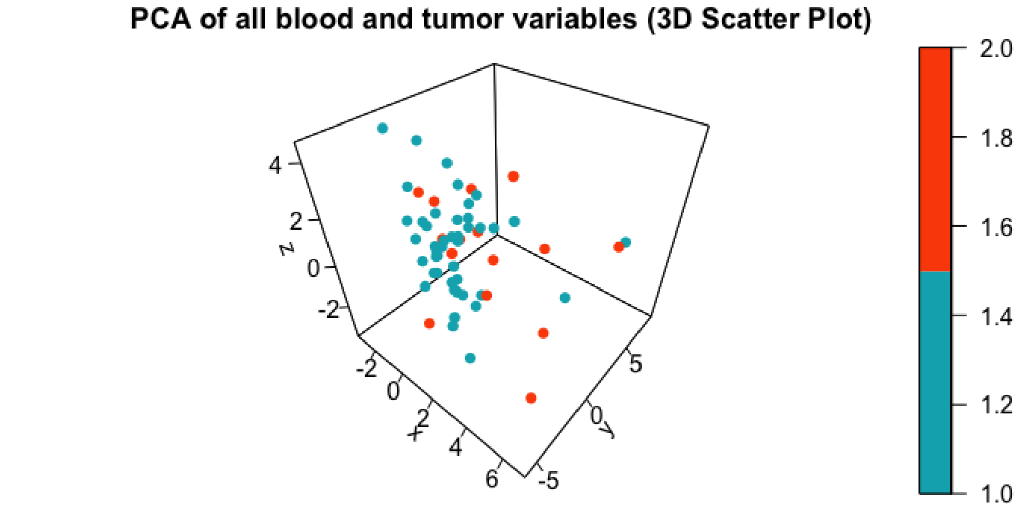


Figure B.419: 3D plot of the all the variables projected onto the first three principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

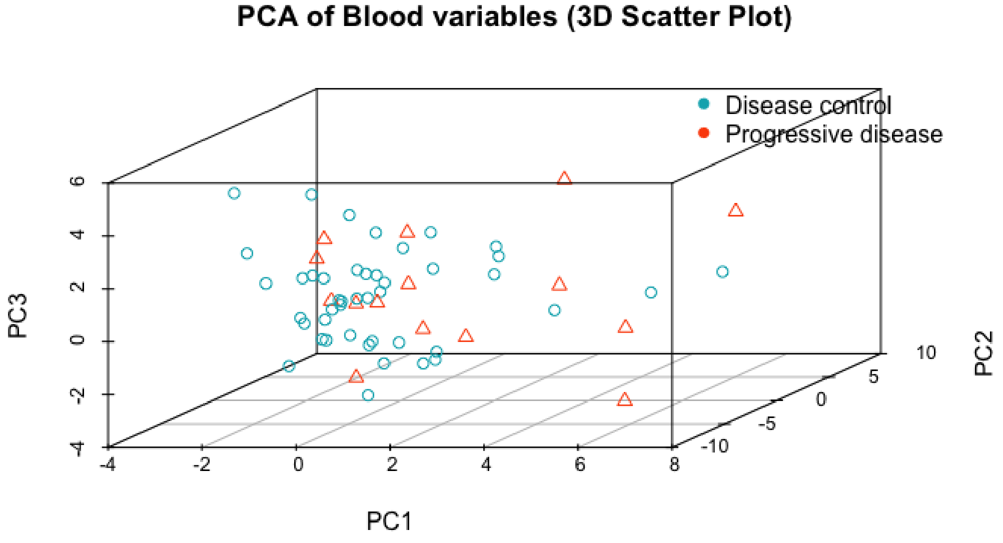


Figure B.420: 3D plot of all the variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=59 (Disease control (n=44), Progressive disease (n=15)).

B.4.9 All blood and tumor markers, Age, BMI and differences

In order to fully assess the efficacy of PCA in reducing dimensionality and transforming original variables, the entire dataset, including the two most significant patient characteristics, Age and BMI, were included in this PCA. Furthermore, the differences between variable measured before and after the first chemotherapy treatment were added to determine which directions of maximum variation would be identified in this case. The dataset comprises of 59 observations of 53 explanatory variables, and the correlation matrix can be found in Figure B.421. As with the previous subsection, the plots in Figure B.422, Figure B.424. Figure B.425 and Figure B.430 should be used primarily for intuition since they involve projections from a 53-dimensional space to a 2-dimensional space. Analysis of Figure B.428 reveals that the highest degree of variation in this case is found in the variables: ALAT, ASAT, GGT, LK, NP, BR, and AF. This finding is distinct from the previous subsection, where differences between measures were not taken into account, suggesting that consideration of differences in values before and after the first chemotherapy treatment may be preferable to simply examining individual values alone. Nevertheless, the ellipses in Figure B.430 fail to provide a clear differentiation between the two final response groups. Neither does the projection onto the first four principal components in 2D, visualized in Figure B.431 provide any more clear separations. Nor is the projection of the scores onto the first three principal components, as presented in Figure B.432 and Figure B.433, of more information.

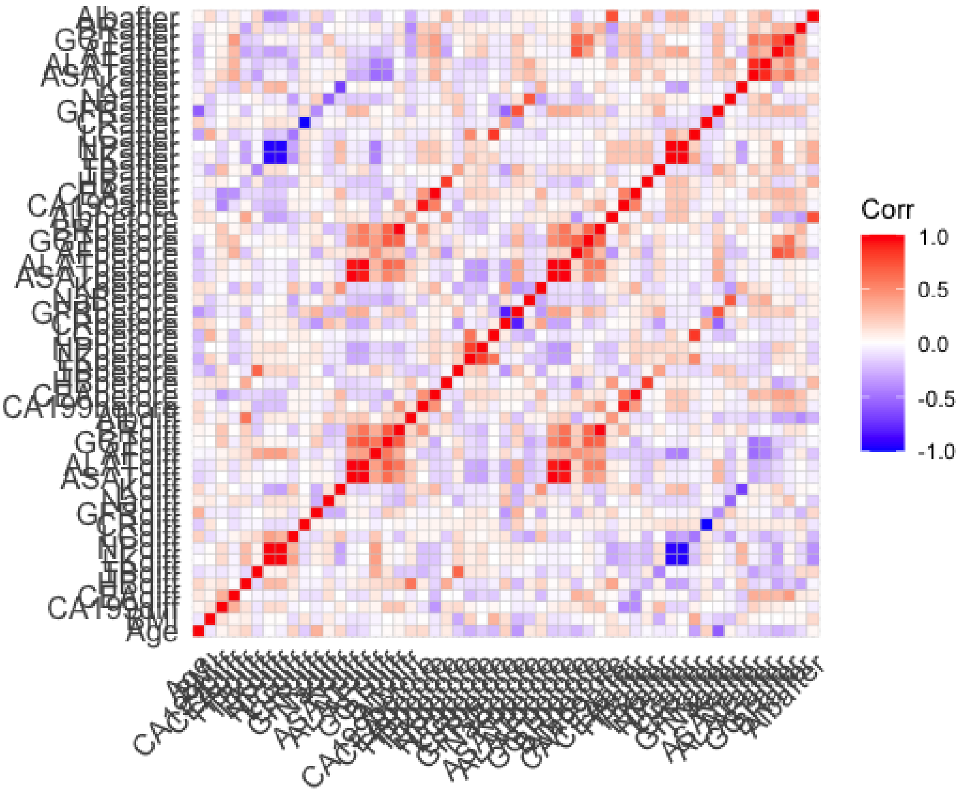


Figure B.421: Correlation Matrix of all the blood and tumor markers, Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

B.4.10 Age, BMI and Differences

Finally, the effect of chemotherapy on blood variables and tumor variables is taken into account by considering only the differences between the before and after treatment values. Additionally, age and BMI are included in the analysis to make it more complete. The dataset in consideration consists of 59 observations of 19 explanatory variables. The correlation matrix in Figure B.434 indicate that most differences are uncorrelated or weakly positively/negatively correlated, with the exception of ASAT/ALAT, AF/GGT, GGT/BR, and LK/NP differences which exhibit strong positive correlation. However, this was seen in earlier analyses as well. PCA is performed, and the screeplot as well as the table in Table B.104 demonstrate that the first few principal components account for a relatively small portion of the total variance in the data, indicating that the data does not possess a clear dominant structure that can be captured by a few principal components. Due to the lack of linear relationships between almost all variables, each variable contributes independently to the overall variance of the data, making it difficult to identify a small number of principal components that capture the majority of the variance.

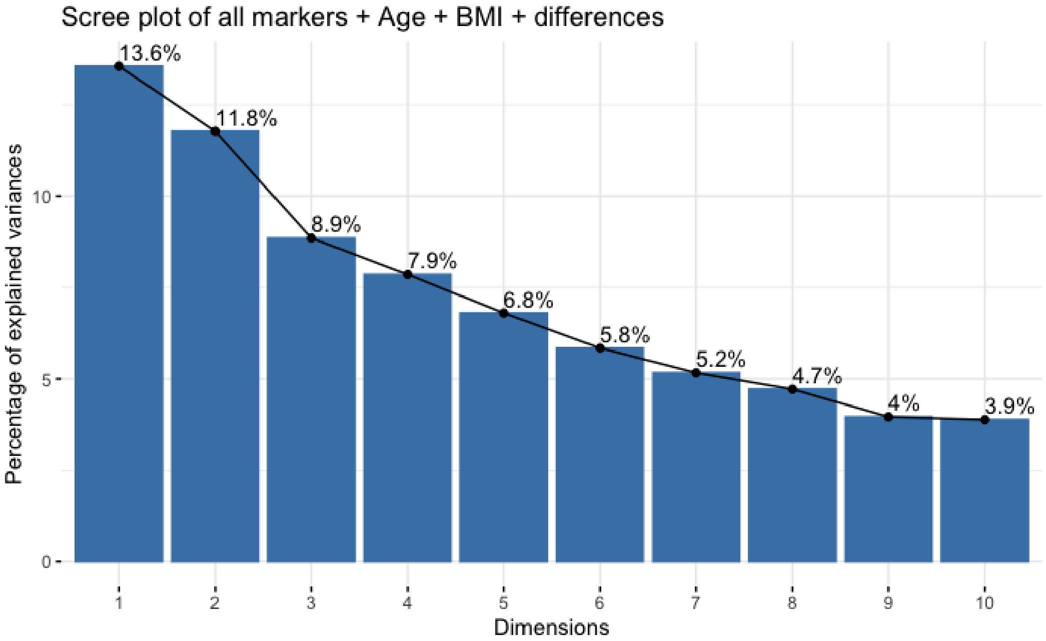


Figure B.422: Scree plot of all the blood and the tumor markers CEA and CA19-9, Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

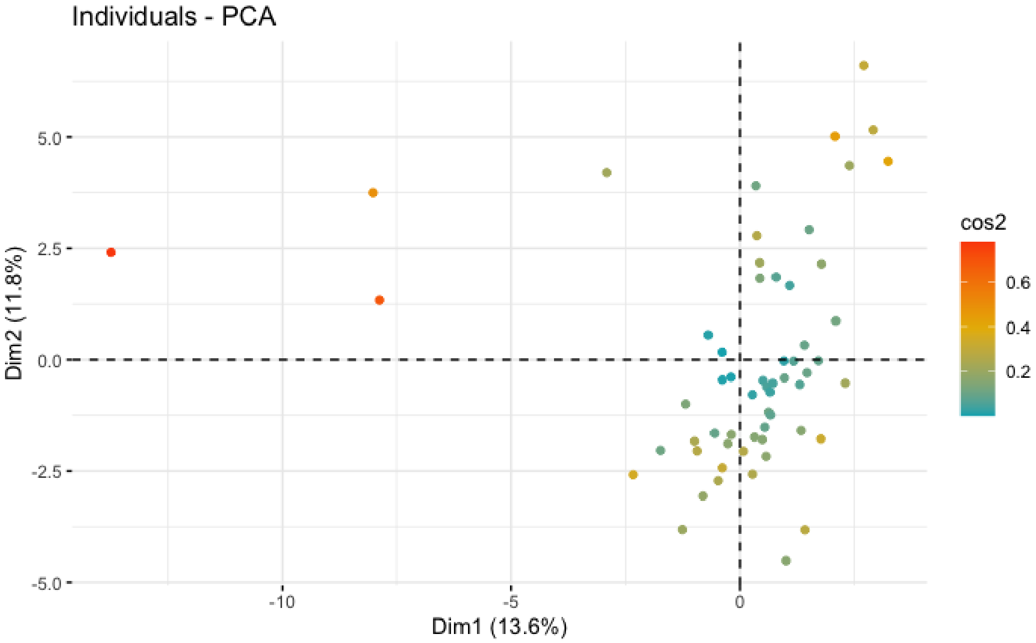


Figure B.423: Scatter plot of all the blood and the tumor markers CEA and CA19-9, Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

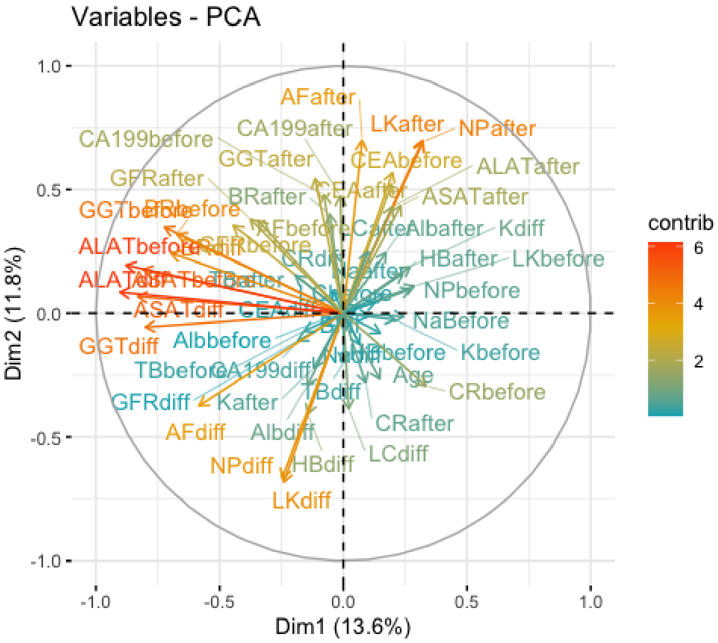


Figure B.424: Loading plot of all the blood and the tumor markers CEA and CA19-9, Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

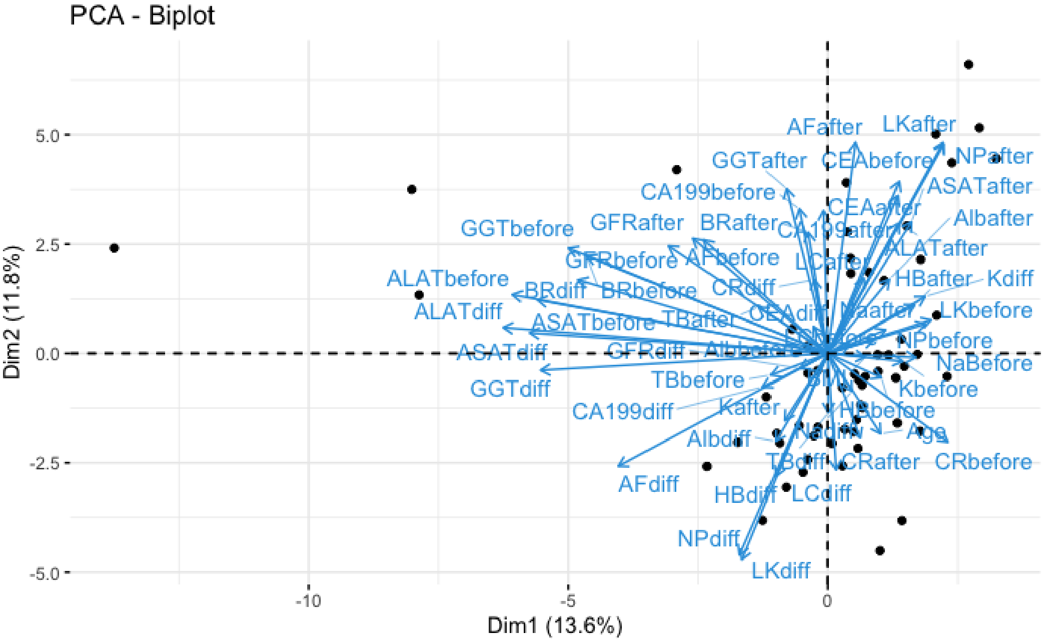
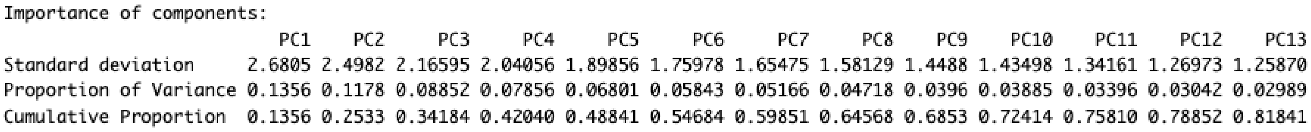


Figure B.425: Biplot plot of all the blood and the tumor markers CEA and CA19-9, Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).



```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8    PC9    PC10    PC11    PC12    PC13
Standard deviation     2.6805 2.4982 2.16595 2.04056 1.89856 1.75978 1.65475 1.58129 1.4488 1.43498 1.34161 1.26973 1.25870
Proportion of Variance 0.1356 0.1178 0.08852 0.07856 0.06801 0.05843 0.05166 0.04718 0.0396 0.03885 0.03396 0.03042 0.02989
Cumulative Proportion  0.1356 0.2533 0.34184 0.42040 0.48841 0.54684 0.59851 0.64568 0.6853 0.72414 0.75810 0.78852 0.81841
```

Figure B.426: PCA summary values of all the blood and the tumor variables, Age, BMI and differences before and after the first chemotherapy treatment of the first 13 principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

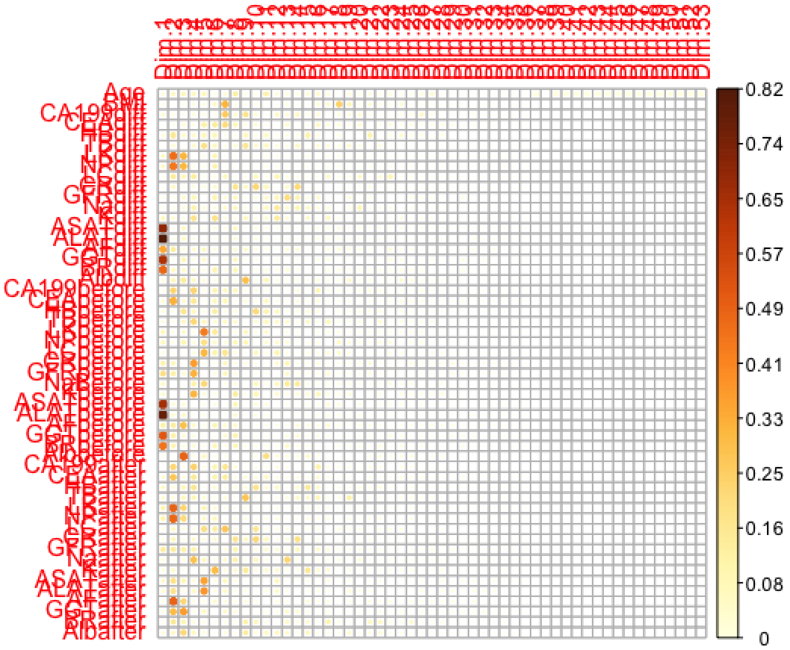


Figure B.427: Correlation plot of the original blood and tumor marker variables, Age, BMI and differences before and after the first chemotherapy treatment with the principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

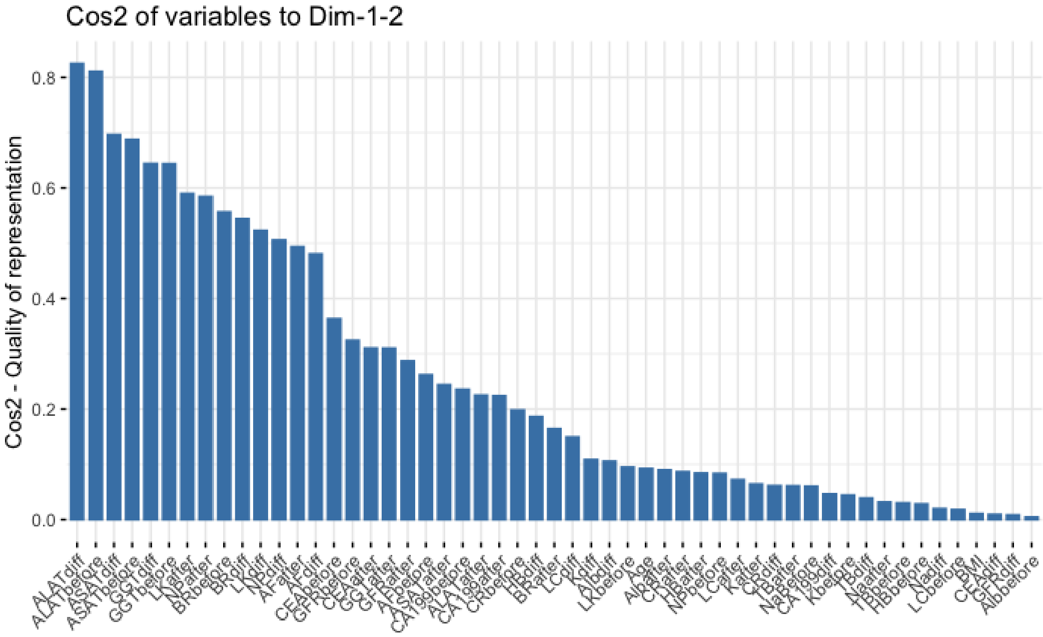


Figure B.428: Cos2 bar chart of the original blood and tumor marker variables, Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

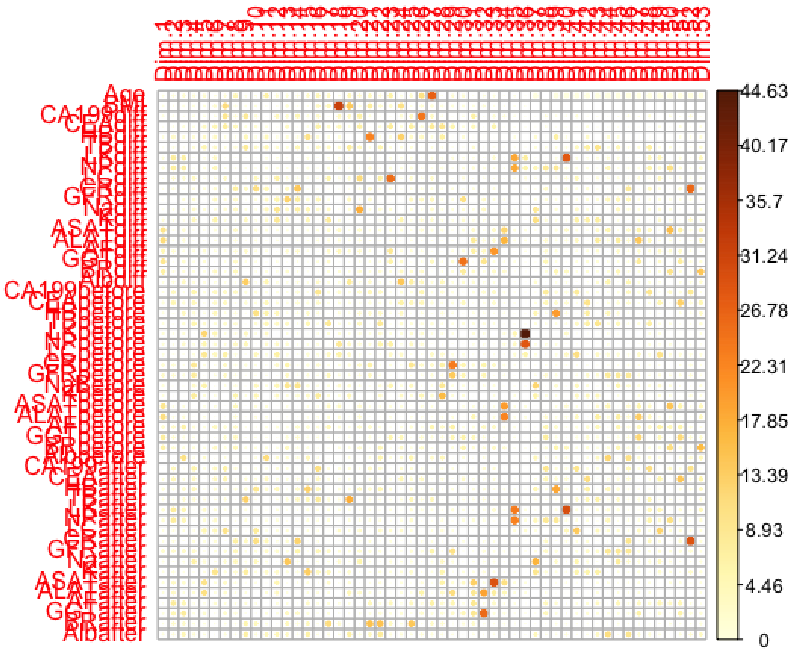


Figure B.429: Contribution of the original blood and tumor marker, Age, BMI and differences before and after the first chemotherapy treatment variables to the principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

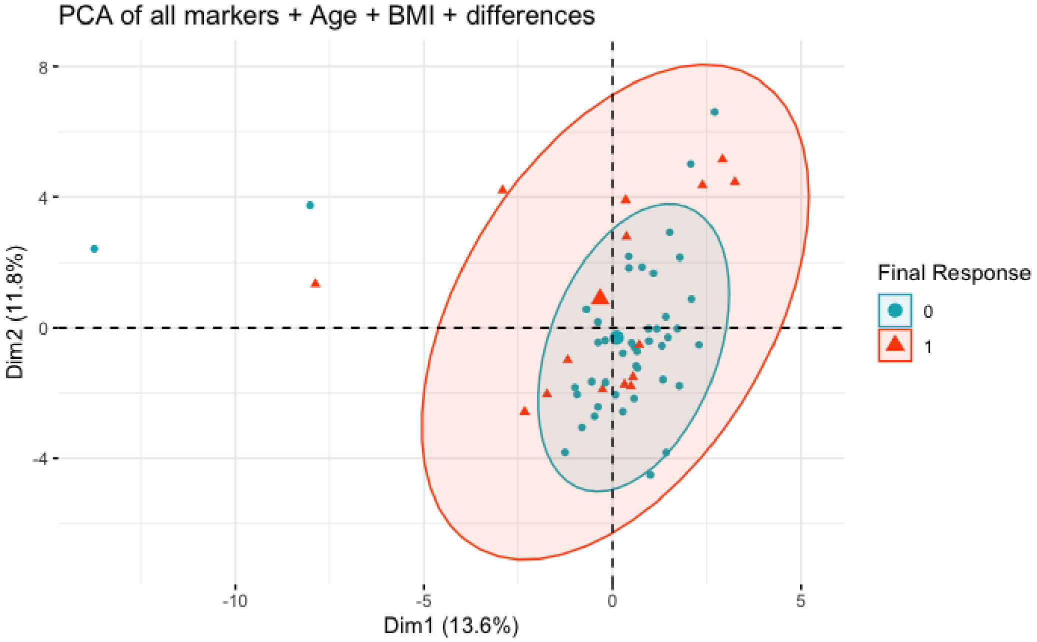


Figure B.430: PCA plot of all the blood markers and the tumor markers CEA and CA19-9, Age, BMI and differences before and after the first chemotherapy treatment with 95% prediction ellipses, n=59 (Disease control (n=44), Progressive disease (n=15)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

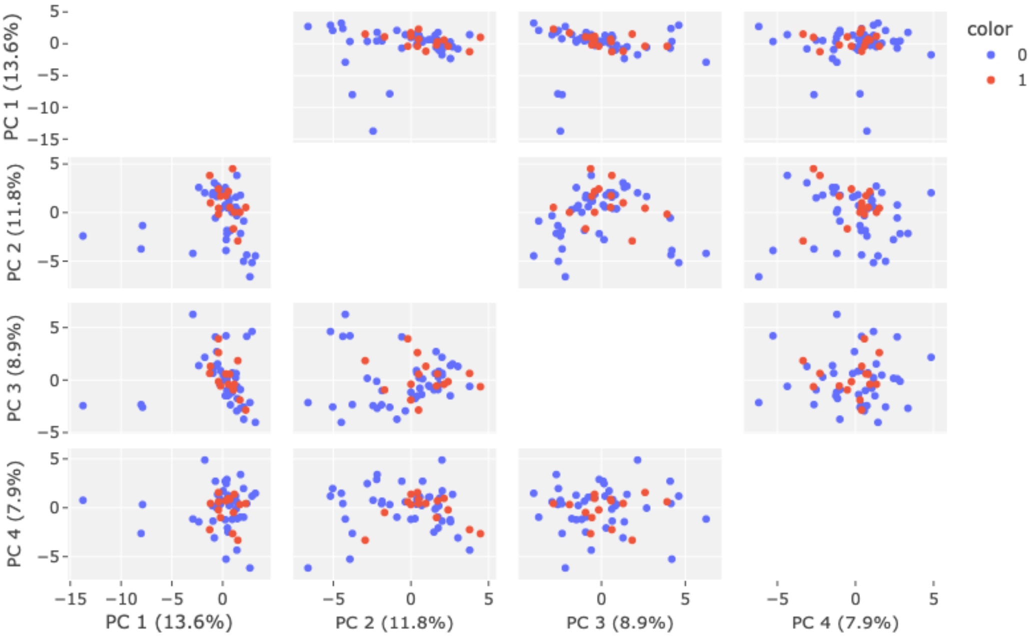


Figure B.431: Scatter matrix plot of the first four principal components. The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing all the measured variables and tumor marker, Age, BMI and differences before/after chemotherapy of the variables with The final dichotomized response is coloured, n=59 (Disease control (n=44), Progressive disease (n=15)), Blue = 0 (Disease control) and Red = 1 (Progressive disease).

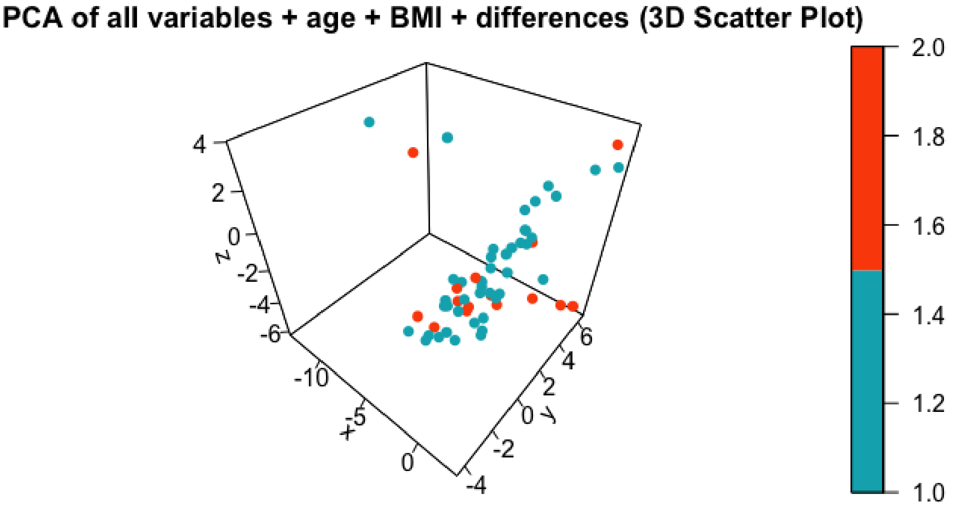


Figure B.432: 3D plot of all the variables, age, BMI and differences projected onto the first three principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

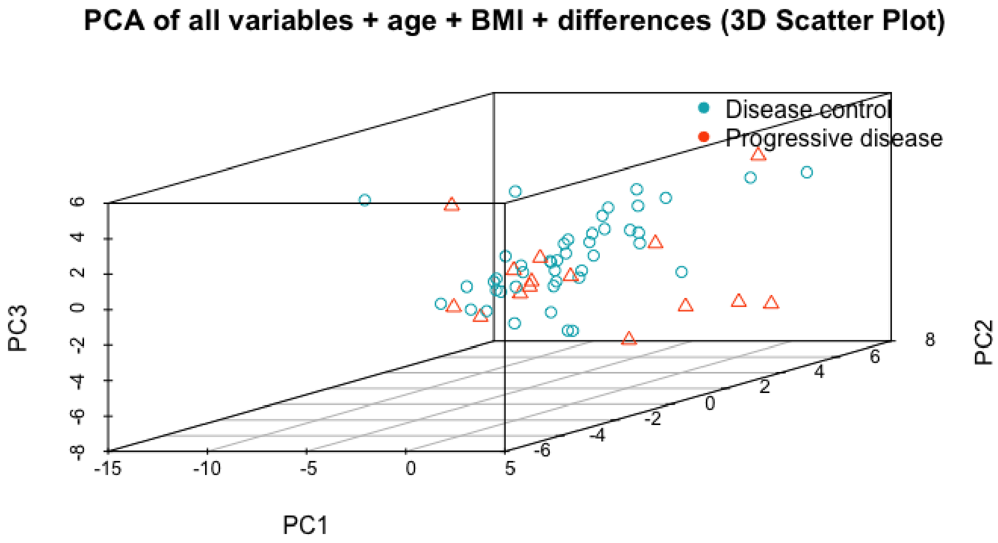


Figure B.433: 3D plot of all the variables, age, BMI and differences with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=59 (Disease control (n=44), Progressive disease (n=15)).

Nonetheless, the plots in Figure B.437, Figure B.438, Figure B.440, Figure B.441, and Figure B.442 may still provide valuable insights. The loading vectors in Figure B.437 and Figure B.438 show that the first two dimensions are primarily determined by LKdiff, NPdiff, GGTdiff, and ALATdiff, which are the best represented variables. LKdiff/NPdiff are strongly correlated with PC2, and GGTdiff/ALATdiff are strongly correlated with PC1. Figure B.440 make evident that the first dimension captures liver measurements (ASAT, ALAT, AF, GGT, BR), the second dimension represents white blood cells (LK, NP, LC), the third dimension is mostly Alb, and the fourth is GFR. Finally, the PCA ellipse plot in Figure B.443 shows that the ellipse of the disease control group is more elongated than that of the progressive disease group, indicating that the observations of the disease group are more diverse than those of the progressive disease group. However, the disease control group ellipse lies within the progressive disease group ellipse, and the plot is only two-dimensional, with PC1 and PC2 accounting for a relatively small portion of the total variation in the data. Hence, any clear conclusions cannot be drawn. Projecting the scores onto the first four principal components does not portray any clear distinctions between the two final response groups. Ultimately, the 3D plots in Figure B.445 and Figure B.446 offer a view of the projection of the scores onto the first three principal components, with no clear patterns either.

Age, BMI and Differences variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	1.906	1.622	1.366	1.270	1.217
Proportion of Variance Explained	0.191	0.139	0.098	0.085	0.078
Cumulative Proportion	0.191	0.330	0.428	0.513	0.591

Table B.104: PCA summary values of the Age, BMI and differences before and after the first chemotherapy treatment of the first 5 PCs. It shows the standard deviation, proportion of variance explained and cumulative proportion of the first five principal components. From the total summary it is seen that using 9 PCs 82% of the total variation is explained and with 10 PCs 86% and only with 12 PCs we have 93% of the total variation, n=59 (Disease control (n=44), Progressive disease (n=15)).

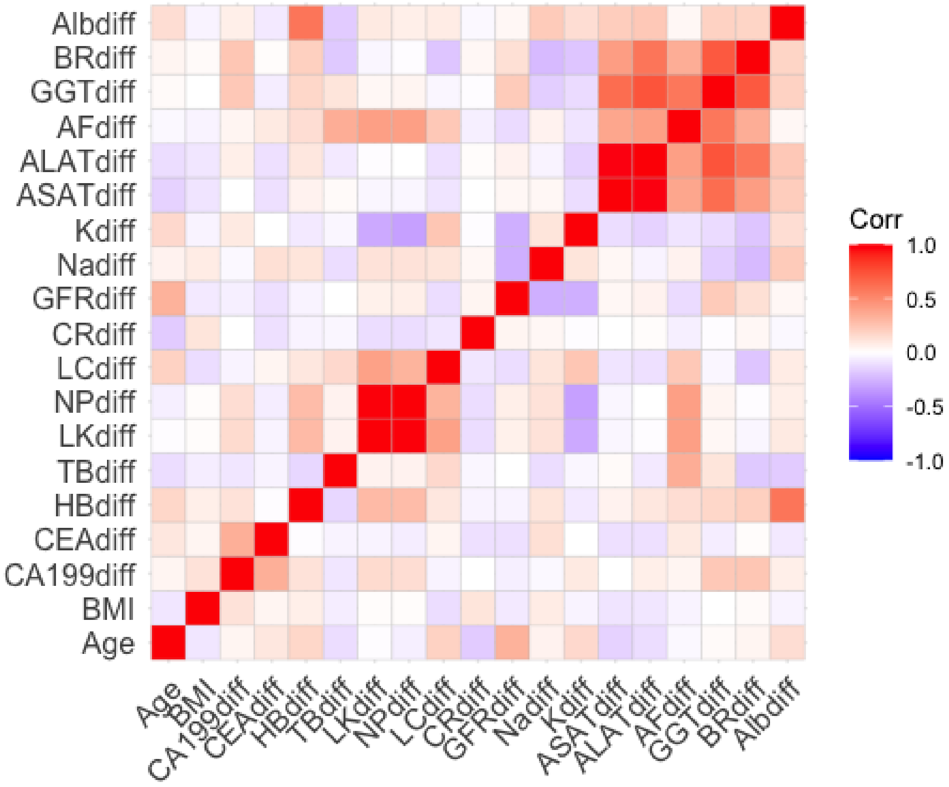


Figure B.434: Correlation Matrix of Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

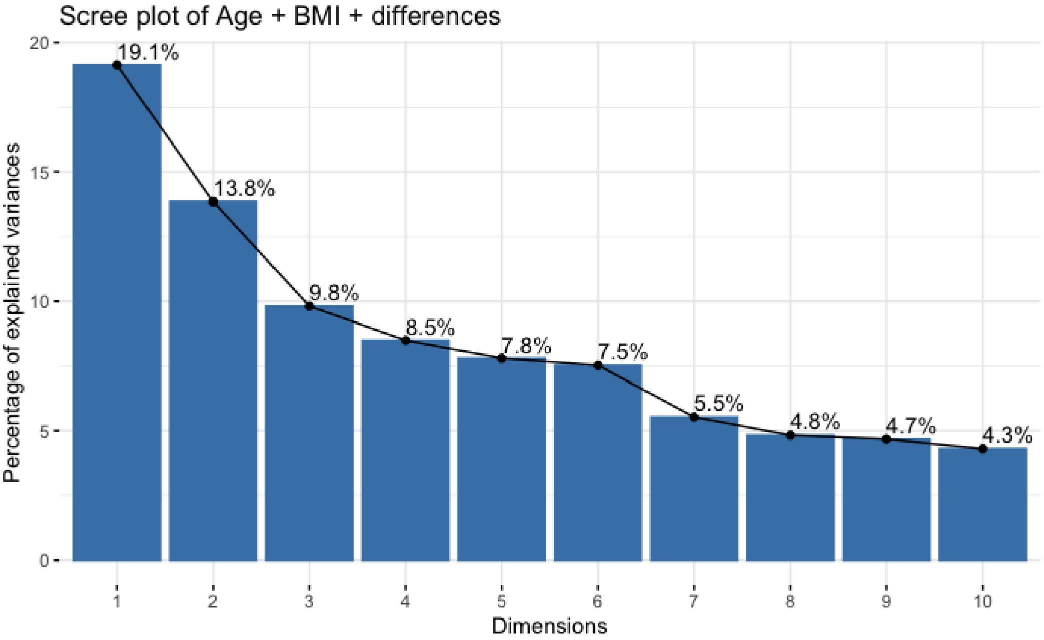


Figure B.435: Scree plot of Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

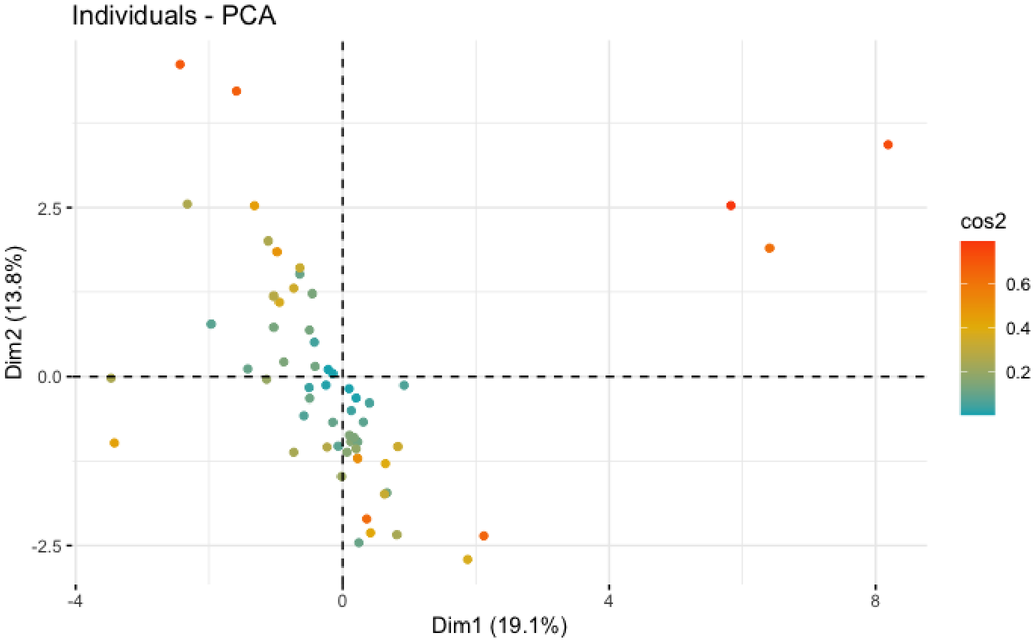


Figure B.436: Scatter plot of Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

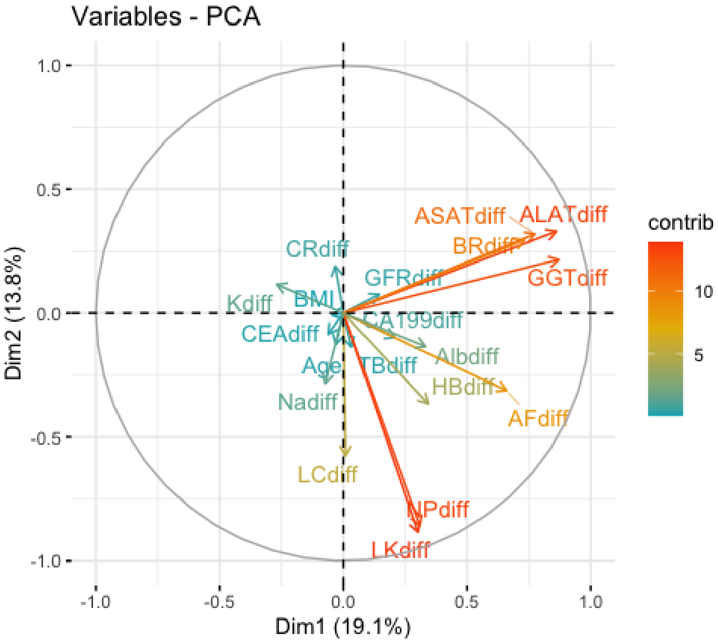


Figure B.437: Loading plot of Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)). The circle reflects how well the variables are described. The longer the loading vector (so closer to the circle), the more the information is captured of that variable. The length of the arrow is proportional to how well the variable is explained.

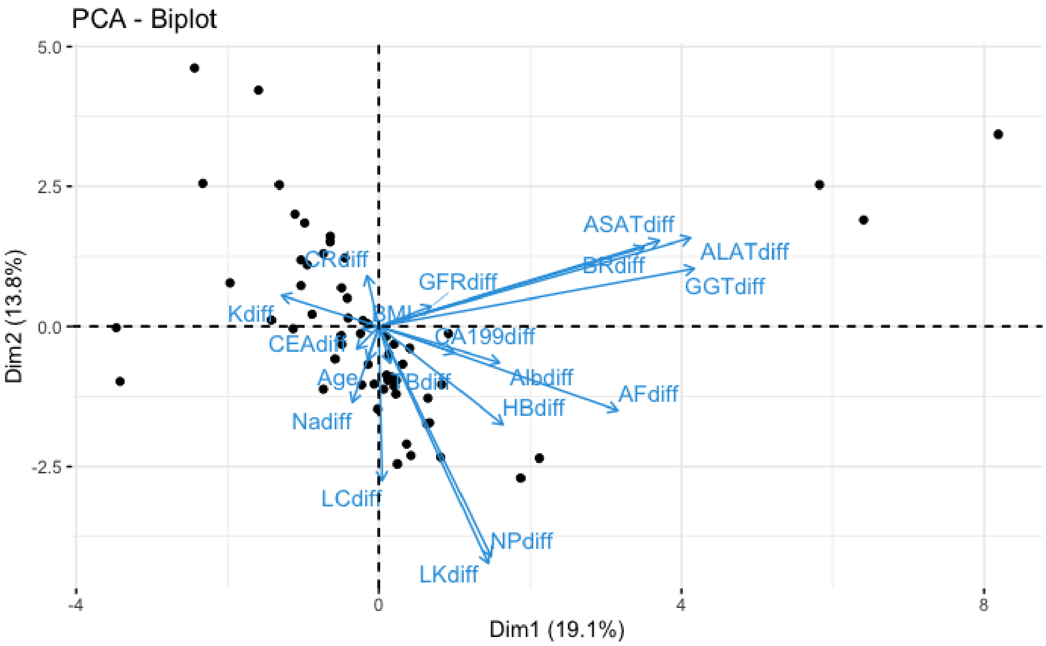
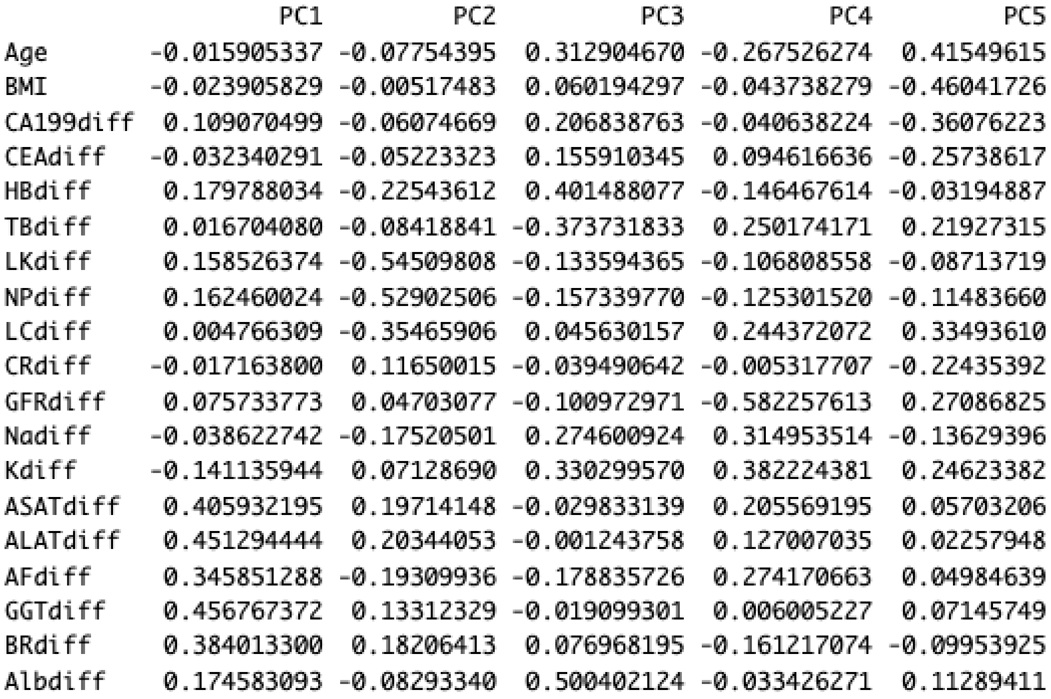


Figure B.438: Biplot plot of Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).



```
                   PC1         PC2          PC3          PC4         PC5
Age       -0.015905337 -0.07754395  0.312904670 -0.267526274  0.41549615
BMI       -0.023905829 -0.00517483  0.060194297 -0.043738279 -0.46041726
CA199diff  0.109070499 -0.06074669  0.206838763 -0.040638224 -0.36076223
CEAdiff   -0.032340291 -0.05223323  0.155910345  0.094616636 -0.25738617
HBdiff     0.179788034 -0.22543612  0.401488077 -0.146467614 -0.03194887
TBdiff     0.016704080 -0.08418841 -0.373731833  0.250174171  0.21927315
LKdiff     0.158526374 -0.54509808 -0.133594365 -0.106808558 -0.08713719
NPdiff     0.162460024 -0.52902506 -0.157339770 -0.125301520 -0.11483660
LCdiff     0.004766309 -0.35465906  0.045630157  0.244372072  0.33493610
CRdiff    -0.017163800  0.11650015 -0.039490642 -0.005317707 -0.22435392
GFRdiff    0.075733773  0.04703077 -0.100972971 -0.582257613  0.27086825
Nadiff    -0.038622742 -0.17520501  0.274600924  0.314953514 -0.13629396
Kdiff     -0.141135944  0.07128690  0.330299570  0.382224381  0.24623382
ASATdiff   0.405932195  0.19714148 -0.029833139  0.205569195  0.05703206
ALATdiff   0.451294444  0.20344053 -0.001243758  0.127007035  0.02257948
AFdiff     0.345851288 -0.19309936 -0.178835726  0.274170663  0.04984639
GGTdiff    0.456767372  0.13312329 -0.019099301  0.006005227  0.07145749
BRdiff     0.384013300  0.18206413  0.076968195 -0.161217074 -0.09953925
Albdiff    0.174583093 -0.08293340  0.500402124 -0.033426271  0.11289411
```

Figure B.439: Loadings of Age, BMI and differences before and after the first chemotherapy treatment of the first 5 principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

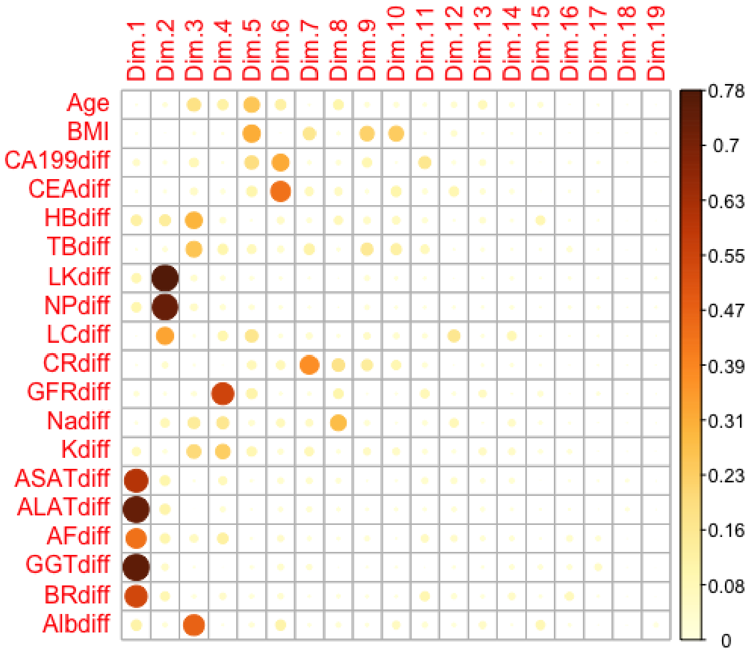


Figure B.440: Correlation plot of the original Age, BMI and differences before and after the first chemotherapy treatment with the principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

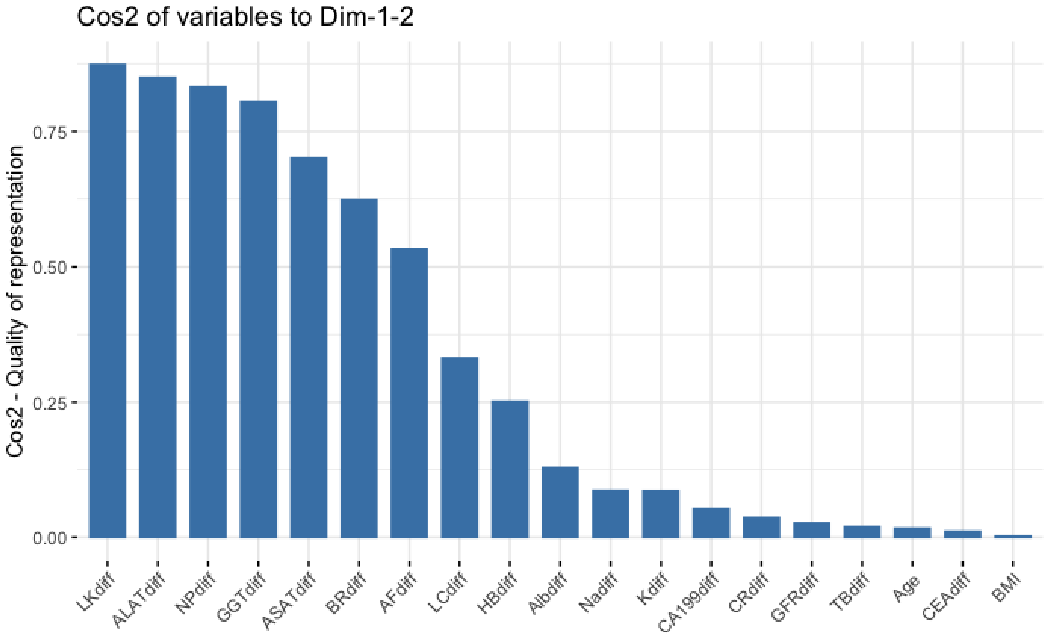


Figure B.441: Cos2 bar chart of the original Age, BMI and differences before and after the first chemotherapy treatment, n=59 (Disease control (n=44), Progressive disease (n=15)).

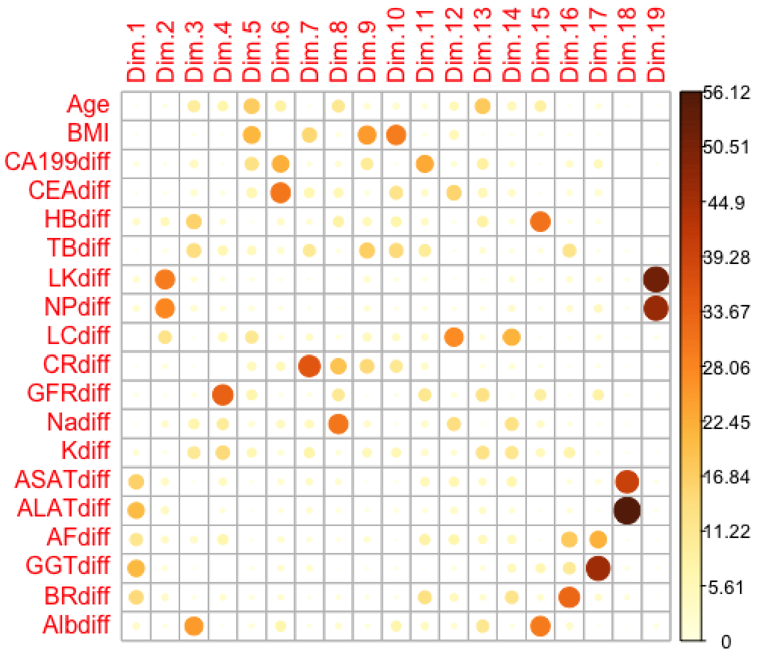


Figure B.442: Contribution of the original Age, BMI and differences before and after the first chemotherapy treatment variables to the principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

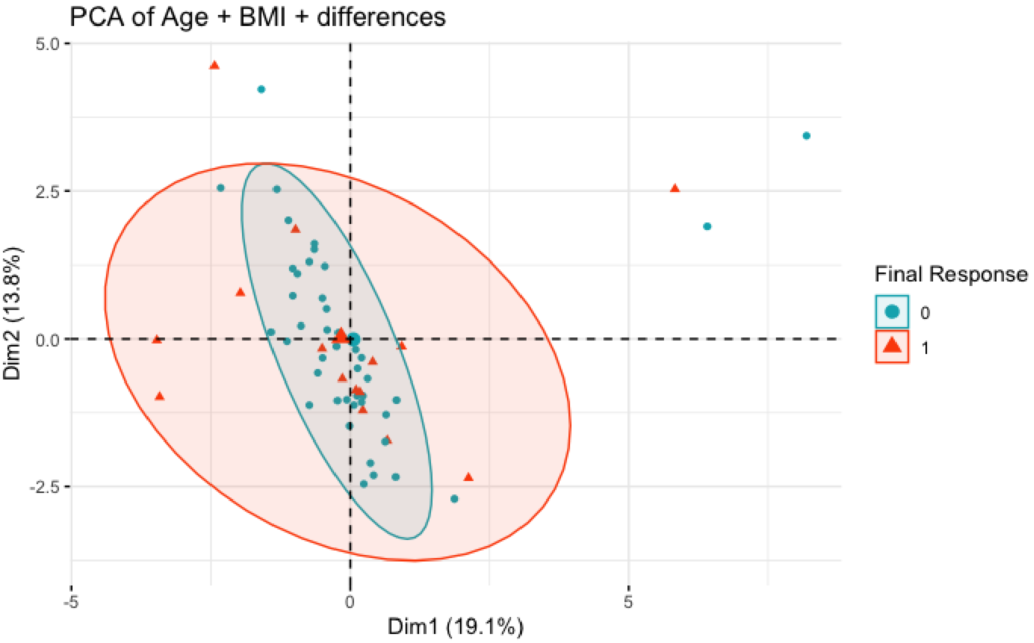


Figure B.443: PCA plot of Age, BMI and differences before and after the first chemotherapy treatment with 95% prediction ellipses, n=59 (Disease control (n=44), Progressive disease (n=15)). The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease).

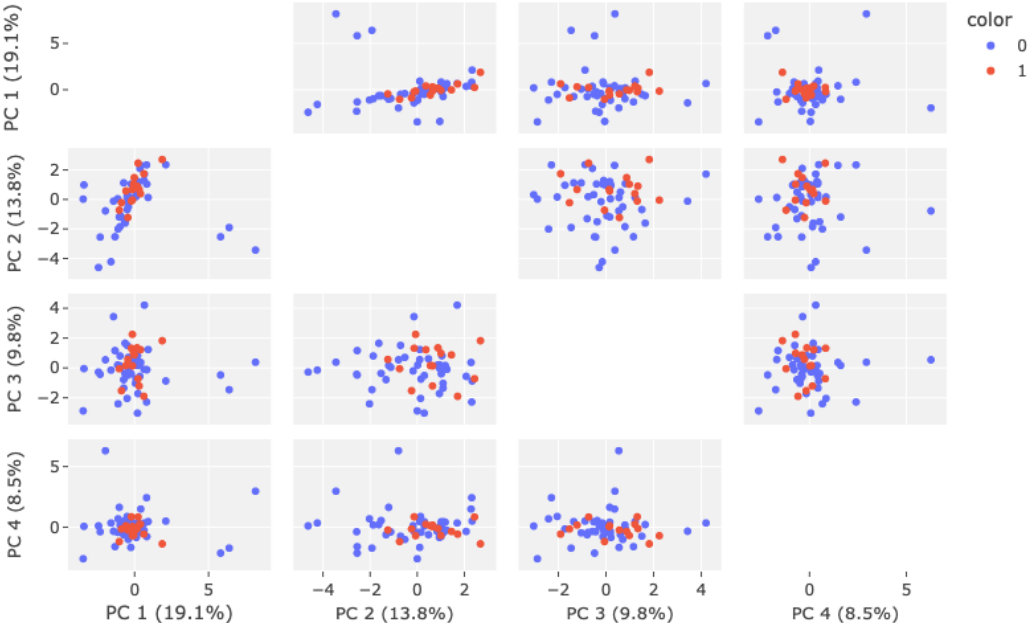


Figure B.444: Scatter matrix plot of the first four principal components. The subplots between PC1, PC2, PC3 and PC4 are provided of the dataset containing Age, BMI and the differences before and after the first chemotherapy cycle of the variables with The final dichotomized response is coloured, Blue = 0 (Disease control) and Red = 1 (Progressive disease), n=59 (Disease control (n=44), Progressive disease (n=15)).

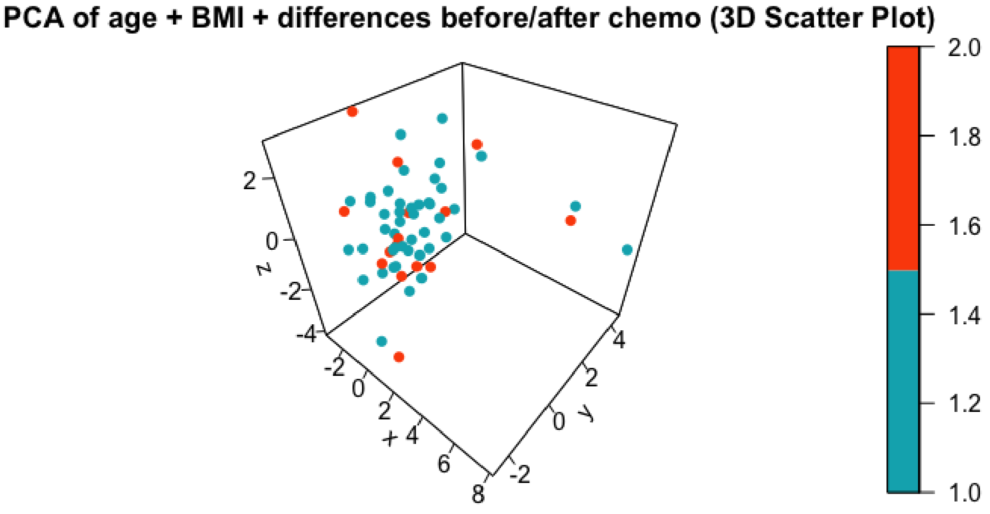


Figure B.445: 3D plot of Age, BMI and the differences before and after the first chemotherapy cycle of the variables projected onto the first three principal components, n=59 (Disease control (n=44), Progressive disease (n=15)).

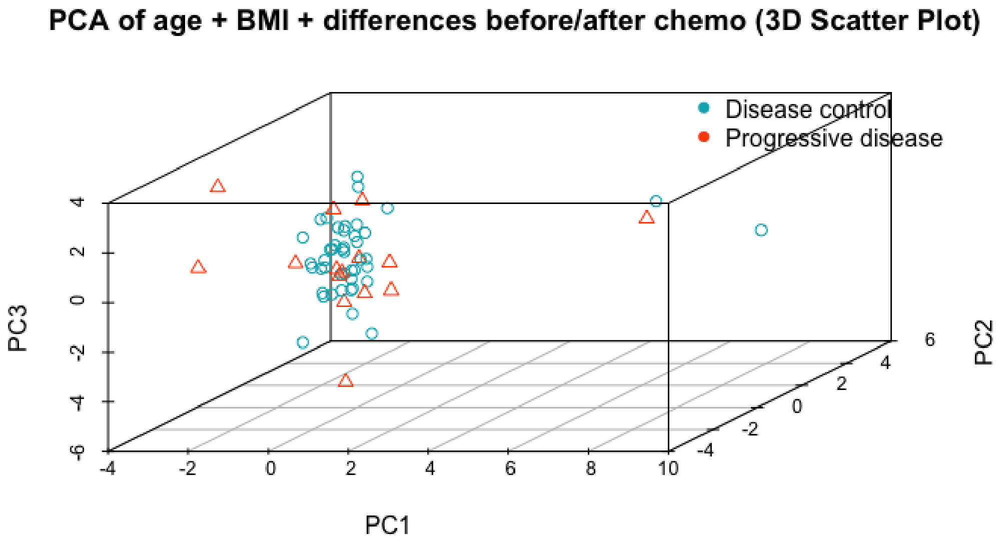


Figure B.446: 3D plot of Age, BMI and the differences before and after the first chemotherapy cycle of the variables with the final dichotomized response coloured and shaped, Blue circle = 0 (Disease control) and Red Triangle = 1 (Progressive disease), n=59 (Disease control (n=44), Progressive disease (n=15)).

B.4.11 Choice of PCA method in R

R has two built-in functions to perform Principal Component Analysis. There are generally two methods:

- Eigen/Spectral decomposition which *princomp()* uses
- Singular Value decomposition which *prcomp()* uses

According to R software, SVD has a slightly better numerical accuracy and therefore, the function prcomp() is preferred compared to princomp().

The prcomp() function returns an object of class 'prcomp', which contains the following components:

- sdev: vector of standard deviations = $\sqrt{eigenvalues}$, determines how much variance is explained by each component
- rotation: matrix of loadings of the principal components, can be used to interpret the direction and strength of he relationships between the variables and principal components
- center: the centered values used for the data
- scale: the scaling values used for the data
- x: the centered and/or scaled data matrix
- y: the scores of the principal components (projections of the original data onto the principal component)

B.5. PCA using covariance matrix

In this section, we will briefly discuss the findings of the same analysis done as in B.3 but using the covariance matrix instead of the correlation matrix. The values in the covariance matrix are centered by their respective means. For each data set, the scree plots were used to determine the number of principal components required to capture a significant amount of the total variation in the data ≥ 90%.

B.5.1 Blood

To begin, PCA without scaling is performed on the blood data set. The screeplot in Figure B.447 depicts that using solely the first two principal components is already sufficient to capture 99% of the total variation in the data. Notably, PCA displays a strong correlation with TBbefore and TBdiff, whereas PC2 exhibits a high correlation with TBafter. These findings suggest that thrombocytes (TB) represent the most significant contributor to variability in the blood data set. However, it should be noted that the prediction ellipses remain overlapping, indicating a need for further investigation.

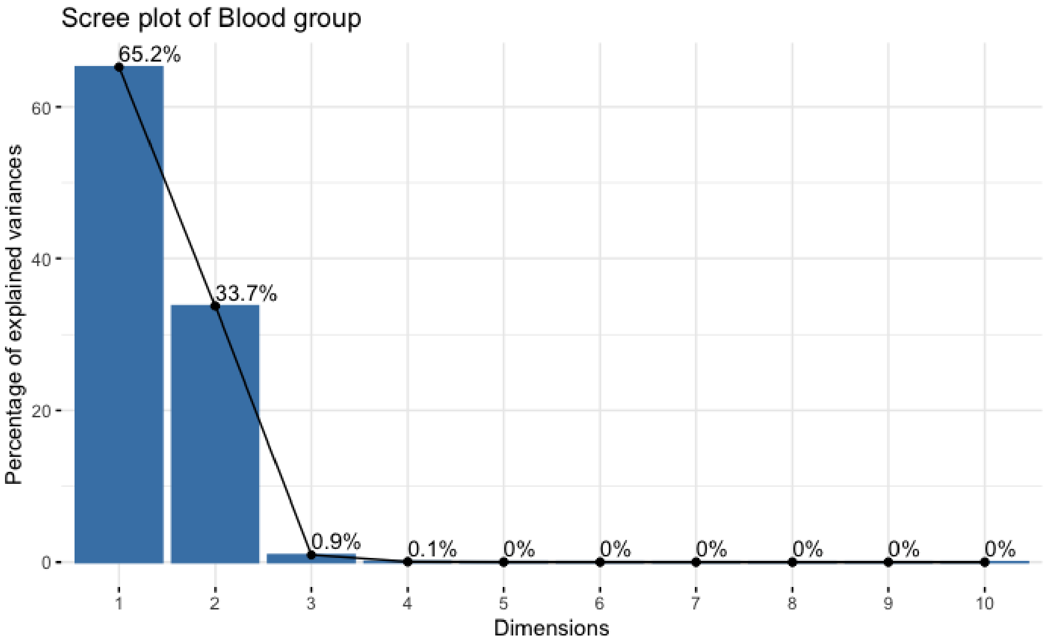


Figure B.447: Screeplot of the blood variables with PCA performed using the covariance matrix

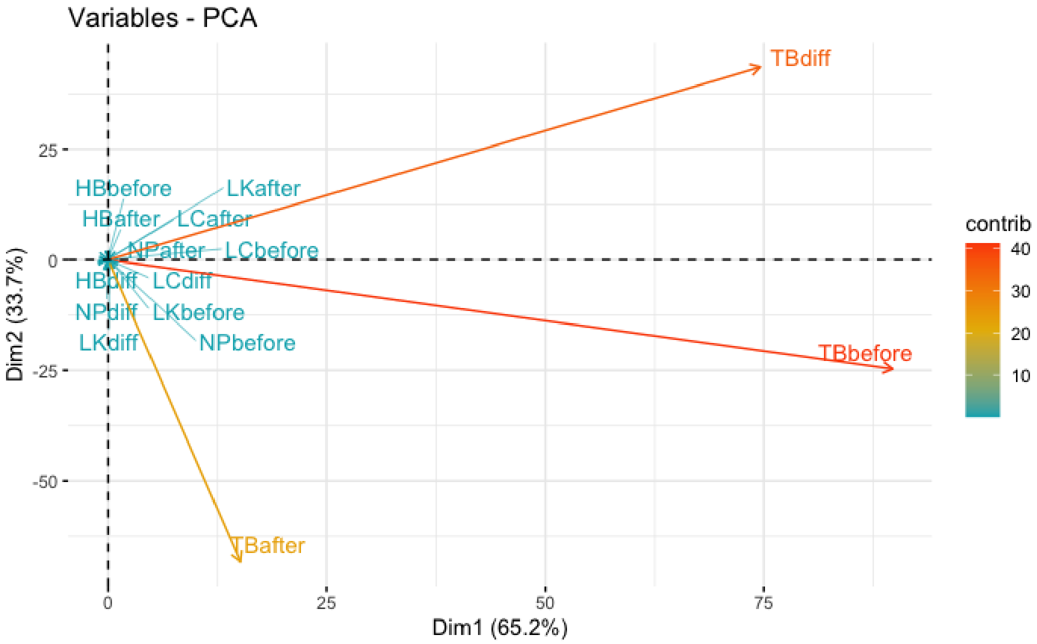


Figure B.448: Loading plot of the blood variables with PCA performed using the covariance matrix

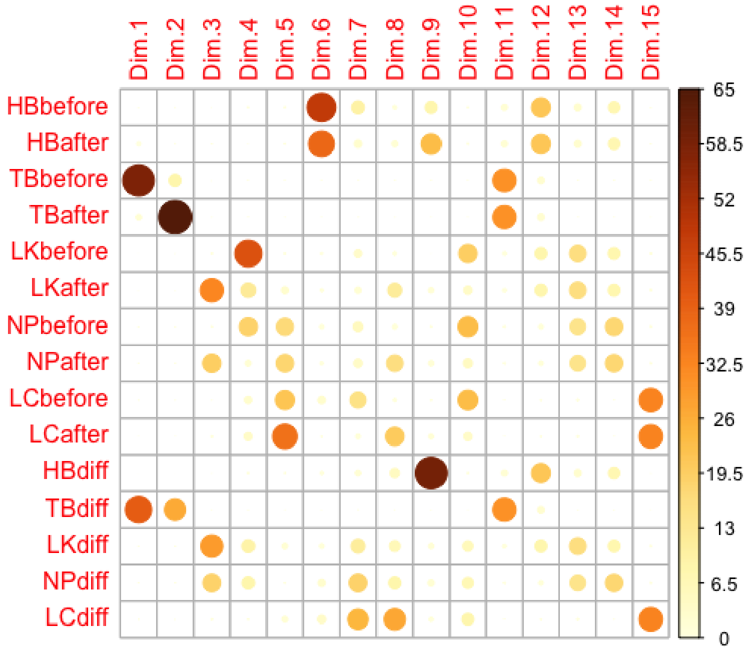


Figure B.449: Contribution plot of the original blood variables with each principal component using PCA performed with the covariance matrix

B.5.2 White Blood Cells

After performing PCA on the centered covariance matrix of the White blood cell data set, the analysis reveals that PC1, PC2 and PC3 account for 92.3% 5.6%, and 1.3% of the total variation in the data, respectively. The cumulative proportion of variance explained of the first two dimensions accounts for 98% of the total variation in the data. The first principal component exhibits a strong correlation with the variables: LKafter, LKdiff, NPafter and NPdiff, indicating that the most variability is observed in the Leukocytes and Neutrophils. It is noteworthy that Neutrophils are a type of Leukocyte, suggesting that the variation in Leukocytes is likely dominated by changes in Neutrophils in most patients. Despite this finding, the prediction ellipses exhibit a high degree of overlap.

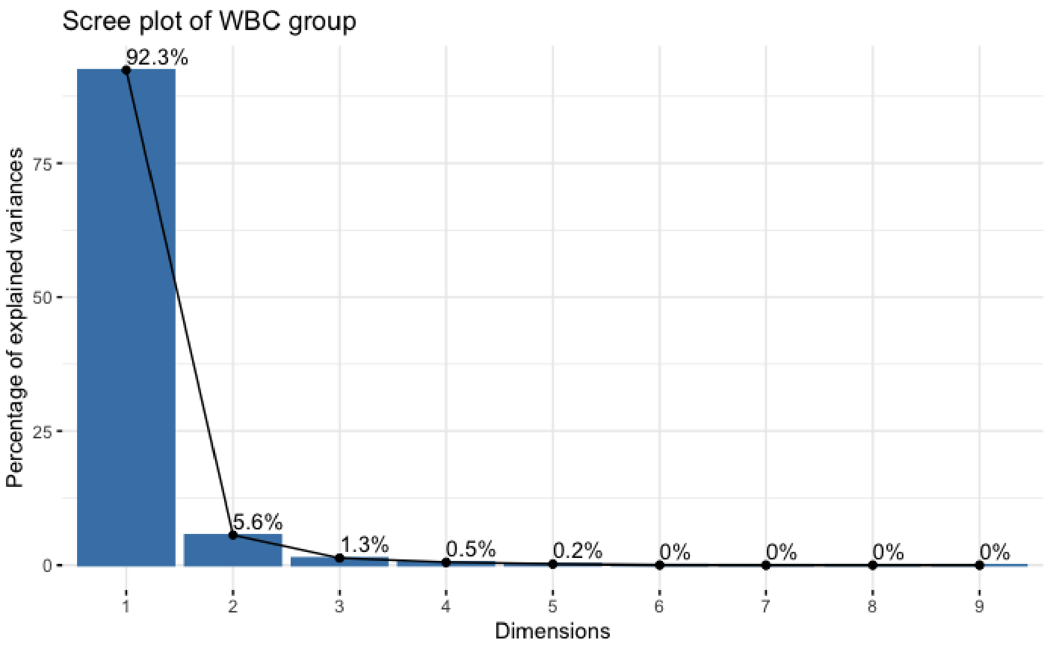


Figure B.450: Screeplot of the white blood cell variables with PCA performed using the covariance matrix

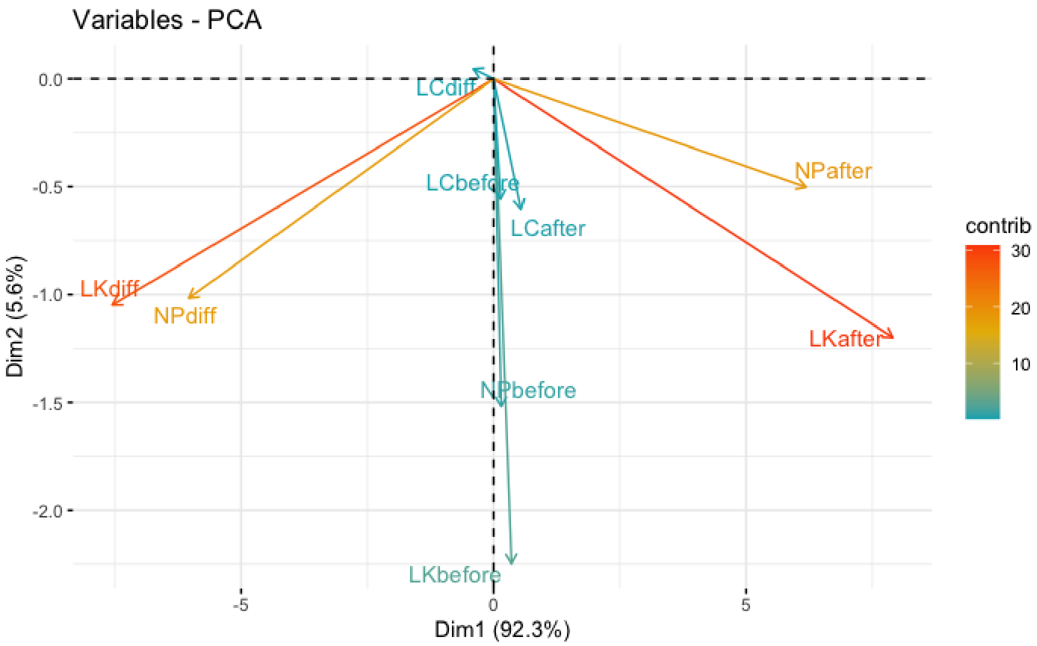


Figure B.451: Loading plot of the white blood cell variables with PCA performed using the covariance matrix

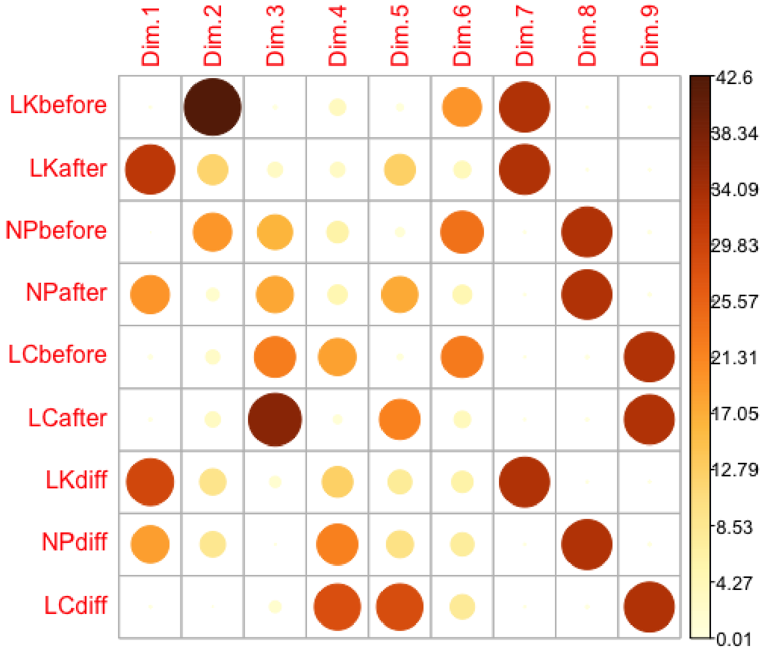


Figure B.452: Contribution plot of the original white blood cell variables with each principal component using PCA performed with the covariance matrix

B.5.3 Kidney

Upon analyzing the kidney group using PCA, it is observed that similar to the white blood cell data set, PC1 captures a significant portion of the total variation in the data. That is, PC1 captures 92.4% of the total variation in the data, followed by PC2 and PC3 capturing 5.2% and 1.5%, respectively. Therefore, PC1 alone would be sufficient in most cases, and the cumulative proportion of variance explained of PC1 and PC2 accounts for 98% of the total variation in the data. The results show that PC1 is highly correlated with CRafter and CRdiff values, while PC2 has the most significant contribution from CRbefore and PC3 from GFRafter and GFRdiff. These findings suggest that creatinine is the most variable value in the kidney data set. The prediction ellipses exhibit an elongated shape along PC2 and significant overlap.

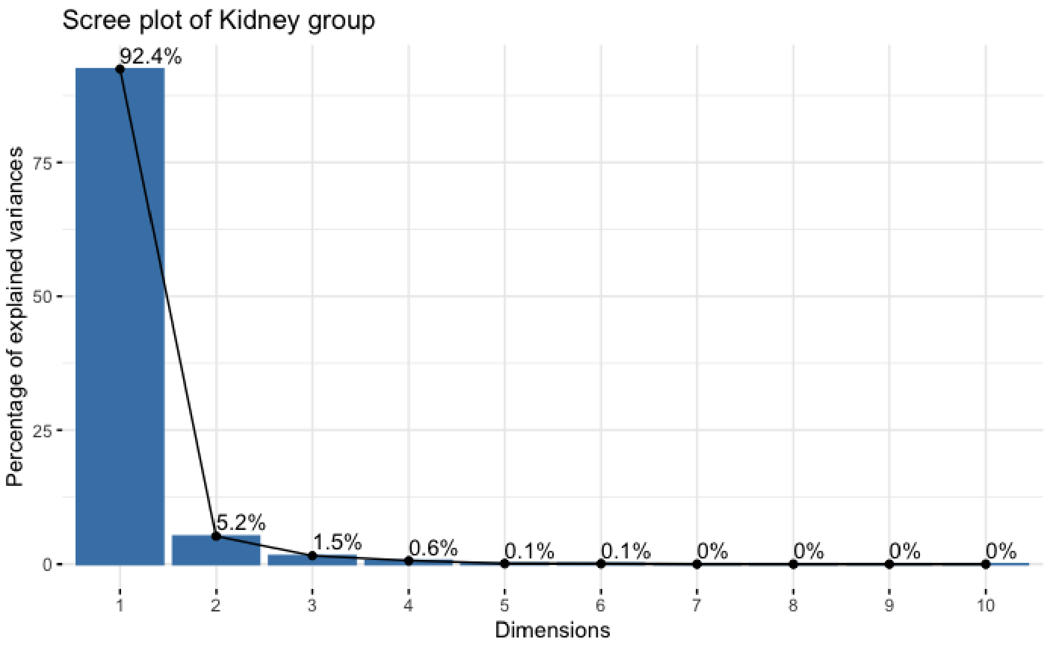


Figure B.453: Screeplot of the kidney variables with PCA performed using the covariance matrix

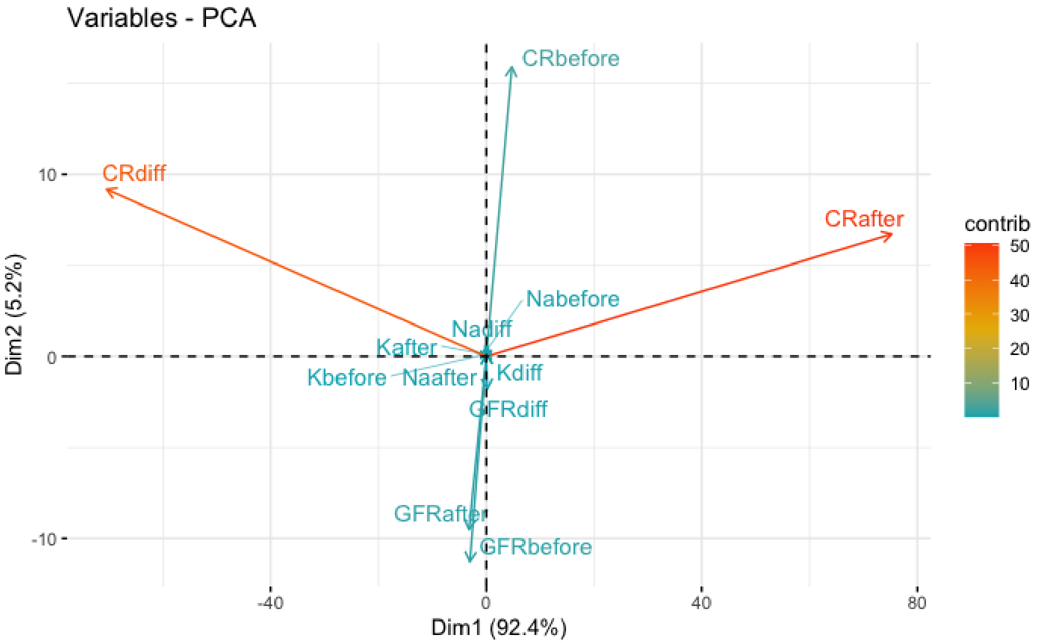


Figure B.454: Loading plot of the kidney variables with PCA performed using the covariance matrix

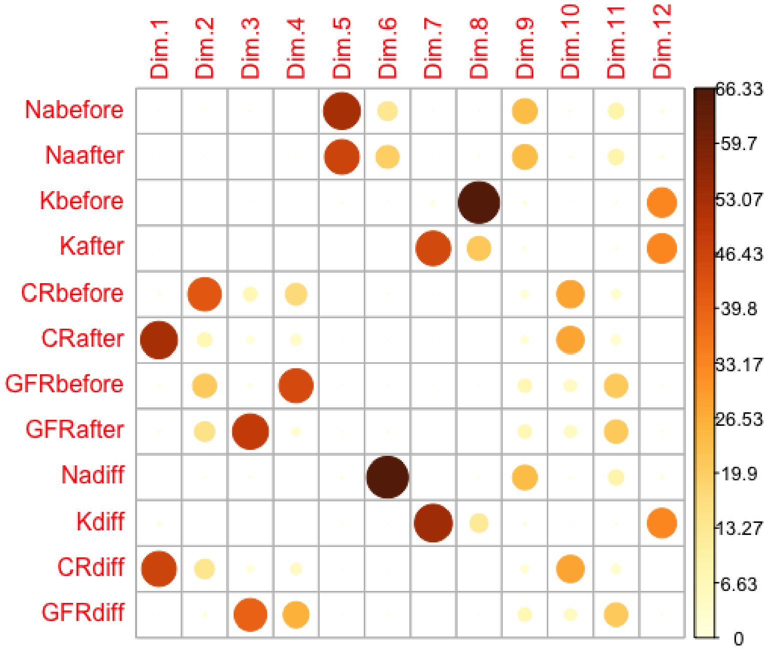


Figure B.455: Contribution plot of the original kidney variables with each principal component using PCA performed with the covariance matrix

B.5.4 Liver

Additionally, in the liver data set, PCA analysis performed on the unscaled variables revealed that PC1 captures 56% of the total variation, PC2 captures 24%, PC3 captures 11% and PC4 captures 4%. PC1 is highly correlated with GGTbefore, indicating that γ-GT is the most variable value in the liver data group, followed by AF and ALAT values. The prediction ellipses still exhibit complete overlap of the disese control group within the progressive disease group ellipse.

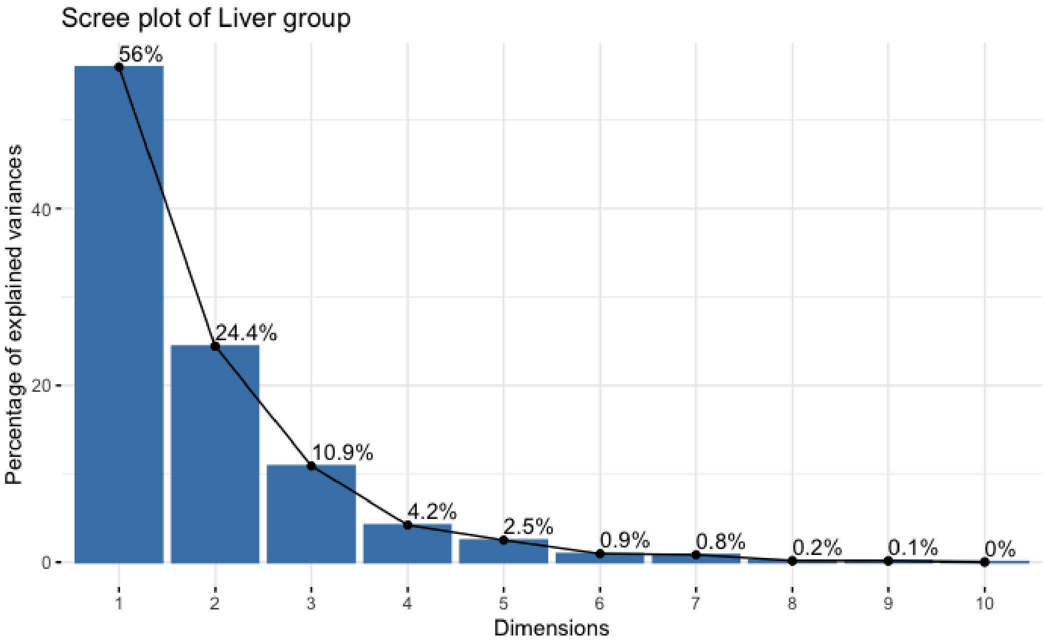


Figure B.456: Screeplot of the liver variables with PCA performed using the covariance matrix

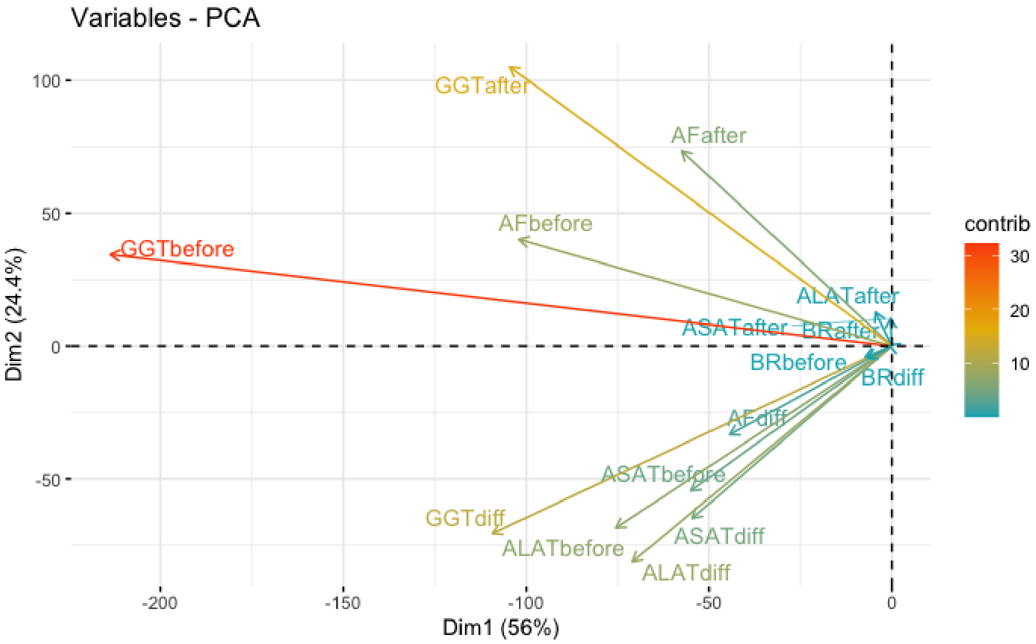


Figure B.457: Loading plot of the liver variables with PCA performed using the covariance matrix

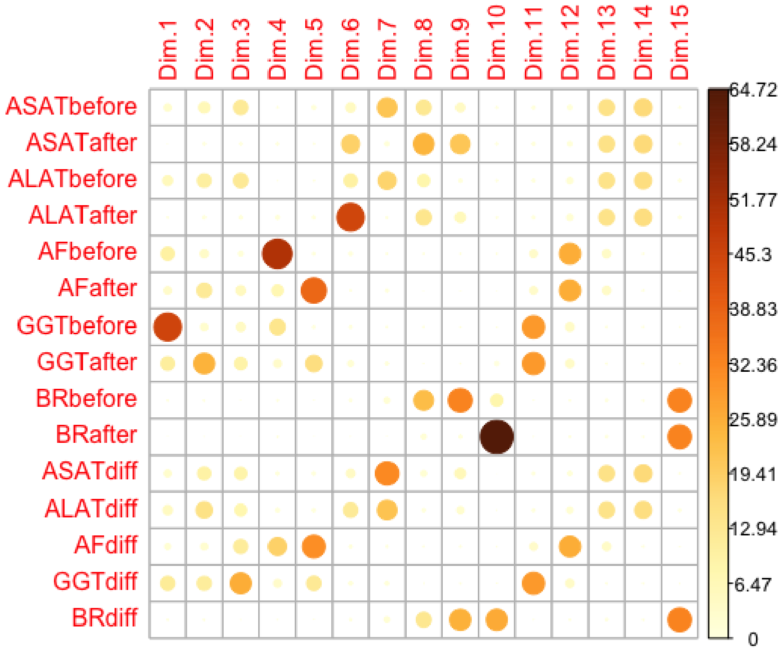


Figure B.458: Contribution plot of the original liver variables with each principal component using PCA performed with the covariance matrix

B.5.5 Nutrition

In the case of the nutrition data set, PCA analysis did not yield a significant improvement using the correlation approach. However, ommitting scaling reveals that PC1 captures 50% of the total variation, followed by PC2 capturing 29.5%, PC3 capturing 11.5% and PC4 capturing 8.4%. Together, using the first three PCs the cumulative proportion is 91% and if PC4 is added it comes to 99% approximately. The results show that PC1 is highly correlated with Albafter, with a smaller but still significant contribution of Albbefore, while PC2 is mainly dominated by Albbefore. This suggests that albumin values exhibit the most significant variation in the nutrition data. The biggest contributors to dimensions 3 and 4 are Nabefore, Naafter for PC3 and Nadiff for PC4, respectively, indicating that sodium values differ second most. The prediction ellipses show almost complete overlap and are of approximately the same size.

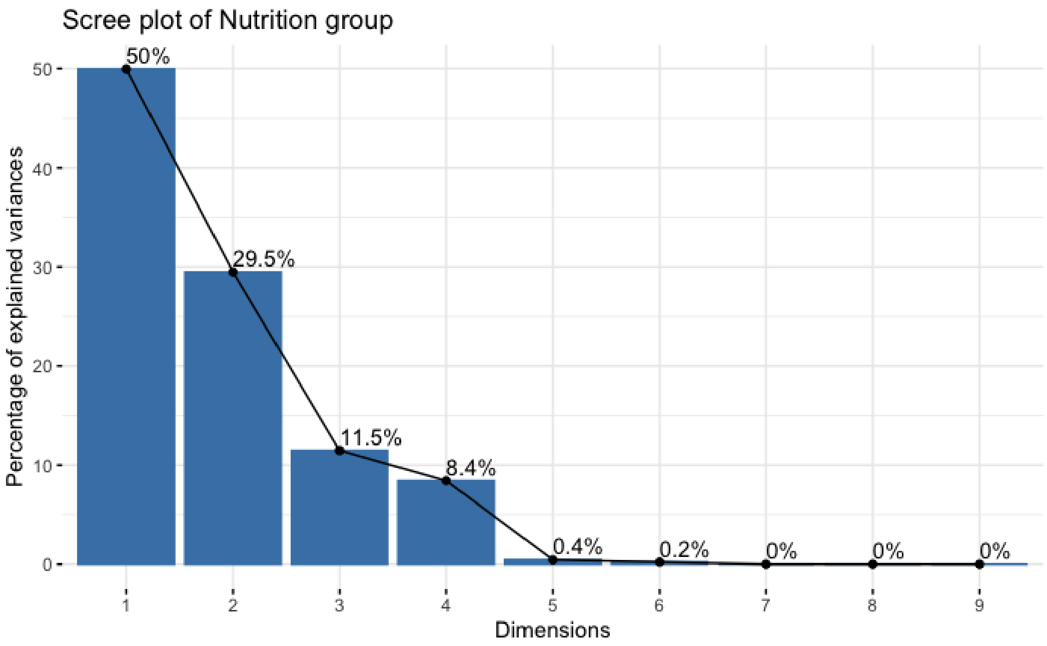


Figure B.459: Screeplot of the nutrition variables with PCA performed using the covariance matrix

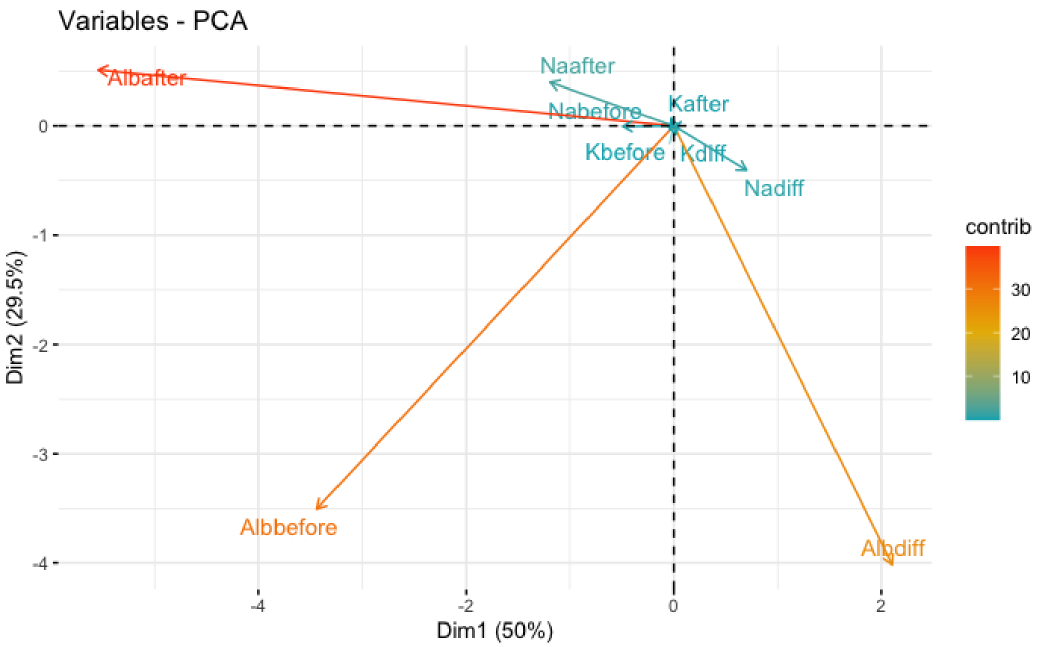


Figure B.460: Loading plot of the nutrition variables with PCA performed using the covariance matrix

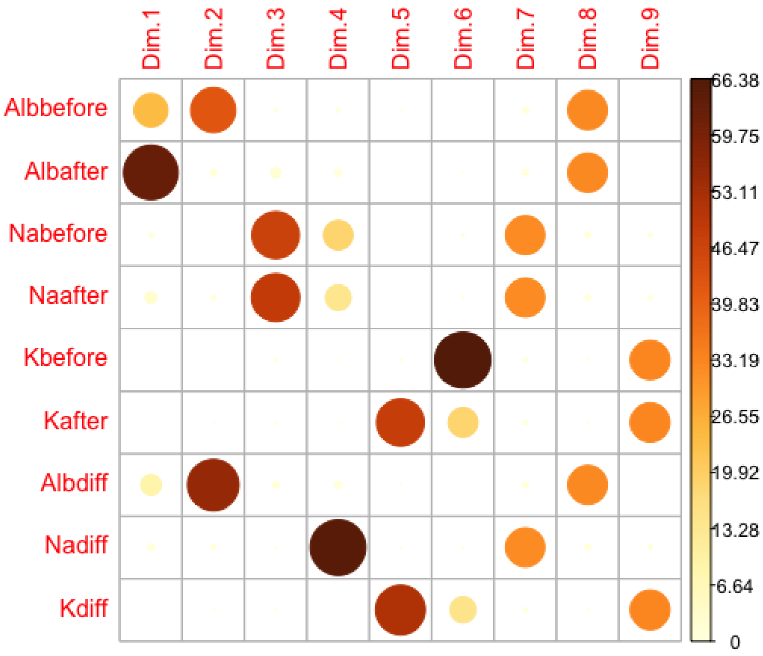


Figure B.461: Contribution plot of the original nutrition variables with each principal component using PCA performed with the covariance matrix

B.5.6 Inflammation

For the inflammation dataset, two principal components were sufficient to capture 99% of the total variation in the data. The scree plot in Figure B.462 shows that PC1 is able to capture 69.7% of the total variation and PC2 29.7%, with the first dimension highly correlated to SIIafter and SIIdiff, and the second dimension highly correlated to SIIbefore. This indicates that inflammation is mainly dictated by the SII values, is a commonly used inflammation index and uses (B.4) values of platelets, neutrophils and lymphocytes. The biggest contributor to PC1 is SII diff, to PC2 is SIIbefore and to PC3 is PLRafter and PLRdiff. The prediction ellipses still show overlap, with the progressive disease ellipse more circular in both directions compared to the disease control group.

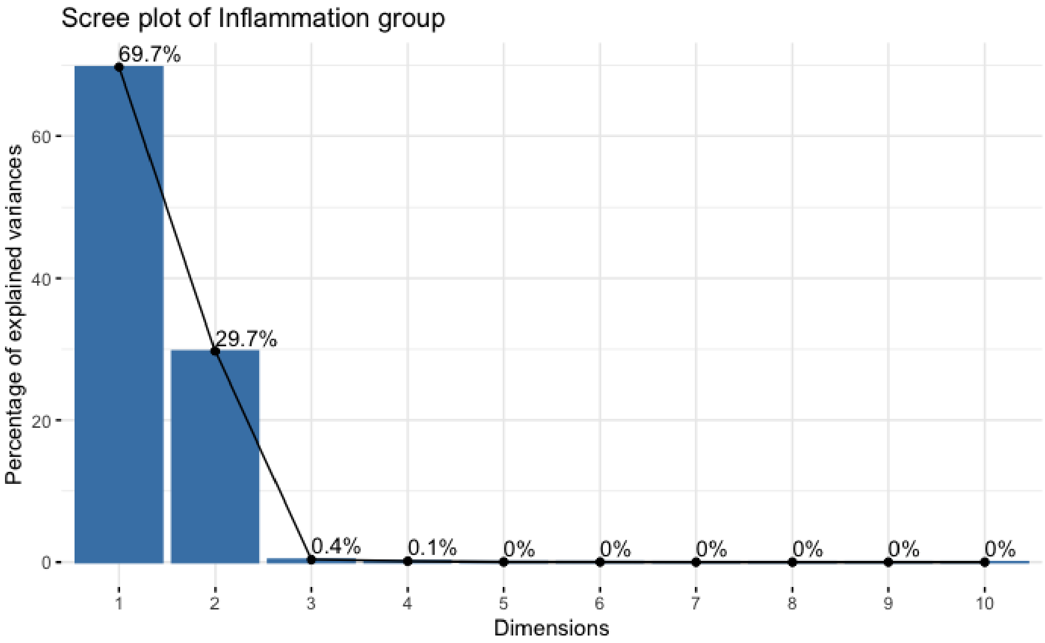


Figure B.462: Screeplot of the inflammation variables with PCA performed using the covariance matrix

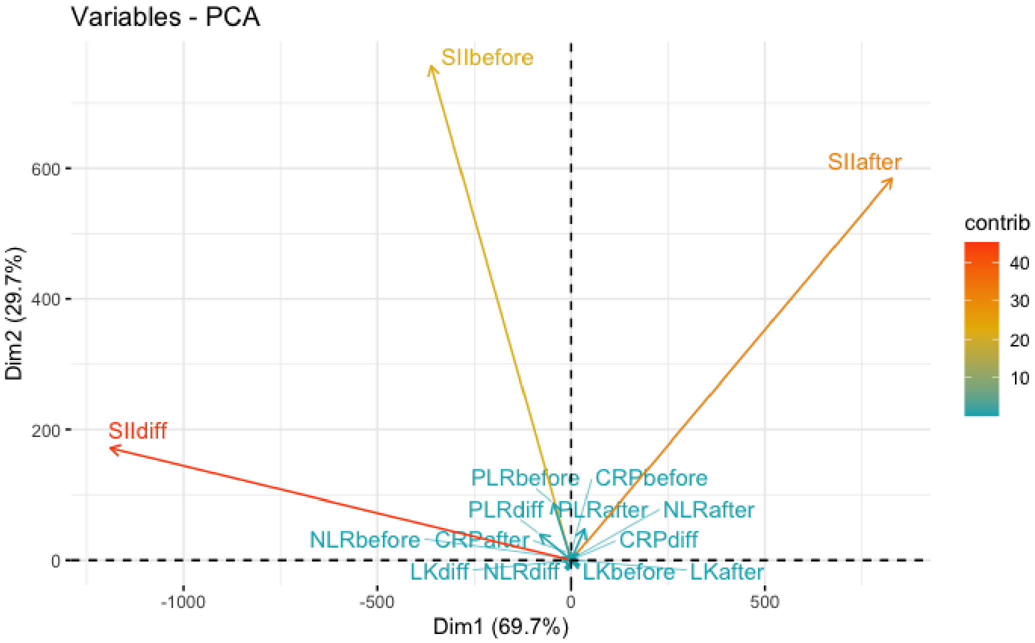


Figure B.463: Loading plot of the inflammation variables with PCA performed using the covariance matrix

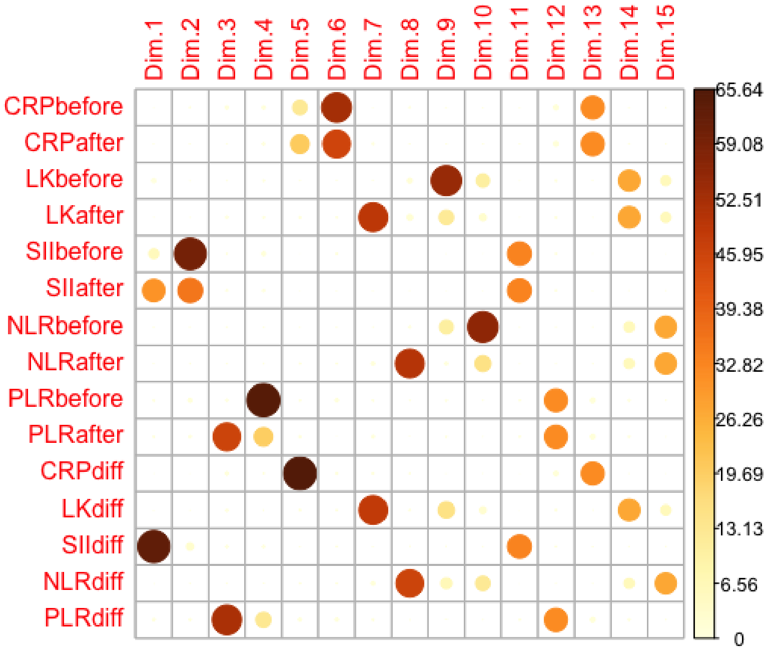


Figure B.464: Contribution plot of the original inflammation variables with each principal component using PCA performed with the covariance matrix

B.5.7 Patient Characteristics

The patient characteristic data set only has four variables and exhibits a similar trend as the inflammation dataset after PCA is performed without scaling. The screeplot in Figure B.465 shows that PC1 is able to capture 67.3% of the total variation, PC2 17.2% and PC3 15.4%. Together these three PCs can explain 99.96% of the total variation, rounded off it would be 100%. Interestingly, the first dimension is extremely correlated to the Weight variable, while the second dimension has the biggest contribution of Age and the third dimension from height. The prediction ellipses also overlap with the progressive disease ellipse almost circular and the disease control ellipse more elongated along the x-axis. This means that disease control has more variation in PC1, which is mostly dominated by Weight.

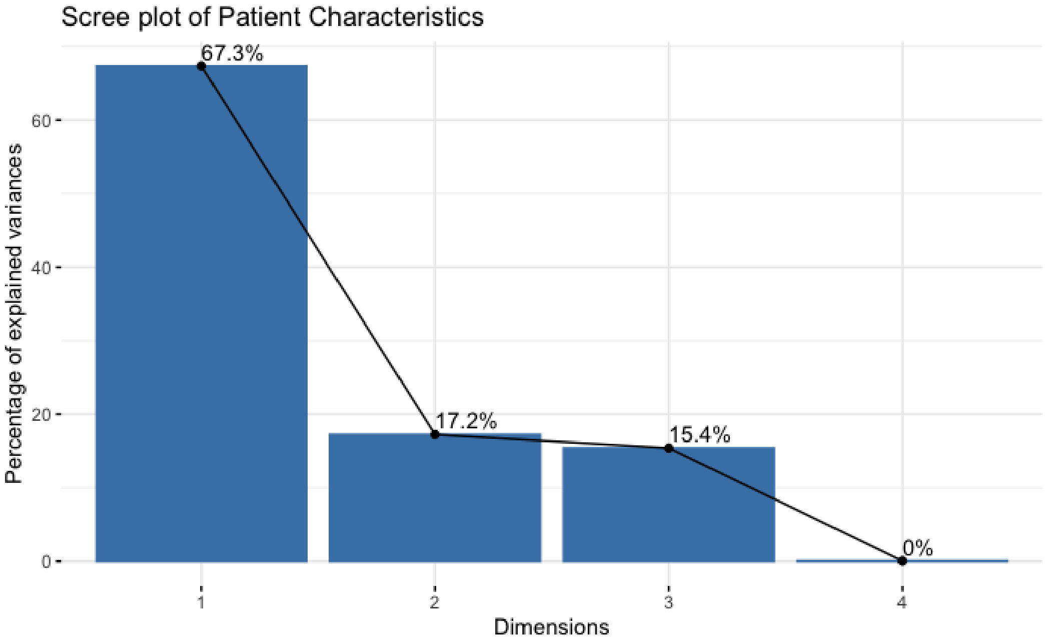


Figure B.465: Screeplot of the patient characteristics variables with PCA performed using the covariance matrix

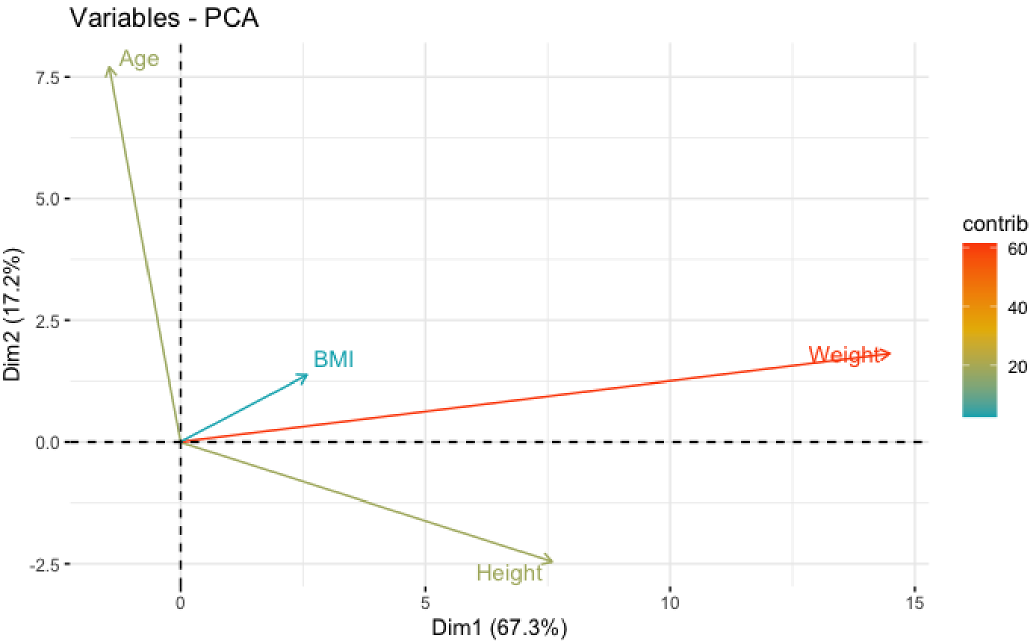


Figure B.466: Loading plot of the patient characteristics variables with PCA performed using the covariance matrix

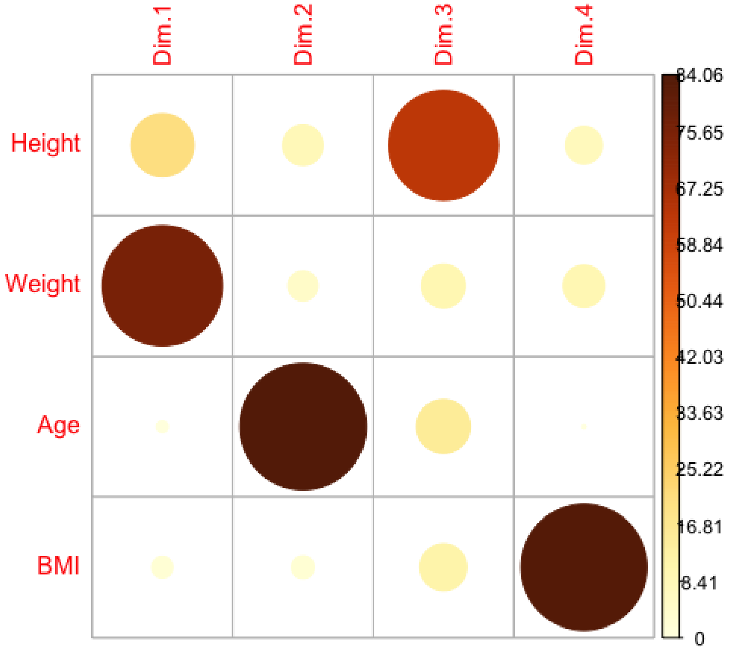


Figure B.467: Contribution plot of the original patient characteristics variables with each principal component using PCA performed with the covariance matrix

B.5.8 Tumor Markers

For the tumor markers dataset, only one principal component was sufficient to capture 97.1% of the total variation in the data, with CA19-9 values after exhibiting the highest correlation. This aligns with previous research and literature findings that CA19-9 is an important marker for pancreatic cancer. In more detail, the screeplot in Figure B.468 shows that PC1 is able to capture 97.1% of the total variation in the data, with PC2 following with a mere 2.4%. PC1 is highly correlated to CA19-9 values after, followed by CA19-9 values before and then CA19-9 values at diagnosis. However, it is important to keep in mind that 10% of the population is unable to produce this marker and these patients have been excluded from this analysis. The contribution to the second dimension is also still from CA19-9 diagnosis, and the third dimension from CA19-9 difference. CEA before chemotherapy is the most variable dimension is CA19-9 is not taken into account. The prediction ellipses do show that the disease control group has a much smaller ellipse compared to the progressive disease group.

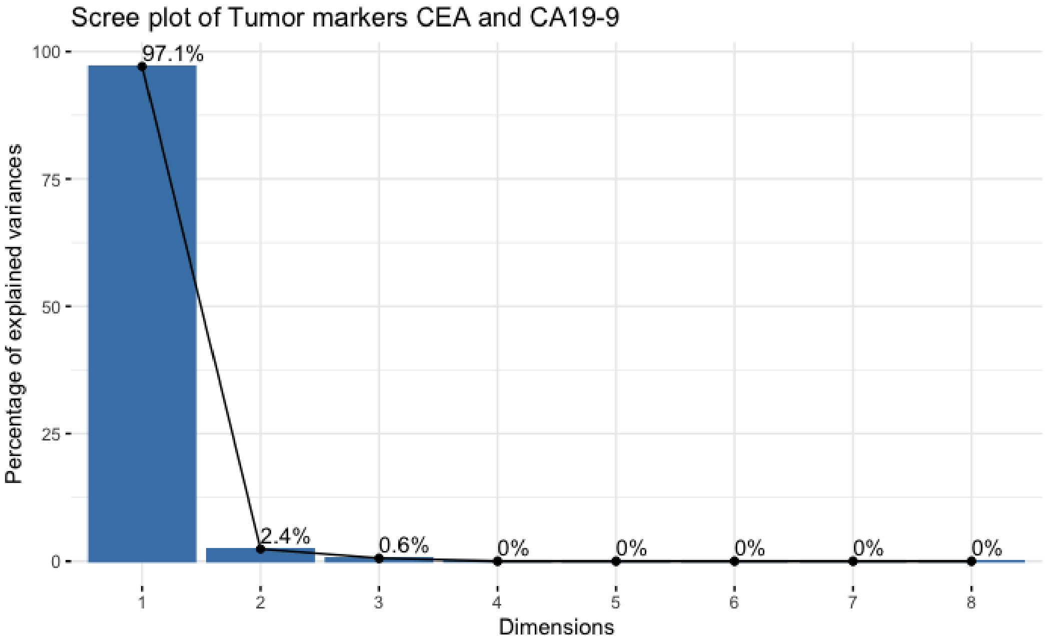


Figure B.468: Screeplot of the tumor marker variables with PCA performed using the covariance matrix

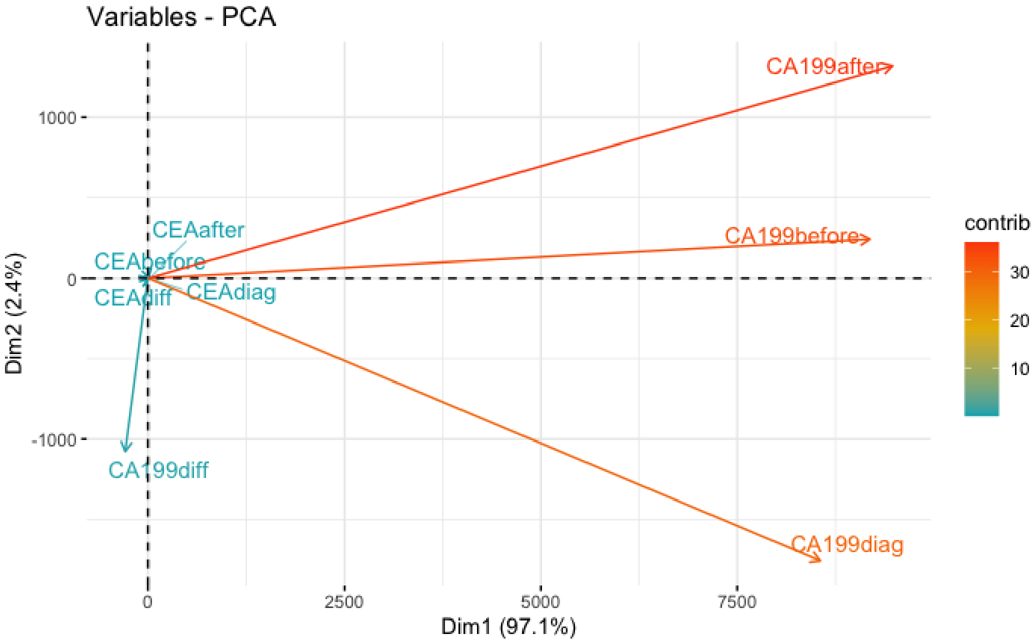


Figure B.469: Loading plot of the tumor marker variables with PCA performed using the covariance matrix

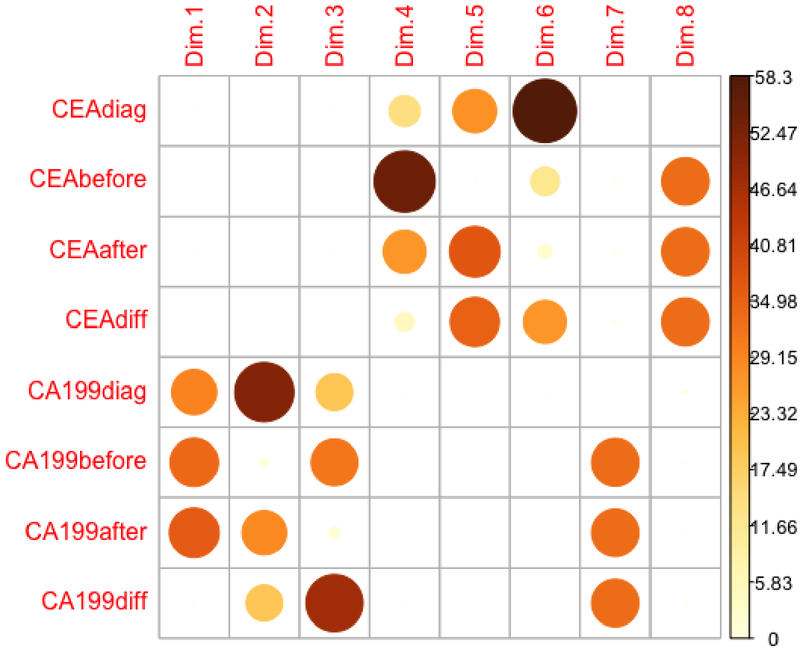


Figure B.470: Contribution plot of the original tumor marker variables with each principal component using PCA performed with the covariance matrix

B.5.9 All markers

Next, consider the all markers data set, which includes all the blood and tumor variables. This data set exhibits a similar trend as the tumor markers dataset, with PC1 capturing 96.9% of the total variation in the data (recall that the data has 34 variables) and PC2 following with 2.7%. The most important variables are CA19-9 after, CA19-9before which are extremely correlated to PC1. Other variables that follow are GGTbefore, GGTafter, AFafter, ALATbefore and TBbefore. Similar to before, the prediction ellipse of the disease control group is completely captured within the prediction ellipse of the progressive disease group.

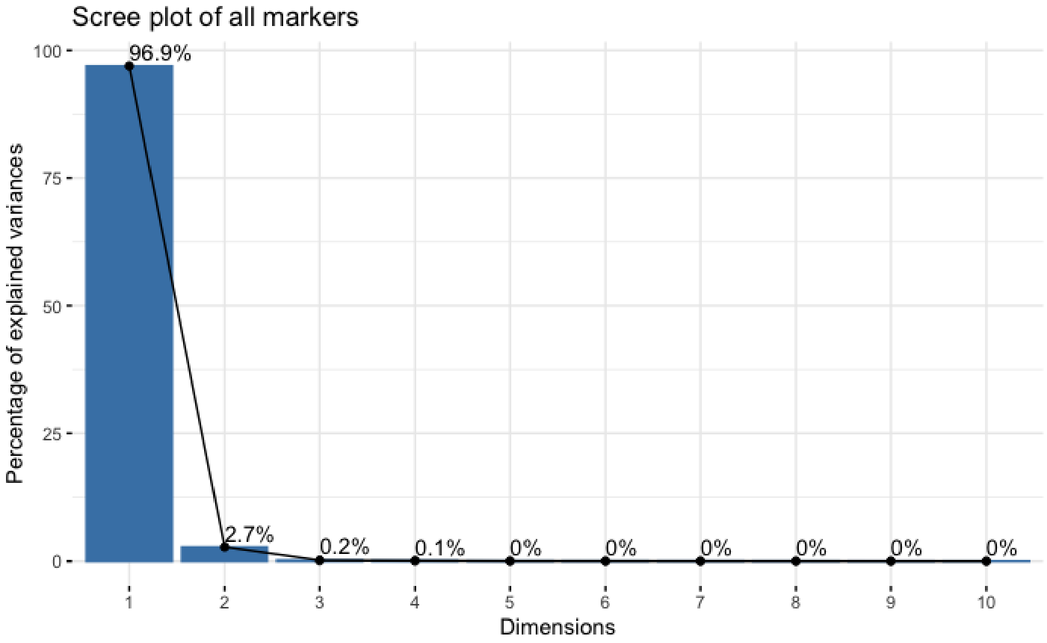


Figure B.471: Screeplot of the all the measured variables with PCA performed using the covariance matrix

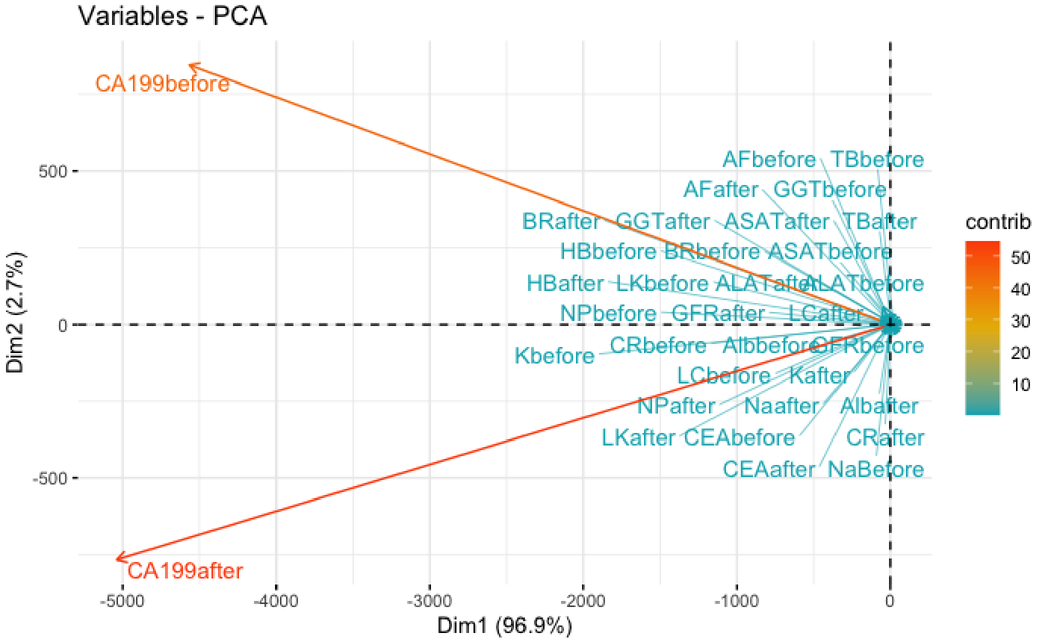


Figure B.472: Loading plot of all the measured variables with PCA performed using the covariance matrix

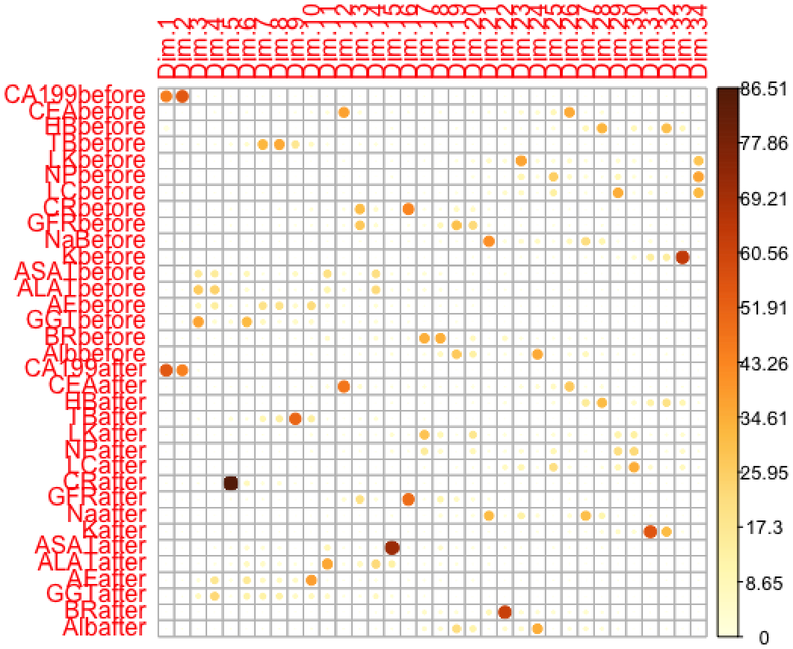


Figure B.473: Contribution plot of the original measured variables with each principal component using PCA performed with the covariance matrix

B.5.10 All markers + Age + BMI + differences

Furthermore, when age, BMI, and the differences before and after chemotherapy were added to the dataset as separate variables, PC1 remained the most important component and could explain 91.8% of the total variation in the data. The most important variables were still CA19-9after, CA19-9before, and CA19-9diff, followed by γ-GT, AF, ALAT, and TB. That is, PC1 is able to capture 91.8% of the total variation in the data, PC2 follows with 7.7% and PC3 0.3%. So PC1 would be sufficient and with PC1 and PC2 together we would explain 99% of the total variation in the data. The prediction ellipses are similar to the ones from the all markers data set. As a side note, is CA19-9 is not taken into the data set, CEAafter and ALbdiff seem to be the most important variables.

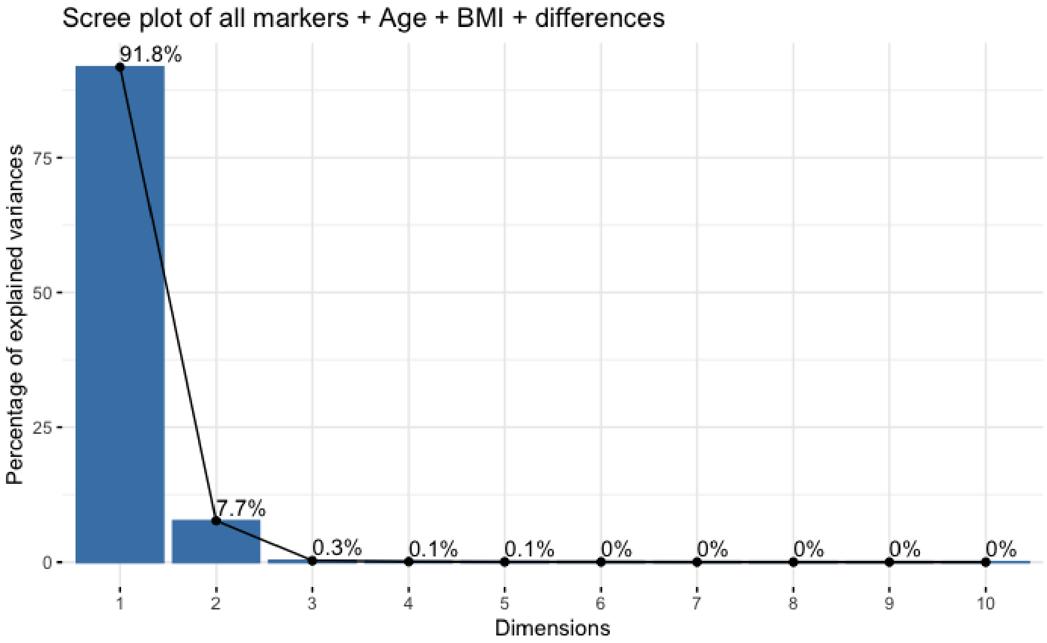


Figure B.474: Screeplot of all the measured variables + Age + BMI + differences between measurements before and after chemotherapy with PCA performed using the covariance matrix

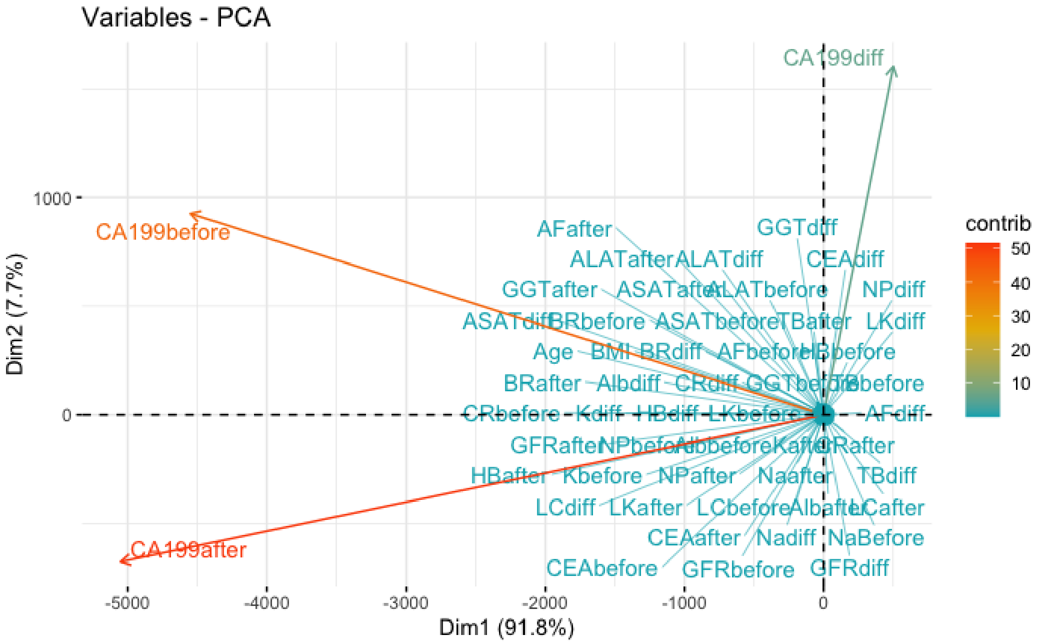


Figure B.475: Loading plot of all the measured variables + Age + BMI + differences between measurements before and after chemotherapy with PCA performed using the covariance matrix

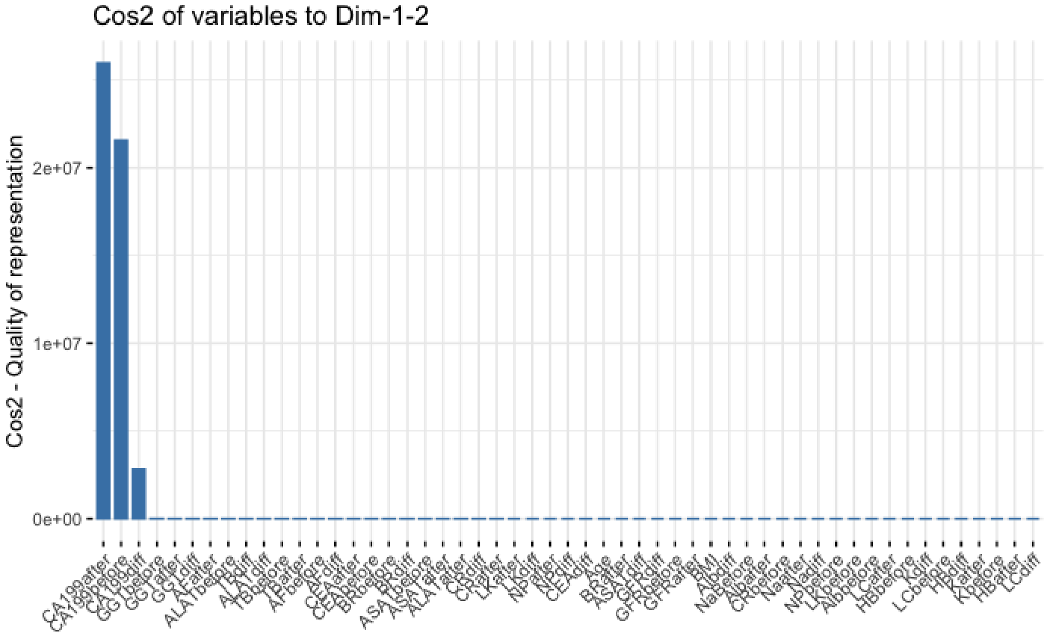


Figure B.476: Cos2 plot of all the original measured variables + Age + BMI + differences between measurements before and after chemotherapy with each principal component using PCA performed with the covariance matrix

B.5.11 Age + BMI + differences

Finally, we consider the data with only the age, BMI and differences before and after chemotherapy as variables. Similar to the previous two analyses, we see that PC1 is able to capture most of the variability in the data, namely 96.3% with PC2 following with 2.5%. Unsurprisingly, the dominant variable is CA19-9diff, followed by the ALATdiff, ASATdiff and GGTdiff. The prediction ellipses show again that the disease control ellipse is centered within the progressive disease ellipse.

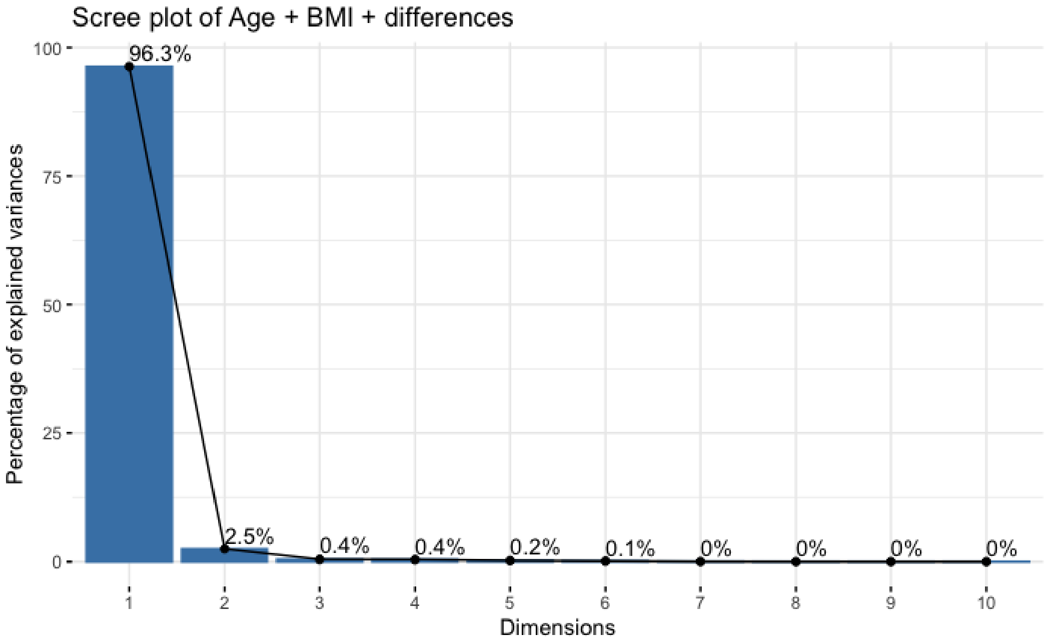


Figure B.477: Screeplot of Age + BMI + differences between measurements before and after chemotherapy variables with PCA performed using the covariance matrix

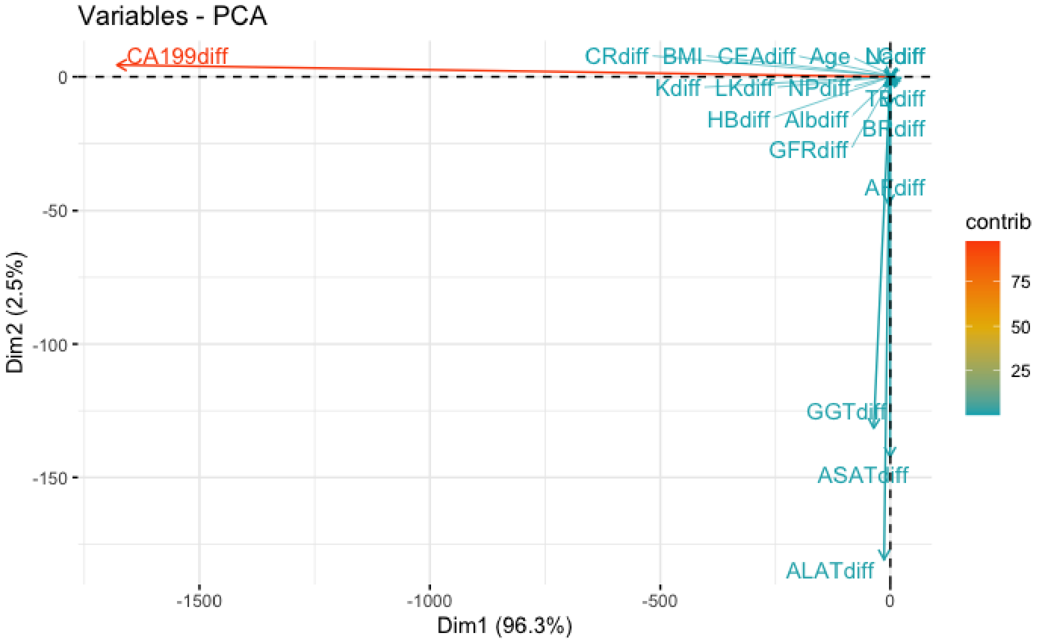


Figure B.478: Loading plot of the Age + BMI + differences between measurements before and after chemotherapy variables with PCA performed using the covariance matrix

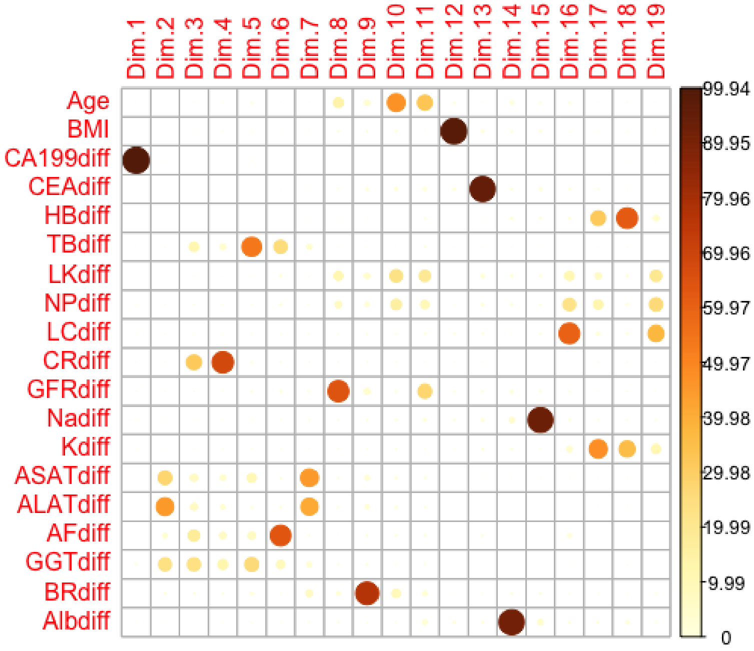


Figure B.479: Contribution plot of the original Age + BMI + differences between measurements before and after chemotherapy variables with each principal component using PCA performed with the covariance matrix

B.6. PCA after removal of outliers (figures)

In this section the plots that have changed after removal of outliers based on GFR and BR values are presented with the interpretation and discussion presented in section 3.3.

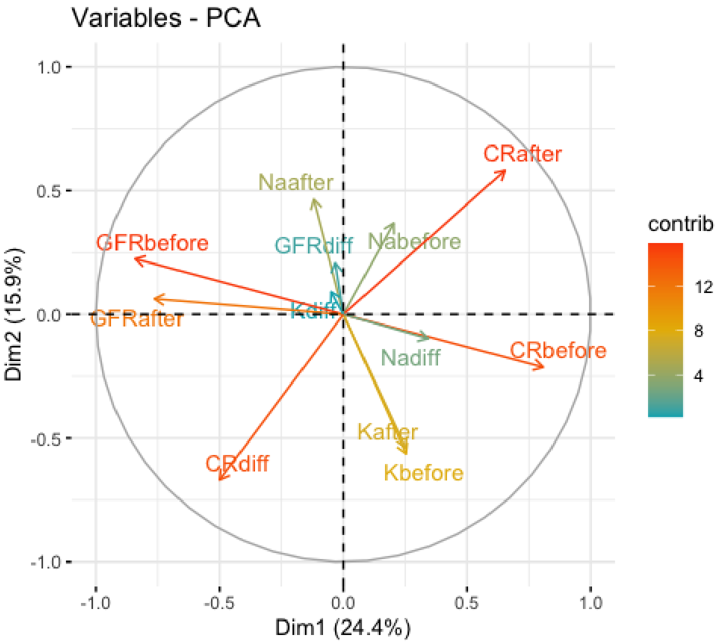


Figure B.480: Loading plot of the kidney variables after removal of outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values prior to chemotherapy, n=139 (Disease control (n=110), Progressive disease (n=29)).

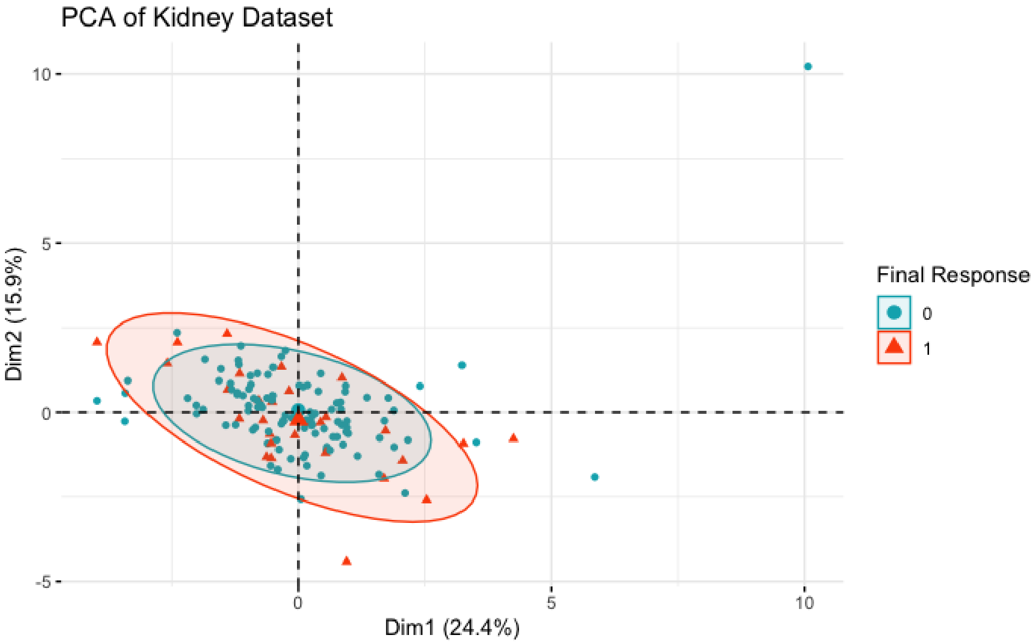


Figure B.481: 95% prediction ellipse of the kidney variables after removal of outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values prior to chemotherapy, n=139 (Disease control (n=110), Progressive disease (n=29)).

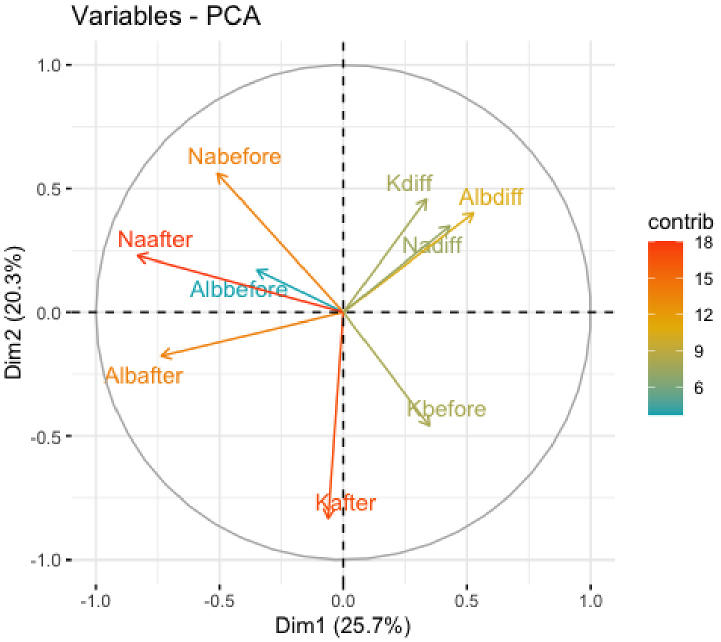


Figure B.482: Loading plot of the nutrition variables after removal of outliers based on BR$> 50 mol/L$ and GFR$< 30 mL/min$ values prior to chemotherapy, n=99 (Disease control (n=78), Progressive disease (n=21)).

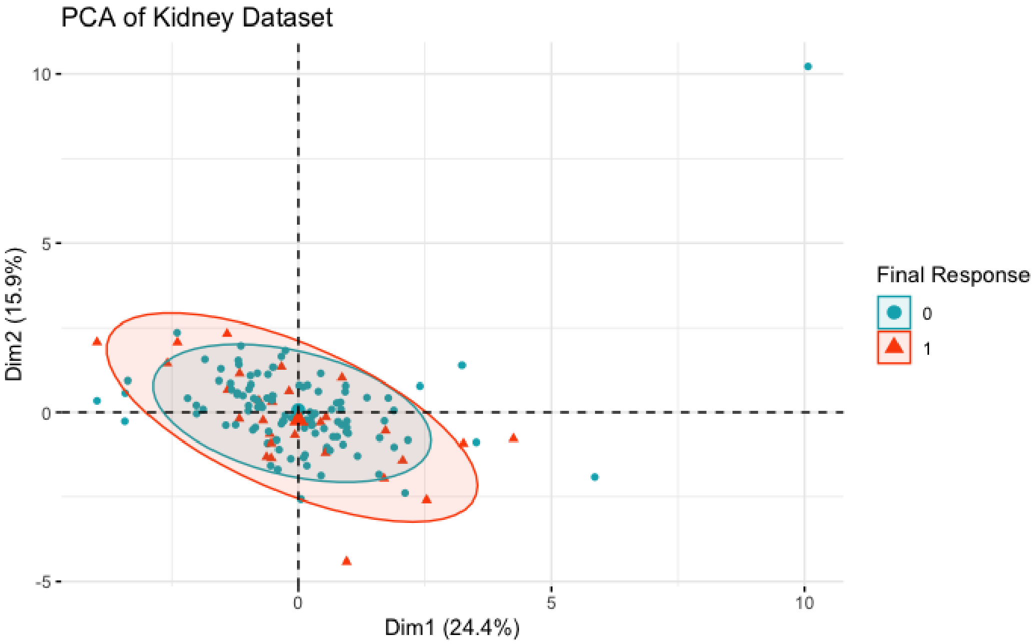


Figure B.483: 95% prediction ellipse of the kidney variables after removal of outliers based on BR$> 50 mol/L$ and GFR$< 30 mL/min$ values prior to chemotherapy, n=99 (Disease control (n=78), Progressive disease (n=21)).

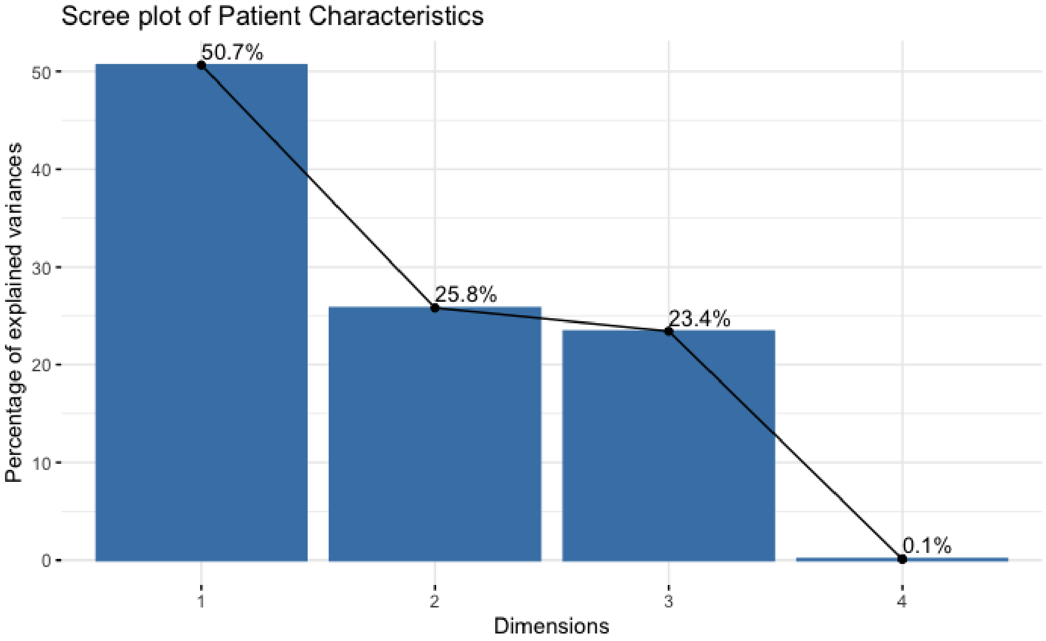


Figure B.484: Scree plot of the patient characteristics after removal of outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values prior to chemotherapy, n=179 (Disease control (n=143), Progressive disease (n=36)).

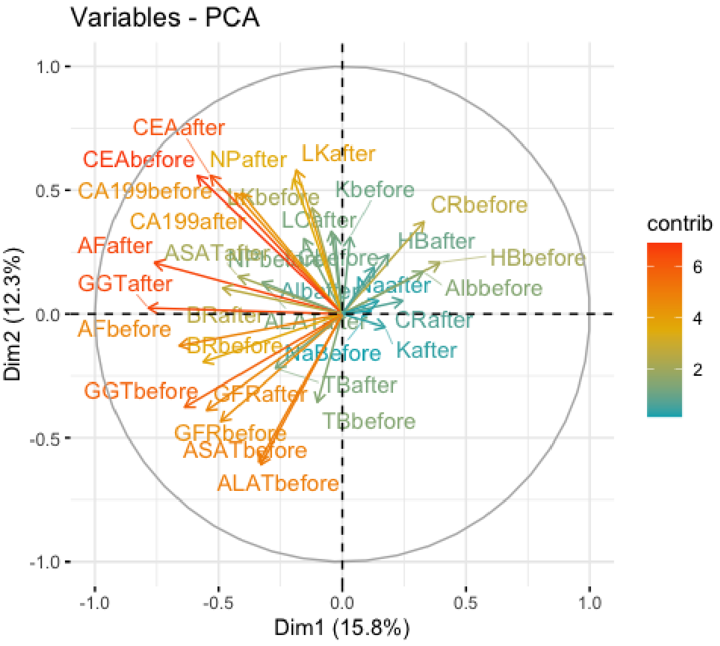


Figure B.485: Loading plot of all the blood and tumor markers after removal of outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values prior to chemotherapy, n=57 (Disease control (n=43), Progressive disease (n=14)).

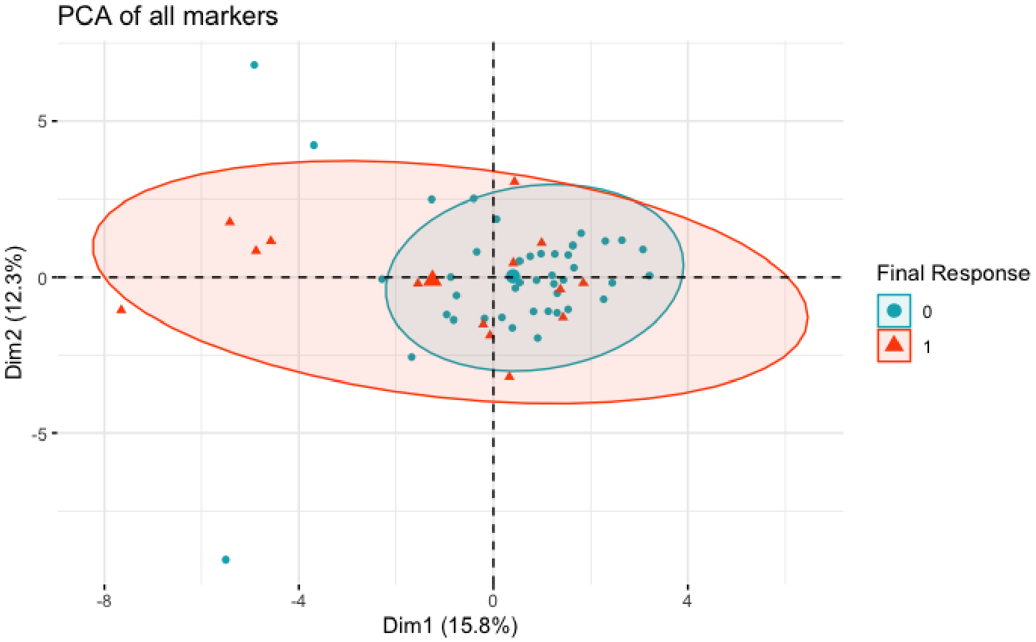


Figure B.486: 95% prediction ellipse of the kidney variables after removal of outliers based on BR> $50mol/L$ and GFR< $30mL/min$ values prior to chemotherapy, n=57 (Disease control (n=43), Progressive disease (n=14)).

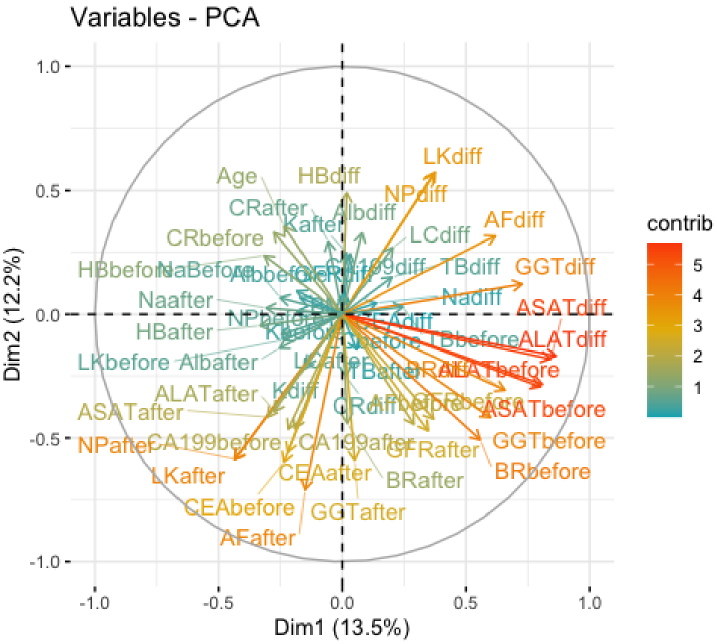


Figure B.487: Loading plot of all the blood and tumor markers including age, BMI and the differences before and after the first chemotherapy cycle after removal of outliers based on BR> $50mol/L$ and GFR< $30mL/min$ values prior to chemotherapy, n=57 (Disease control (n=43), Progressive disease (n=14)).

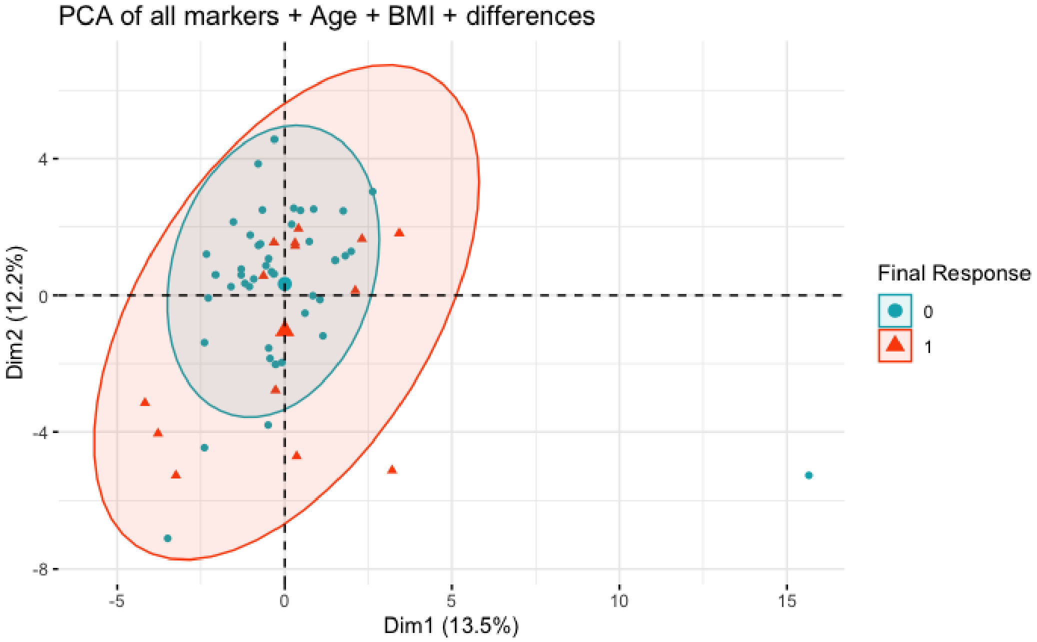


Figure B.488: 95% prediction ellipse of all the blood and tumor markers including age, BMI and the differences before and after the first chemotherapy cycle after removal of outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values prior to chemotherapy, n=57 (Disease control (n=43), Progressive disease (n=14)).

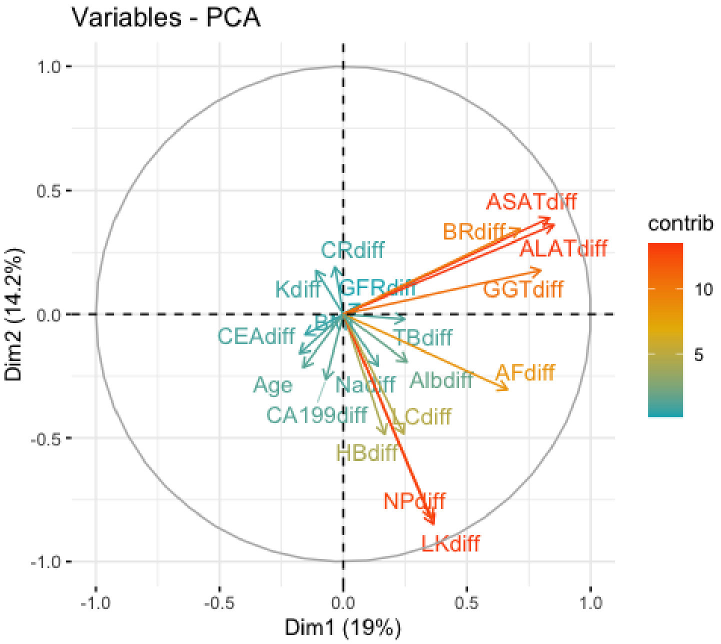


Figure B.489: Loading plot of all age, BMI and the differences before and after the first chemotherapy cycle after removal of outliers based on BR$> 50mol/L$ and GFR$< 30mL/min$ values prior to chemotherapy, n=57 (Disease control (n=43), Progressive disease (n=14)).

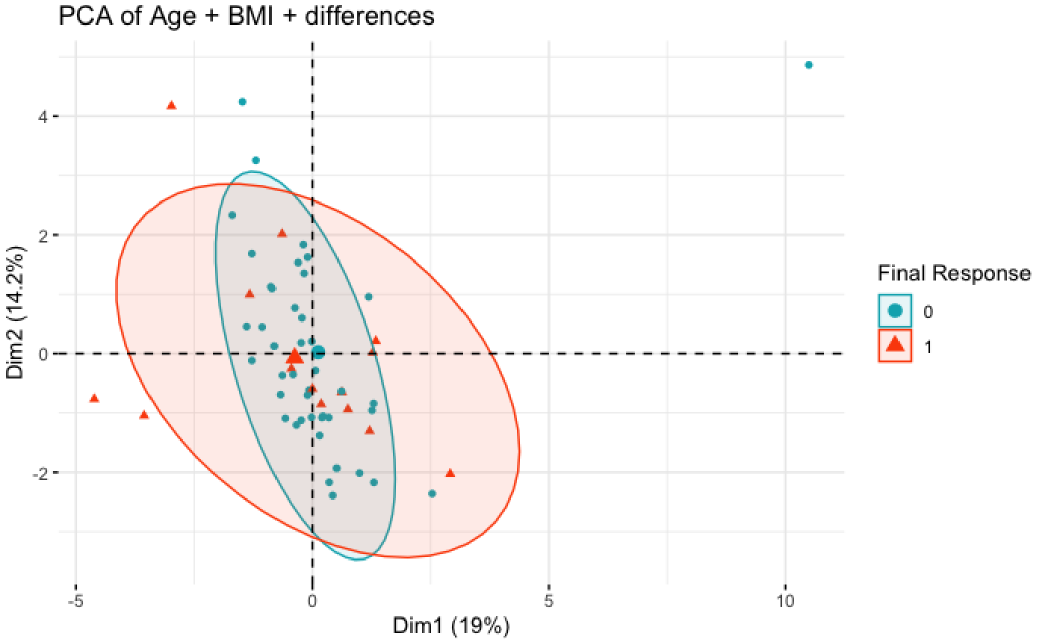


Figure B.490: 95% prediction ellipse of age, BMI and the differences before and after the first chemotherapy cycle after removal of outliers based on BR$> 50 mol/L$ and GFR$< 30 mL/min$ values prior to chemotherapy, n=57 (Disease control (n=43), Progressive disease (n=14)).

C | Random Forest

C.1. Decision Trees

This Appendix contains examples of decisions trees created using the different training datasets considered in section 5.

C.1.1 Decision Trees (full dataset)

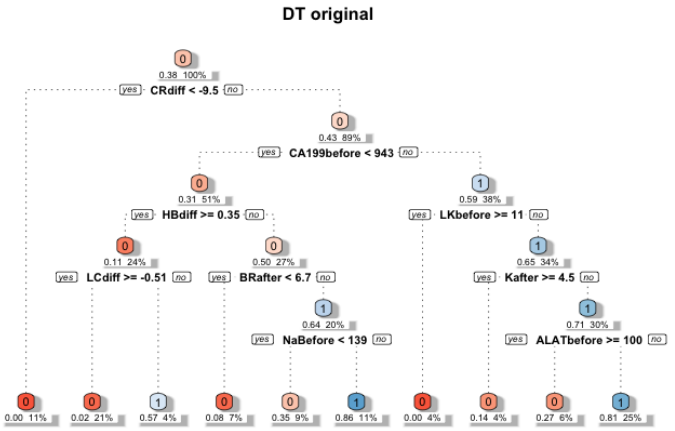


Figure C.1: Decision tree created using the original training dataset, n=194, 0=Disease Control (n=120), 1=Progressive disease (n=74).

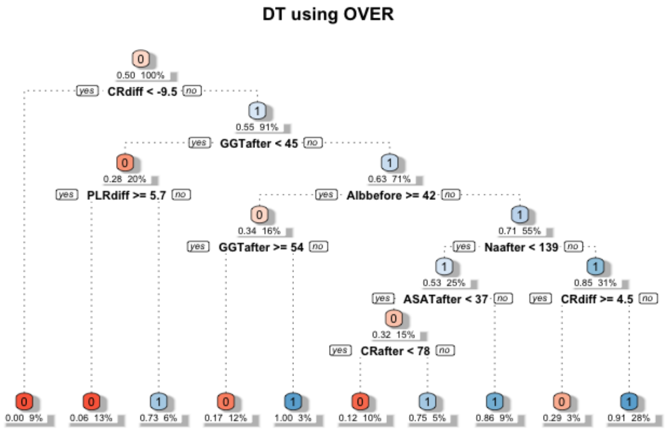


Figure C.2: Decision tree created using the oversampled training dataset, n=240, 0=Disease Control (n=120), 1=Progressive disease (n=120).

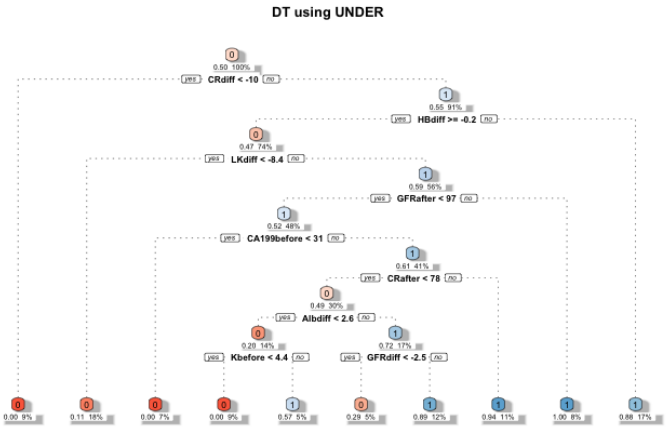


Figure C.3: Decision tree created using the undersampled training dataset, n=148, 0=Disease Control (n=74), 1=Progressive disease (n=74).

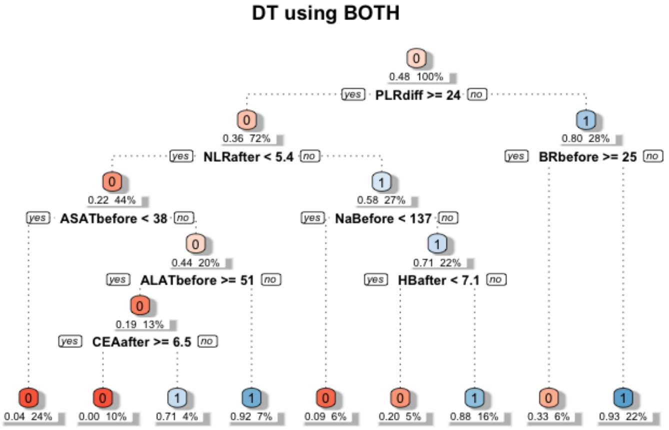


Figure C.4: Decision tree created using both under- and oversampled training dataset, n=194, 0=Disease Control (n=100), 1=Progressive disease (n=94).

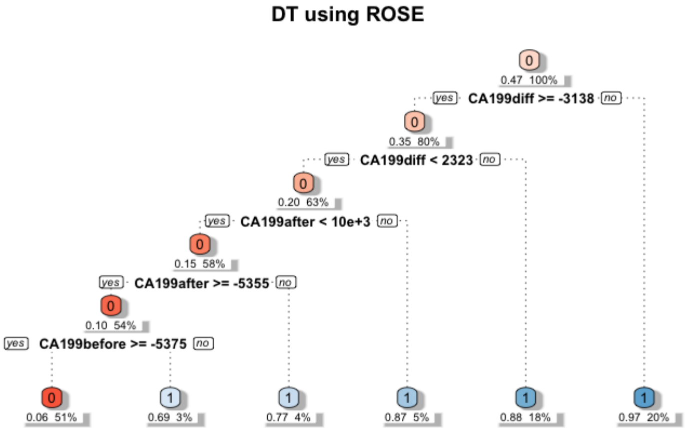


Figure C.5: Decision tree created using the ROSE training dataset, n=500, 0=Disease Control (n=265), 1=Progressive disease n=235.

C.1.2 Decision Trees (no outliers)

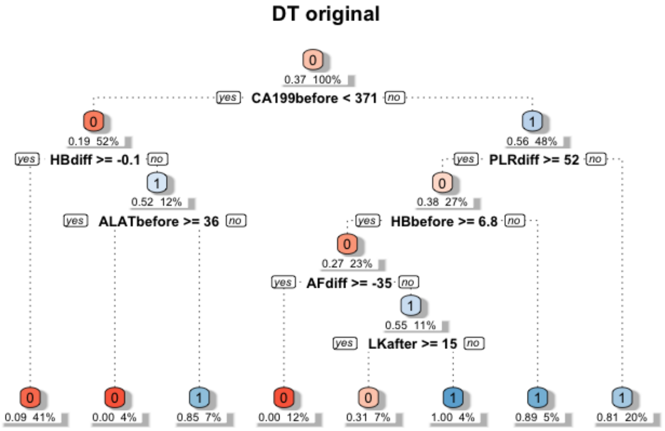


Figure C.6: Decision tree created using the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, 0=Disease Control (n=115), 1=Progressive disease (n=67).

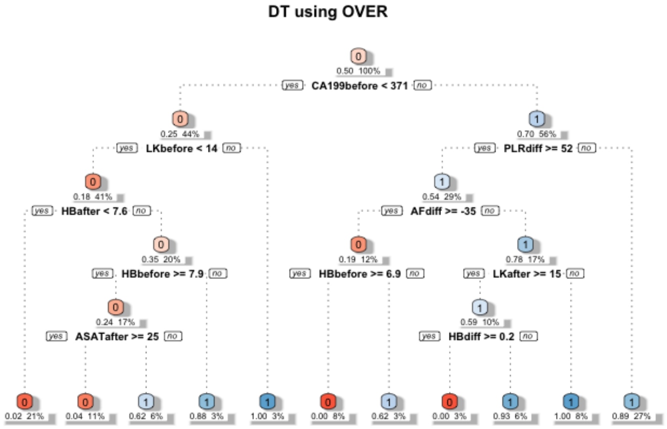


Figure C.7: Decision tree created using the oversampled training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=230, 0=Disease Control (n=115), 1=Progressive disease (n=115).

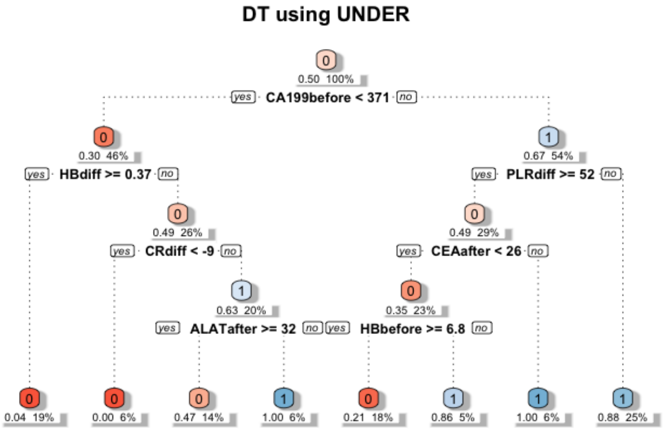


Figure C.8: Decision tree created using the undersampled training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=134, 0=Disease Control (n=67), 1=Progressive disease (n=67).

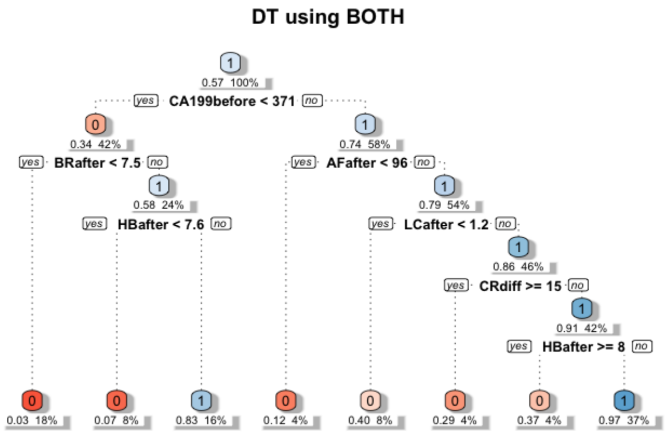


Figure C.9: Decision tree created using both under- and oversampled training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, 0=Disease Control (n=95), 1=Progressive disease (n=87).

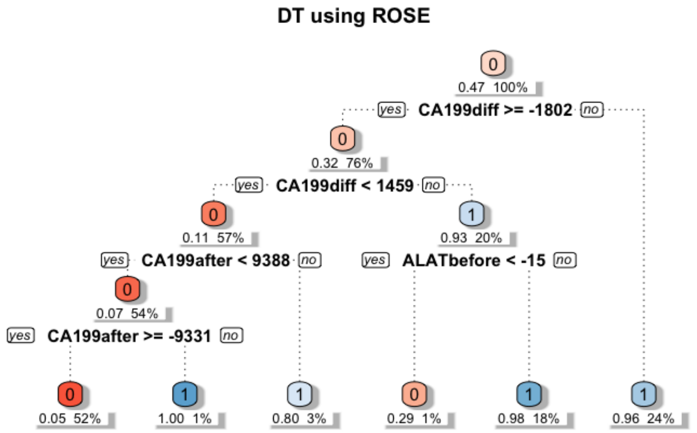


Figure C.10: Decision tree created using the ROSE training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=500, 0=Disease Control (n=265), 1=Progressive disease (n=235).

C.2. Random Forest Plots

C.2.1 Random Forest (full dataset)

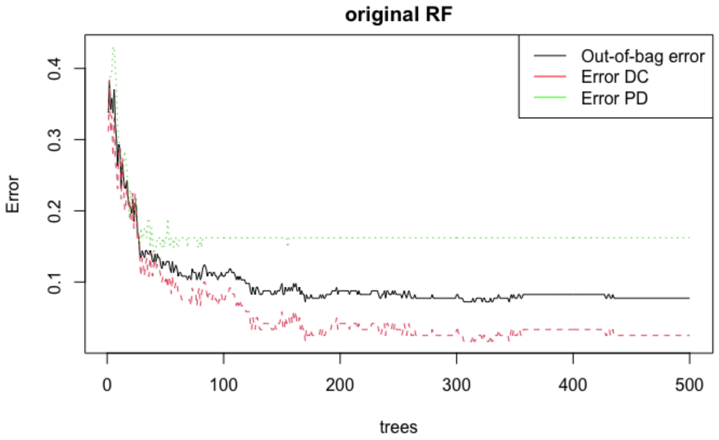


Figure C.11: Out-Of-Bag Error and Training error of the random forest created using the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

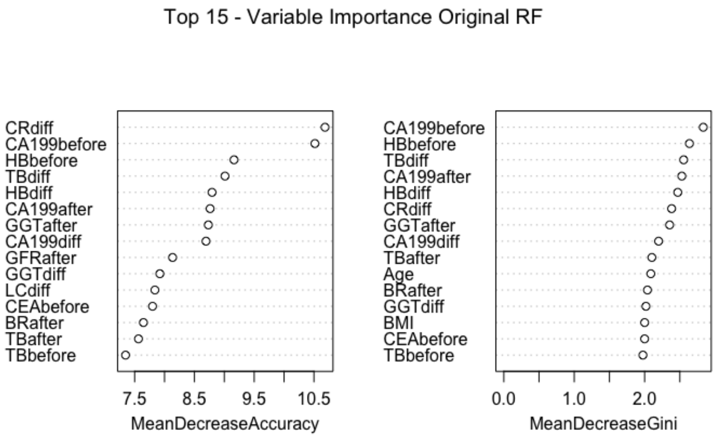


Figure C.12: Top 15 variable importance plot of the random forest created using the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

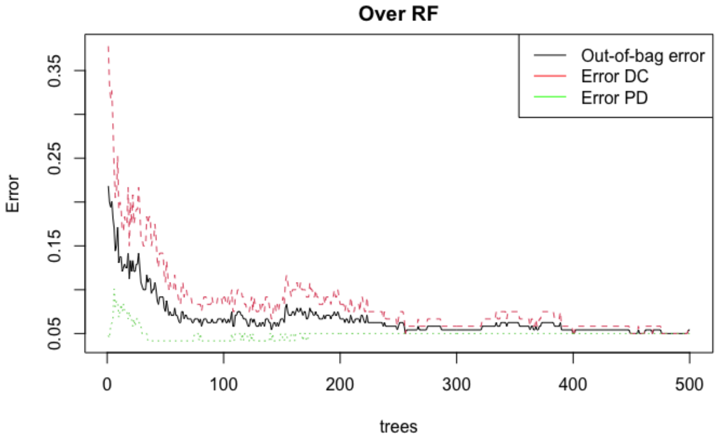


Figure C.13: Out-Of-Bag Error and Training error of the random forest created using the oversampled training dataset, n=240, Disease Control (n=120), Progressive disease (n=120).

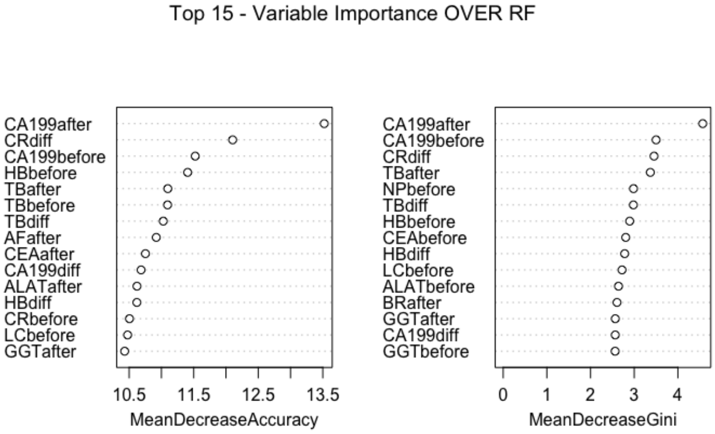


Figure C.14: Top 15 variable importance plot of the random forest created using the oversampled training dataset, n=240, Disease Control (n=120), Progressive disease (n=120).

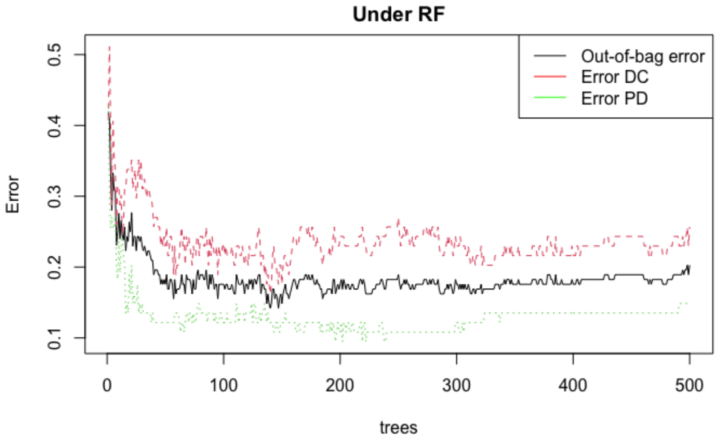


Figure C.15: Out-Of-Bag Error and Training error of the random forest created using the undersampled training dataset, n=148, Disease Control (n=74), Progressive disease (n=74).

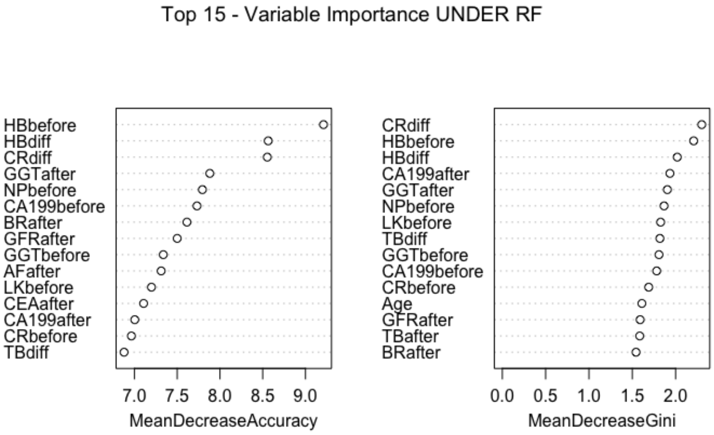


Figure C.16: Top 15 variable importance plot of the random forest created using the undersampled training dataset, n=200, Disease Control (n=90), Progressive disease (110).

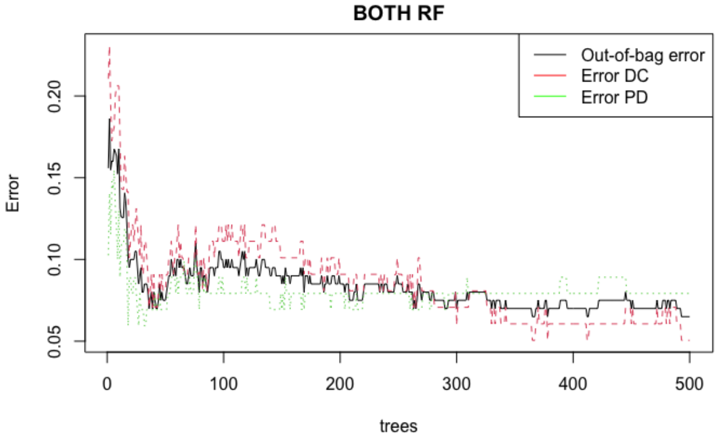


Figure C.17: Out-Of-Bag Error and Training error of the random forest created using both over- and undersampled training dataset, n=200, Disease Control (n=90), Progressive disease (110).

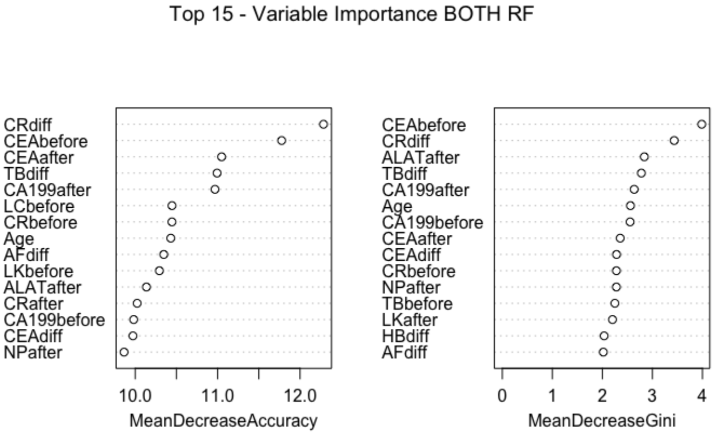


Figure C.18: Top 15 variable importance plot of the random forest created using the original training dataset, n=194, Disease Control (n=120), Progressive disease (n=74).

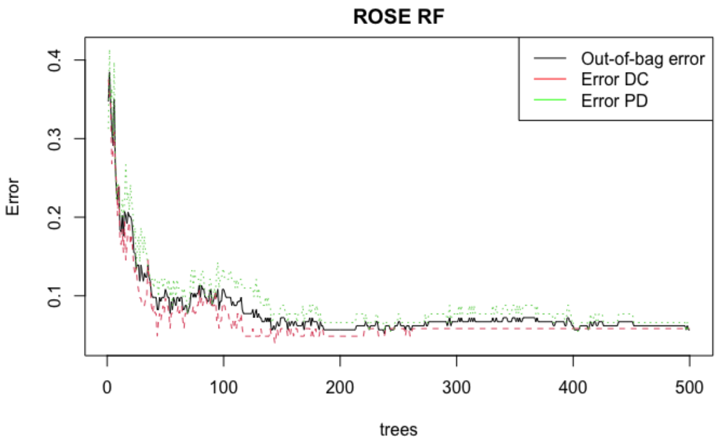


Figure C.19: Out-Of-Bag Error and Training error of the random forest created using the ROSE training dataset, n=204, Disease Control (n=97), Progressive disease (n=97).

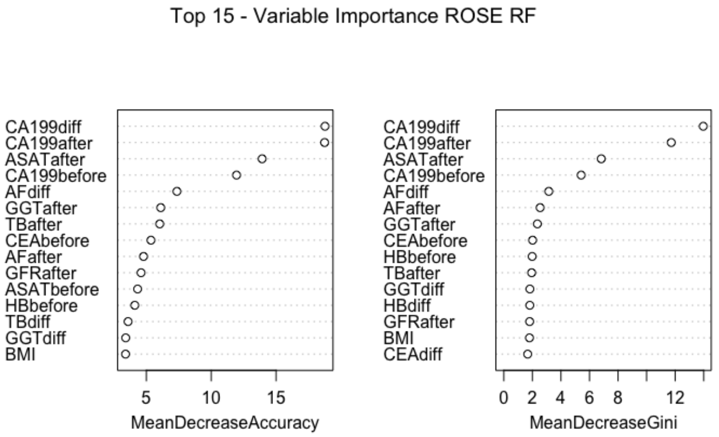
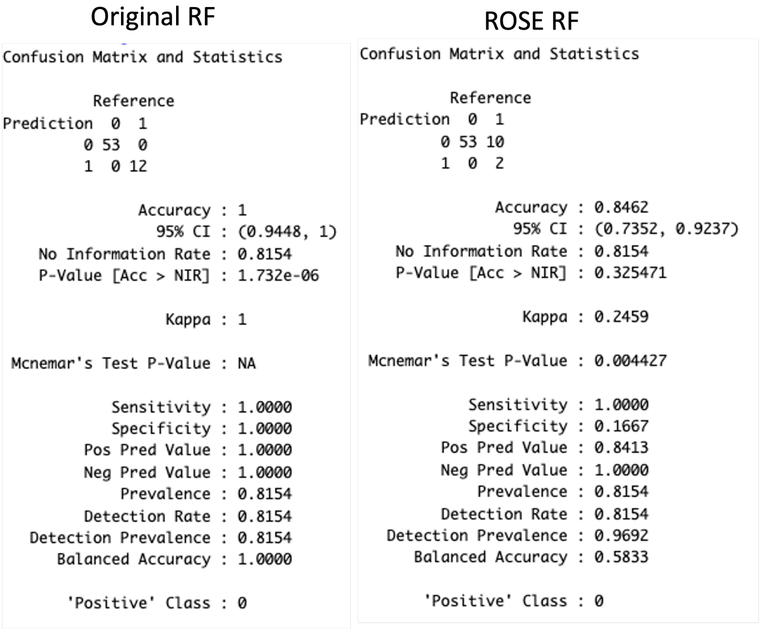


Figure C.20: Top 15 variable importance plot of the random forest created using the ROSE training dataset, n=204, Disease Control (n=97), Progressive disease (n=97).

C.2.2 Random Forest (full dataset) confusion matrices



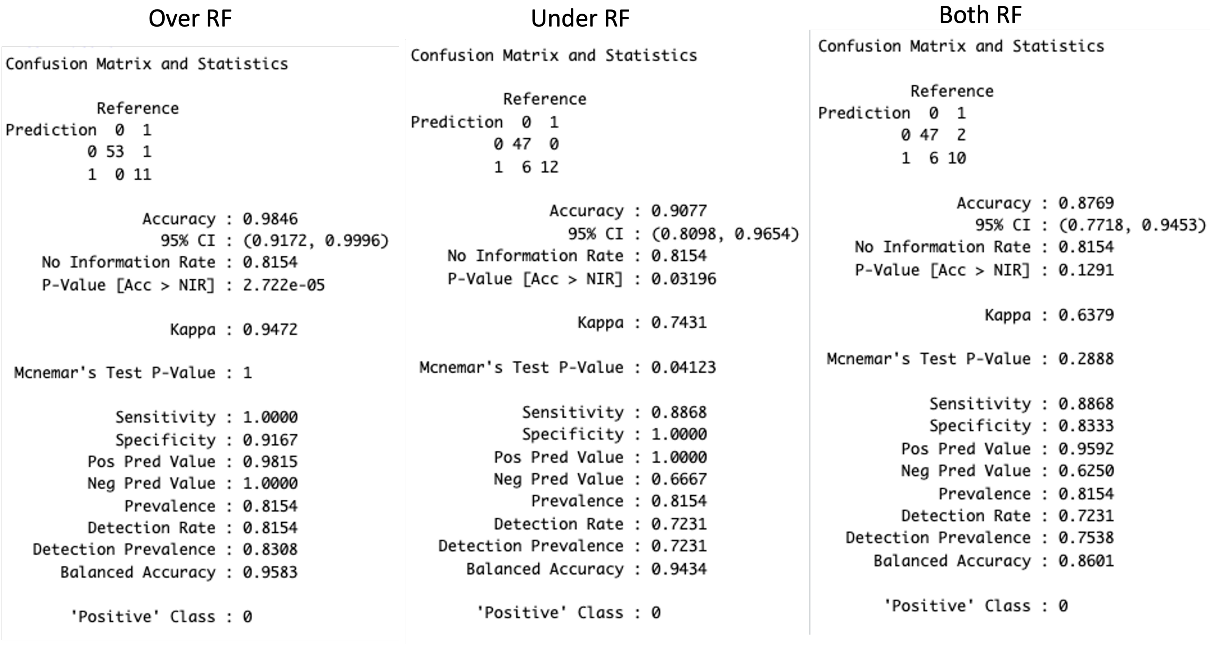
Original RF

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 53  0
         1  0 12

               Accuracy : 1
                 95% CI : (0.9448, 1)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : 1.732e-06

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.8154
         Detection Rate : 0.8154
   Detection Prevalence : 0.8154
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

ROSE RF

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 53 10
         1  0  2

               Accuracy : 0.8462
                 95% CI : (0.7352, 0.9237)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : 0.325471

                  Kappa : 0.2459

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 1.0000
            Specificity : 0.1667
         Pos Pred Value : 0.8413
         Neg Pred Value : 1.0000
             Prevalence : 0.8154
         Detection Rate : 0.8154
   Detection Prevalence : 0.9692
      Balanced Accuracy : 0.5833

       'Positive' Class : 0
```

Figure C.21: Confusion matrix and evaluation metrics of the random forest created using the original training dataset (n=194, Disease Control (n=120), Progressive disease (n=74)) and ROSE training dataset (n=500, Disease Control (n=265), Progressive disease (n=235)).



Over RF

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 53  1
         1  0 11

               Accuracy : 0.9846
                 95% CI : (0.9172, 0.9996)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : 2.722e-05

                  Kappa : 0.9472

 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9167
         Pos Pred Value : 0.9815
         Neg Pred Value : 1.0000
             Prevalence : 0.8154
         Detection Rate : 0.8154
   Detection Prevalence : 0.8308
      Balanced Accuracy : 0.9583

       'Positive' Class : 0
```

Under RF

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 47  0
         1  6 12

               Accuracy : 0.9077
                 95% CI : (0.8098, 0.9654)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : 0.03196

                  Kappa : 0.7431

 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 0.8868
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.6667
             Prevalence : 0.8154
         Detection Rate : 0.7231
   Detection Prevalence : 0.7231
      Balanced Accuracy : 0.9434

       'Positive' Class : 0
```

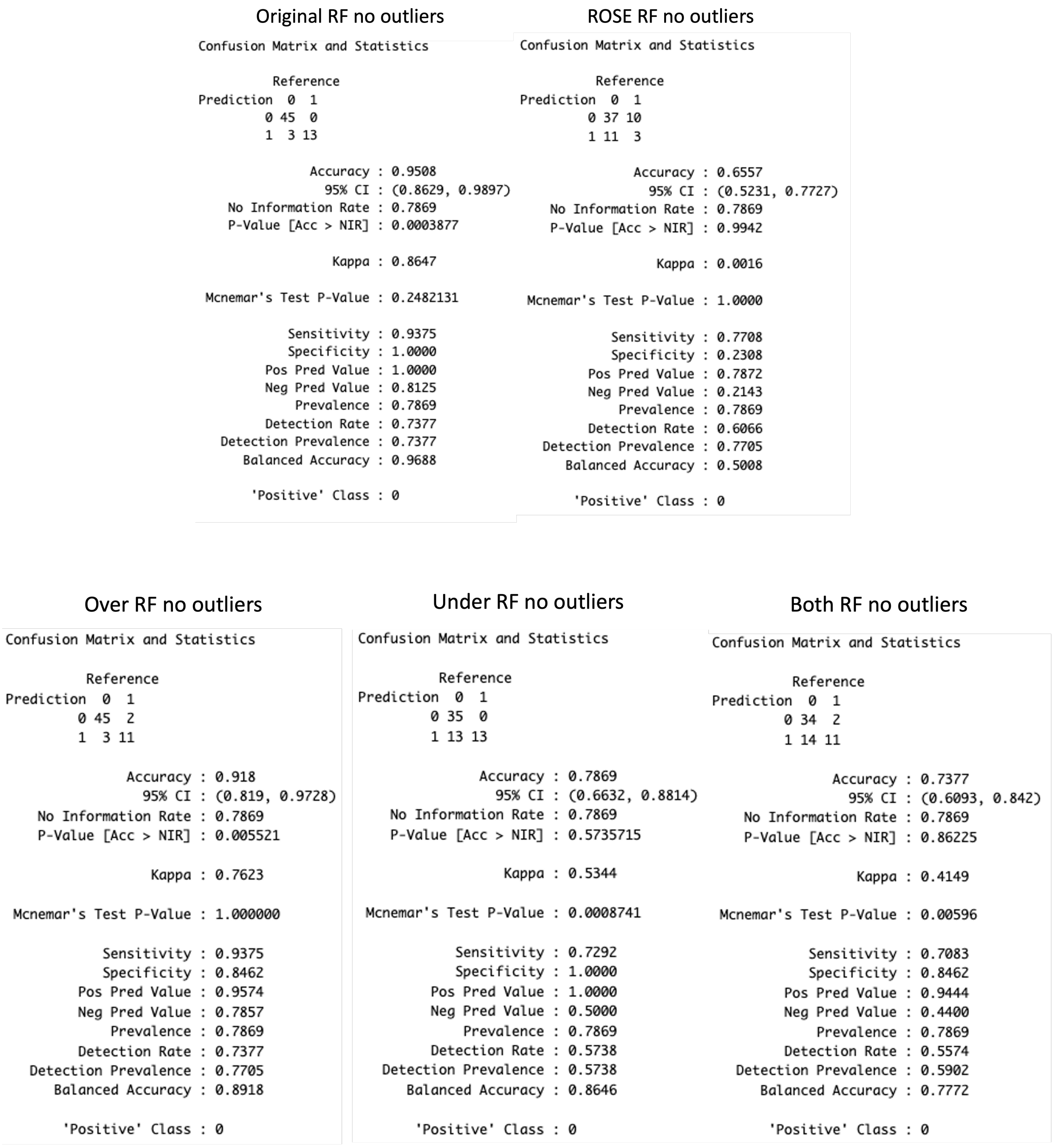
Both RF

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 47  2
         1  6 10

               Accuracy : 0.8769
                 95% CI : (0.7718, 0.9453)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : 0.1291

                  Kappa : 0.6379

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.8868
            Specificity : 0.8333
         Pos Pred Value : 0.9592
         Neg Pred Value : 0.6250
             Prevalence : 0.8154
         Detection Rate : 0.7231
   Detection Prevalence : 0.7538
      Balanced Accuracy : 0.8601

       'Positive' Class : 0
```

Figure C.22: Confusion matrix and evaluation metrics of the random forest created using the oversampled training dataset (n=240, Disease Control (n=120), Progressive disease (n=120)) undersampled training dataset (n=148, Disease Control (n=74), Progressive disease (n=74)) and both over- and undersampled training dataset (n=194, Disease Control (n=100), Progressive disease (n=94)).

Note the 'Positive' Class: 0 refers to the fact that the class labeled as '0' (DC) is referred to as the positive class in the model. It is not a numeric value.

C.2.3 Random Forest (no outliers) confusion matrices



Original RF no outliers

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 45  0
         1  3 13

               Accuracy : 0.9508
                 95% CI : (0.8629, 0.9897)
    No Information Rate : 0.7869
    P-Value [Acc > NIR] : 0.0003877

                  Kappa : 0.8647

 Mcnemar's Test P-Value : 0.2482131

            Sensitivity : 0.9375
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8125
             Prevalence : 0.7869
         Detection Rate : 0.7377
   Detection Prevalence : 0.7377
      Balanced Accuracy : 0.9688

       'Positive' Class : 0
```

ROSE RF no outliers

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 37 10
         1 11  3

               Accuracy : 0.6557
                 95% CI : (0.5231, 0.7727)
    No Information Rate : 0.7869
    P-Value [Acc > NIR] : 0.9942

                  Kappa : 0.0016

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.7708
            Specificity : 0.2308
         Pos Pred Value : 0.7872
         Neg Pred Value : 0.2143
             Prevalence : 0.7869
         Detection Rate : 0.6066
   Detection Prevalence : 0.7705
      Balanced Accuracy : 0.5008

       'Positive' Class : 0
```

Over RF no outliers

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 45  2
         1  3 11

               Accuracy : 0.918
                 95% CI : (0.819, 0.9728)
    No Information Rate : 0.7869
    P-Value [Acc > NIR] : 0.005521

                  Kappa : 0.7623

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.9375
            Specificity : 0.8462
         Pos Pred Value : 0.9574
         Neg Pred Value : 0.7857
             Prevalence : 0.7869
         Detection Rate : 0.7377
   Detection Prevalence : 0.7705
      Balanced Accuracy : 0.8918

       'Positive' Class : 0
```

Under RF no outliers

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 35  0
         1 13 13

               Accuracy : 0.7869
                 95% CI : (0.6632, 0.8814)
    No Information Rate : 0.7869
    P-Value [Acc > NIR] : 0.5735715

                  Kappa : 0.5344

 Mcnemar's Test P-Value : 0.0008741

            Sensitivity : 0.7292
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.5000
             Prevalence : 0.7869
         Detection Rate : 0.5738
   Detection Prevalence : 0.5738
      Balanced Accuracy : 0.8646

       'Positive' Class : 0
```

Both RF no outliers

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 34  2
         1 14 11

               Accuracy : 0.7377
                 95% CI : (0.6093, 0.842)
    No Information Rate : 0.7869
    P-Value [Acc > NIR] : 0.86225

                  Kappa : 0.4149

 Mcnemar's Test P-Value : 0.00596

            Sensitivity : 0.7083
            Specificity : 0.8462
         Pos Pred Value : 0.9444
         Neg Pred Value : 0.4400
             Prevalence : 0.7869
         Detection Rate : 0.5574
   Detection Prevalence : 0.5902
      Balanced Accuracy : 0.7772

       'Positive' Class : 0
```

Figure C.23: Confusion matrix and evaluation metrics of the random forest created using the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values (n=182, Disease Control (n=115), Progressive disease (n=67)) and ROSE training dataset (n=500, Disease Control (n=265), Progressive disease (n=235)). Confusion matrix and evaluation metrics of the random forest created using the oversampled training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values (n=230, Disease Control (n=115), Progressive disease (n=115)) undersampled training dataset (n=134, Disease Control (n=67), Progressive disease (n=67)) and both over- and undersampled training dataset (n=182, Disease Control (n=95), Progressive disease (n=87)).

C.2.4 Random forest (no outliers) top 15 variables

1

Variable	Count
CA199before	98
CA199diff	2

2

Variable	Count
CA199after	42
CA199diff	33
HBdiff	10
TBdiff	8
GGTbefore	3
GGTafter	2
AFafter	1
CA199before	1

3

Variable	Count
CA199diff	32
CA199after	29
TBdiff	13
HBdiff	11
GGTbefore	9
BRafter	3
GGTafter	2
AFafter	1

4

Variable	Count
TBdiff	19
HBdiff	19
CA199diff	17
CA199after	16
GGTbefore	15
AFafter	5
BRafter	4
GGTafter	2
HBbefore	1
Kbefore	1
CA199before	1

5

Variable	Count
HBdiff	21
TBdiff	21
GGTbefore	18
CA199diff	8
BRafter	8
GGTafter	8
CA199after	7
AFafter	4
Kbefore	2
PLRdiff	2
CRdiff	1

6

Variable	Count
TBdiff	17
GGTbefore	15
GGTafter	14
HBdiff	14
BRafter	12
AFafter	10
Kbefore	6
CA199diff	4
HBbefore	2
CA199after	2
NPbefore	2
NLRafter	1
PLRdiff	1

7

Variable	Count
BRafter	14
HBdiff	14
AFafter	13
GGTbefore	12
TBdiff	11
Kbefore	9
GGTafter	5
PLRdiff	4
NLRafter	4
NPbefore	4
HBbefore	3
BMI	2
CRdiff	2
CA199after	2
CA199diff	1

8

Variable	Count
GGTbefore	17
AFafter	15
BRafter	14
GGTafter	11
Kbefore	8
HBdiff	7
PLRdiff	6
HBbefore	5
TBdiff	4
CRdiff	2
BMI	2
NPbefore	2
NLRafter	2
CA199after	1
CA199diff	1
Age	1
SIIdiff	1
AFdiff	1

9

Variable	Count
GGTafter	18
Kbefore	12
AFafter	12
CRdiff	10
NLRafter	9
HBbefore	9
NPbefore	9
PLRdiff	7
BRafter	5
BMI	2
Age	1
SIIbefore	1
SIIdiff	1
TBdiff	1
HBdiff	1
AFdiff	1
CA199diff	1

10

Variable	Count
HBbefore	13
GGTafter	13
Kbefore	10
BRafter	10
NPbefore	10
PLRdiff	8
CRdiff	7
NLRafter	5
AFafter	5
Age	5
BMI	4
GGTbefore	4
SIIdiff	1
AFdiff	1
HBdiff	1
CA199after	1
ALATbefore	1
GGTdiff	1

11

Variable	Count
Kbefore	15
AFafter	12
NLRafter	10
HBbefore	10
PLRdiff	9
NPbefore	7
BMI	7
GGTafter	6
GGTbefore	5
TBdiff	4
BRafter	4
CRdiff	3
TBbefore	2
AFbefore	2
Age	1
CA199diff	1
HBdiff	1
AFdiff	1

12

Variable	Count
BRafter	13
PLRdiff	11
NPbefore	10
HBbefore	9
BMI	9
NLRafter	8
Kbefore	7
CRdiff	7
AFafter	5
GGTafter	4
AFdiff	3
Age	2
AFbefore	2
CEAafter	2
SIIbefore	2
ALATbefore	1
HBdiff	1
TBdiff	1
SIIdiff	1
TBbefore	1
GGTbefore	1

13

Variable	Count
NLRafter	18
PLRdiff	12
NPbefore	10
HBbefore	8
GGTafter	7
BMI	7
Kbefore	7
CRdiff	6
TBbefore	5
Age	4
AFbefore	3
AFafter	3
SIIdiff	3
GFRafter	2
BRafter	2
PLRafter	1
SIIbefore	1
ASATbefore	1

14

Variable	Count
PLRdiff	13
CRdiff	12
NPbefore	11
HBbefore	11
BMI	10
NLRafter	7
Kbefore	6
Age	5
AFafter	4
SIIbefore	4
CEAafter	4
AFdiff	3
BRafter	2
GGTafter	2
GGTdiff	1
AFbefore	1
TBbefore	1
TBafter	1
SIIafter	1
CEAbefore	1

15

Variable	Count
CRdiff	13
NPbefore	12
AFdiff	8
Kbefore	8
PLRdiff	7
BMI	7
HBbefore	7
NLRafter	6
CEAbefore	5
Age	4
TBbefore	3
BRafter	3
TBafter	2
CEAafter	2
SIIbefore	2
AFbefore	2
AFafter	1
NaBefore	1
Kdiff	1
HBafter	1
TBdiff	1
GGTafter	1
GGTbefore	1
GFRafter	1
SIIdiff	1

Table C.1: Top 15 most important variables and their respective count on that ‘rank’ measured using the mean decrease in Gini in each of the optimal random forest (ntree = 300) tuned to their respective optimal mtry value across 100 runs trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$) and BR ($> 50\mu mol/L$) values, n=182, Disease Control (n=115), Progressive disease (n=67).

D | PDP and ALE

D.1. Quantiles and Finite Difference in 1D

D.1.1 Quantiles:

Quantiles are a statistical concept used to divide a dataset or probability distribution into equally sized groups, providing insights into the distribution of values and facilitating the assessment of the relative position of specific values within the dataset. To calculate quantiles, the dataset or distribution is first sorted in ascending order, arranging the values from smallest to largest. A quantile represents a threshold value below which a certain percentage of the data falls. For example, the median is a common quantile that divides the data into two equal parts, with 50% of the values falling above it and 50% of the values below it. In the case of dividing each data point into its own group, we assign a quantile to each individual observation. The first data point corresponds to the 0^{th} quantile, indicating that no values in the dataset are below it. The second data point represents the $\frac{1}{n}^{th}$ quantile, where n is the total number of observations. Similarly, the third observation corresponds to the $\frac{2}{n}^{th}$ quantile, and so on. Each observation is thus assigned its specific quantile value based on its position in the sorted dataset.

D.1.2 Finite difference in 1D

In the case of the first-order ALE method using the forward difference formula, finite differences is used to approximate the partial derivative. Finite differences refer to the differences between the values of a function at specific points. In the context of one-dimensional functions, finite differences provide an approximation of the derivative of a function at different grid points. Consider a function $f(z)$ defined on a set of points $z_0, z_1, z_2, \ldots, z_n$ in a one-dimensional space, where $h = z_i - z_{i-1}$ represents the grid spacing. The finite difference at a particular point z_i is computed by subtracting the value of the function at that point from the value of the function at a neighboring point. Specifically, the forward difference formula is used in this case. The general formula for the forward difference is as follows:

$$f'(z) \approx \frac{f(z+h) - f(x)}{h} = \frac{f(z_{i+1} - f(z_i)}{h}. \tag{D.1}$$

This is exactly what is used in the the ALE method as can be seen in Equation (4.52).

D.2. PDP ALE plots (top 10 important variables)

Based on the top 10 variables identified in section 5 the PDP and ALE plots that have not been presented in section 5.2 are presented in this Appendix.

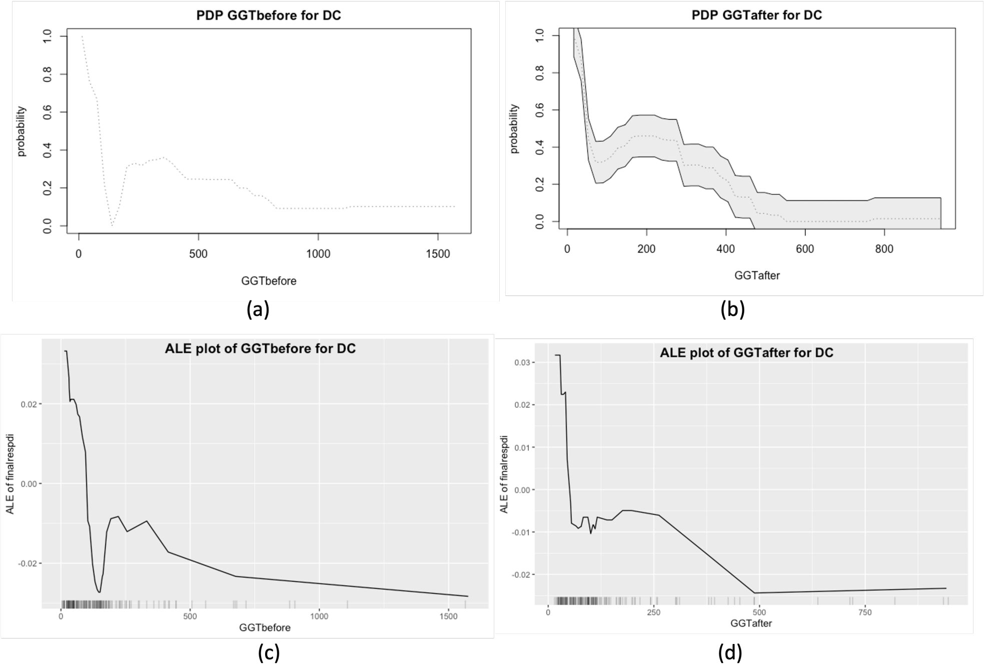
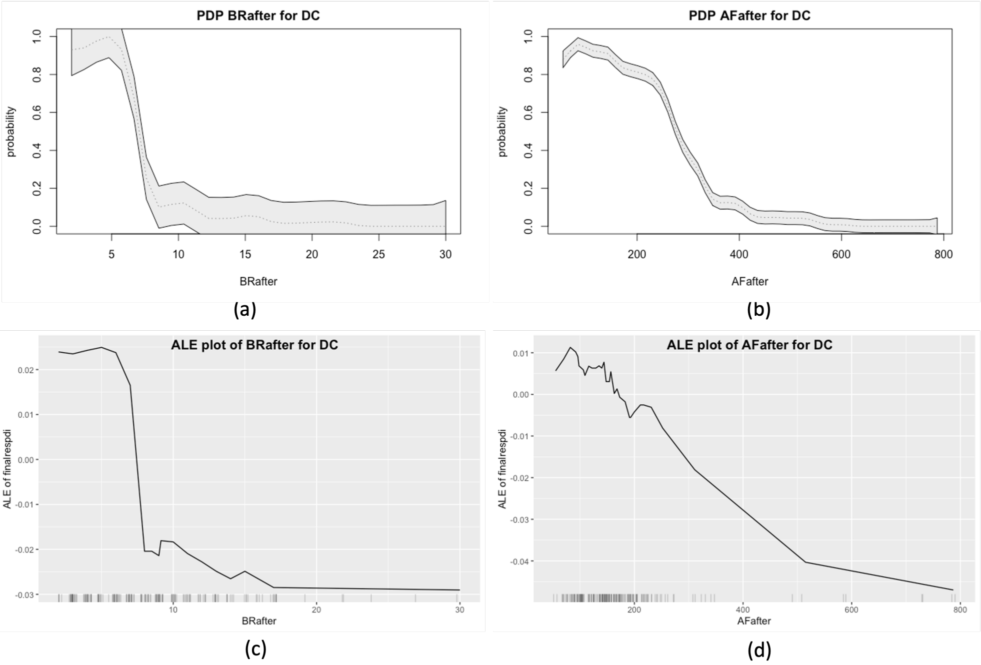


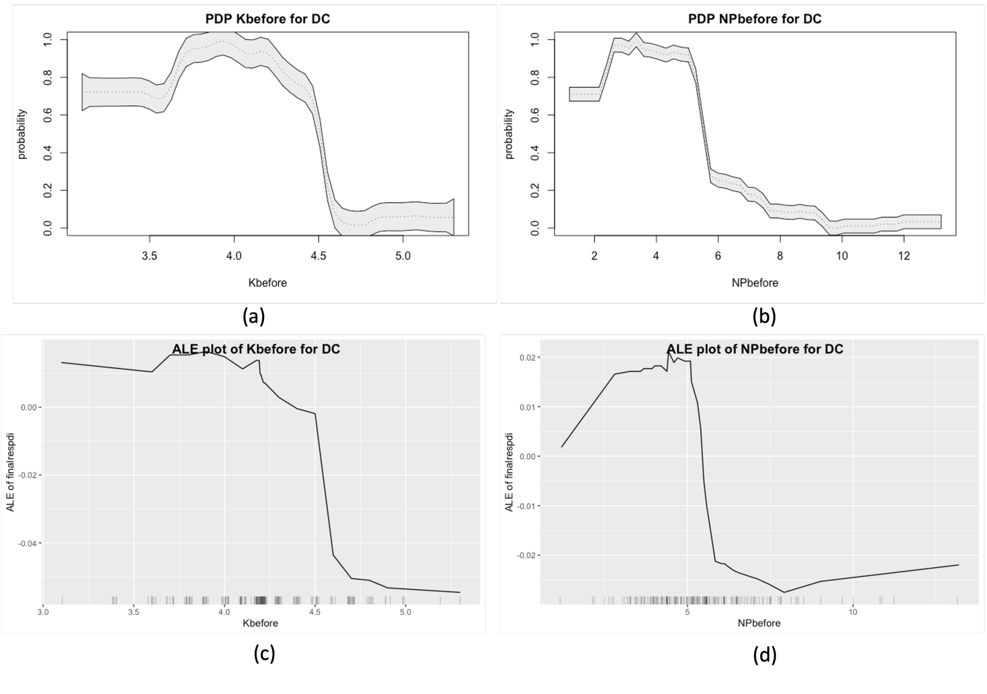
Figure D.1: Partial Dependence Plot and Accumulated Local Effect Plots of the γ-Glutamyl Transferase before and after the first chemotherapy cycle (U/L) values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of GGT before (U/L)for the Disease Control as response group (b) PDP plot with 95% confidence interval of GGT after (U/L) for the Disease Control as response group (c) ALE plot of GGT before (U/L) for the Disease Control as response group (grid size = 30) (d) ALE plot of GGT after (U/L)) for the Disease Control as response group (grid size = 30).



(a) (b)

(c) (d)

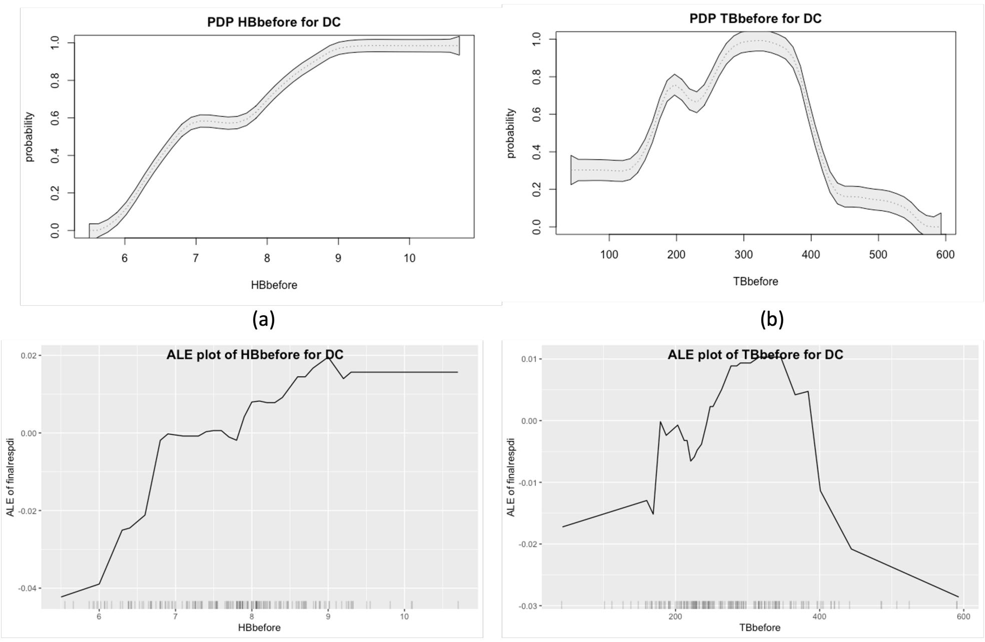
Figure D.2: Partial Dependence Plot and Accumulated Local Effect Plots of the Bilirubin and Alkaline Phosphatase after the first chemotherapy cycle values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of BR after ($\mu mol/L$) for the Disease Control as response group (b) PDP plot with 95% confidence interval of AF after (U/L) for the Disease Control as response group (c) ALE plot of BR after ($\mu mol/L$) for the Disease Control as response group (grid size = 30) (d) ALE plot of AF after (U/L) for the Disease Control as response group (grid size = 30).



(a) (b)

(c) (d)

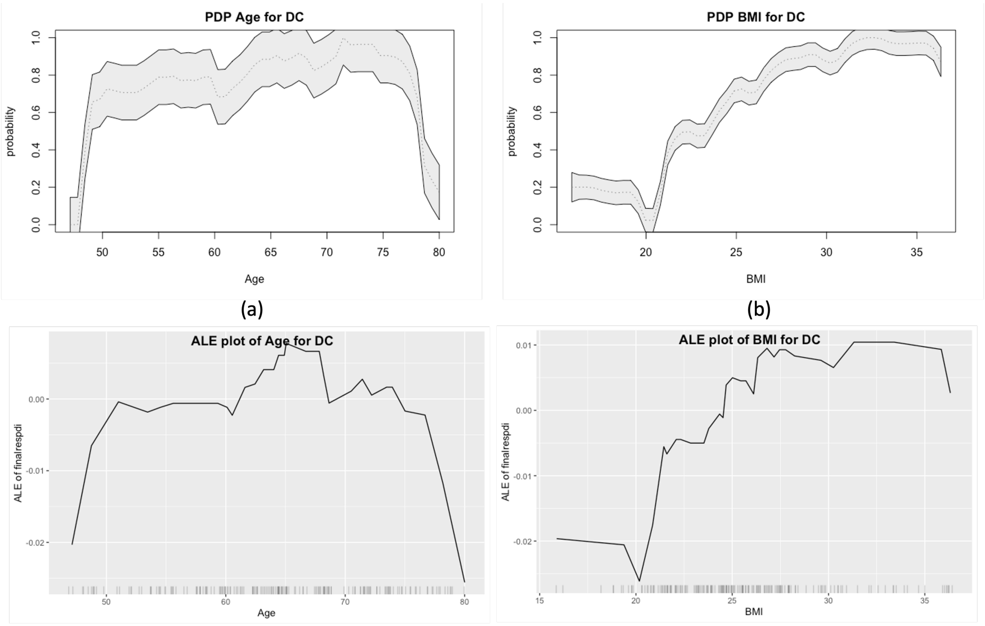
Figure D.3: Partial Dependence Plot and Accumulated Local Effect Plots of the potassium and neutrophils before chemotherapy values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of K before ($mmol/L$) for the Disease Control as response group (b) PDP plot with 95% confidence interval of NP before ($10^9/L$) for the Disease Control as response group (c) ALE plot of K before ($mmol/L$) for the Disease Control as response group (grid size = 30) (d) ALE plot of NP before ($10^9/L$) for the Disease Control as response group (grid size = 30).



(a) (b)

(c) (d)

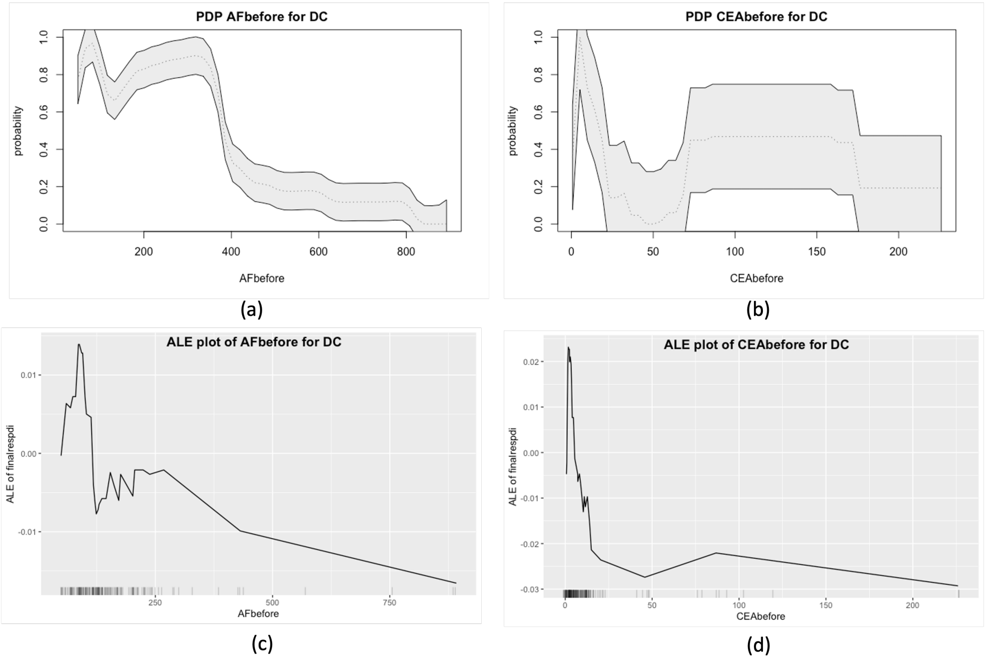
Figure D.4: Partial Dependence Plot and Accumulated Local Effect Plots of the Hemoglobin and Thrombocyte before chemotherapy values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of HB before ($mmol/L$) for the Disease Control as response group (b) PDP plot with 95% confidence interval of TB before ($10^9/L$) for the Disease Control as response group (c) ALE plot of HB before ($mmol/L$) for the Disease Control as response group (grid size = 30) (d) ALE plot of TB before ($10^9/L$) for the Disease Control as response group (grid size = 30).



(a) (b)

(c) (d)

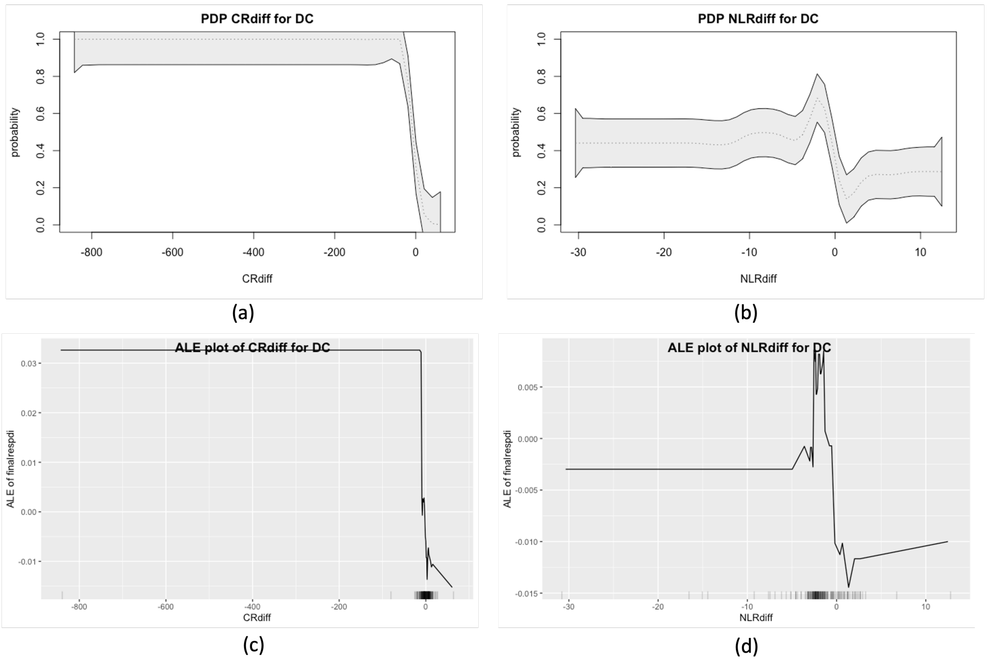
Figure D.5: Partial Dependence Plot and Accumulated Local Effect Plots of the Age and BMI values based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of Age (years) for the Disease Control as response group (b) PDP plot with 95% confidence interval of BMI for the Disease Control as response group (c) ALE plot of Age (years) for the Disease Control as response group (grid size = 30) (d) ALE plot of BMI for the Disease Control as response group (grid size = 30).



(a) (b)

(c) (d)

Figure D.6: Partial Dependence Plot and Accumulated Local Effect Plots of the Alkaline Phosphatase and CEA values before chemotherapy based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of AF before (U/L) for the Disease Control as response group (b) PDP plot with 95% confidence interval of CEA before ($\mu g/L$) for the Disease Control as response group (c) ALE plot of AF before (U/L) for the Disease Control as response group (grid size = 30) (d) ALE plot of CEA before ($\mu g/L$) for the Disease Control as response group (grid size = 30).



(a) (b)

(c) (d)

Figure D.7: Partial Dependence Plot and Accumulated Local Effect Plots of the Creatinin and Neutrophil-to-Lymphocyte Ratio differences based on the random forest trained on the original training dataset after removal of outliers based on GFR ($< 30mL/min$), BR ($> 50\mu mol/L$) (n=182, Disease Control (n=115), Progressive disease (n=67)): (a) PDP plot with 95% confidence interval of CR difference ($\mu mol/L$) for the Disease Control as response group (b) PDP plot with 95% confidence interval of NLR difference for the Disease Control as response group (c) ALE plot of CR difference ($\mu mol/L$) for the Disease Control as response group (grid size = 30) (d) ALE plot of NLR difference for the Disease Control as response group (grid size = 30).

E | Additional Background Information

E.1. Additional background information

E.1.1 Diagnostic, predictive and prognostic

The terms “diagnostic," “prognostic," and “predictive" are frequently used in medical literature, particularly in the context of biomarkers. When applied to biomarkers, these terms can be defined as follows:

- **Diagnostic:** a diagnostic biomarker determines the presence and type of cancer
- **Prognostic:** a prognostic biomarker gives information about which outcomes are likely or unlikely with or without (standard) treatment
- **Predictive:** a predictive biomarker gives information about the treatment benefit and helps to identify which treatment the patient is most likely to respond to or benefit from.

To further clarify, a prognostic biomarker provides information about the overall cancer prognosis for patients, regardless of the therapy given. On the other hand, a predictive biomarker specifically informs about the impact of a therapeutic intervention. It can serve as a target for therapy. This thesis primarily focuses on predictive biomarkers. [109]

E.1.2 Clinical trial (phases)

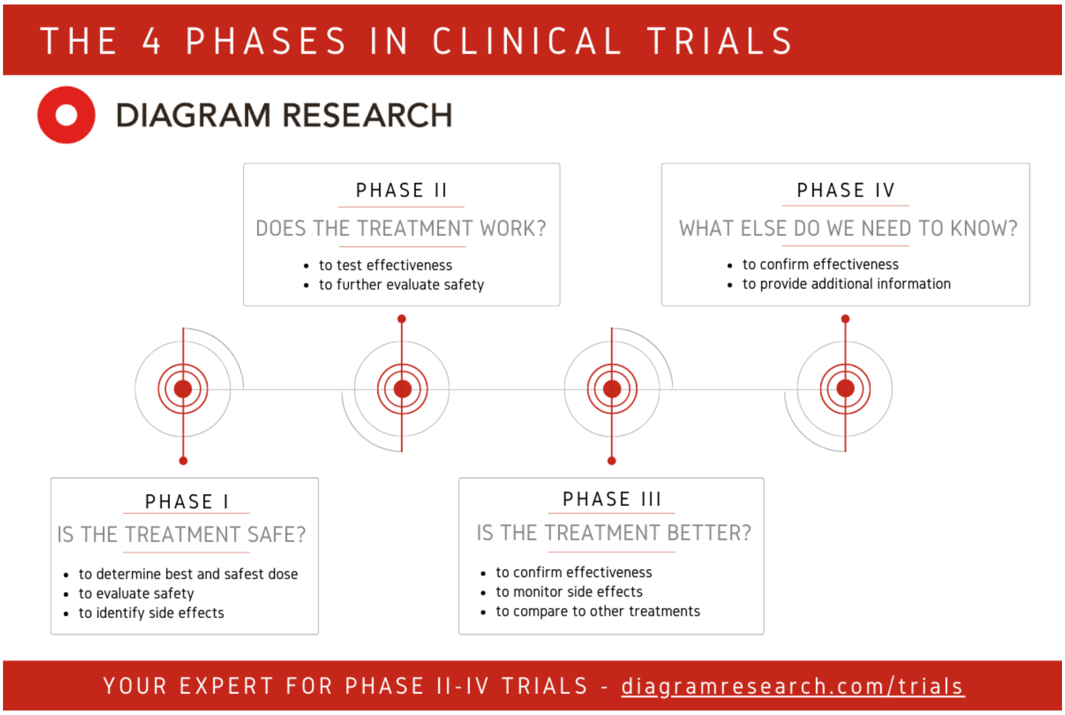
Clinical trials are typically categorized into different phases, which represent specific stages in the research process. Each phase of a clinical trial is designed with distinct objectives and criteria that need to be met before a potential new treatment or medical device can progress to the next phase.

- Phase I trials are the initial stage of testing in humans and primarily focus on evaluating the safety and dosage levels of the intervention. These trials involve a small number of participants and aim to determine how the treatment or device is metabolized and tolerated by the body.
- Phase II trials involve a larger group of participants and are conducted to further assess the treatment’s safety and effectiveness. These trials may also investigate optimal dosages and potential side effects. The primary goal of phase II trials is to gather preliminary data on the treatment’s efficacy in treating the targeted condition or disease.
- Phase III trials involve a larger population and are conducted to confirm the treatment’s effectiveness, monitor side effects, and compare it to standard treatments or a placebo. These trials provide crucial evidence to support the approval of a new treatment or device and are often randomized and controlled.
- Phase IV trials, also known as post-marketing surveillance trials, occur after the treatment or device has received regulatory approval. These trials aim to gather additional information on the treatment’s long-term risks, benefits, and optimal use in larger and more diverse populations.

By following this phased approach, clinical trials ensure that new treatments and medical devices undergo rigorous evaluation to determine their safety, efficacy, and overall impact on human health before they are made widely available [110].

E.1.3 ECOG Performance Status Scale

In order to ensure consistent conduct of clinical trials for cancer treatment across multiple healthcare institutions, including hospitals, cancer centers, and clinics, it is necessary to employ standardized criteria for assessing the impact of the disease on a patient’s daily functioning. This evaluation, commonly referred to as a patient’s performance status, is crucial for physicians and researchers. The ECOG Performance Status Scale represents one such standardized measurement. This scale provides a comprehensive assessment of a patient’s functional capacity, encompassing their ability to perform self-care, engage in daily activities, and carry out physical tasks such as walking or working.



THE 4 PHASES IN CLINICAL TRIALS

DIAGRAM RESEARCH

PHASE II

DOES THE TREATMENT WORK?

- to test effectiveness
- to further evaluate safety

PHASE IV

WHAT ELSE DO WE NEED TO KNOW?

- to confirm effectiveness
- to provide additional information

PHASE I

IS THE TREATMENT SAFE?

- to determine best and safest dose
- to evaluate safety
- to identify side effects

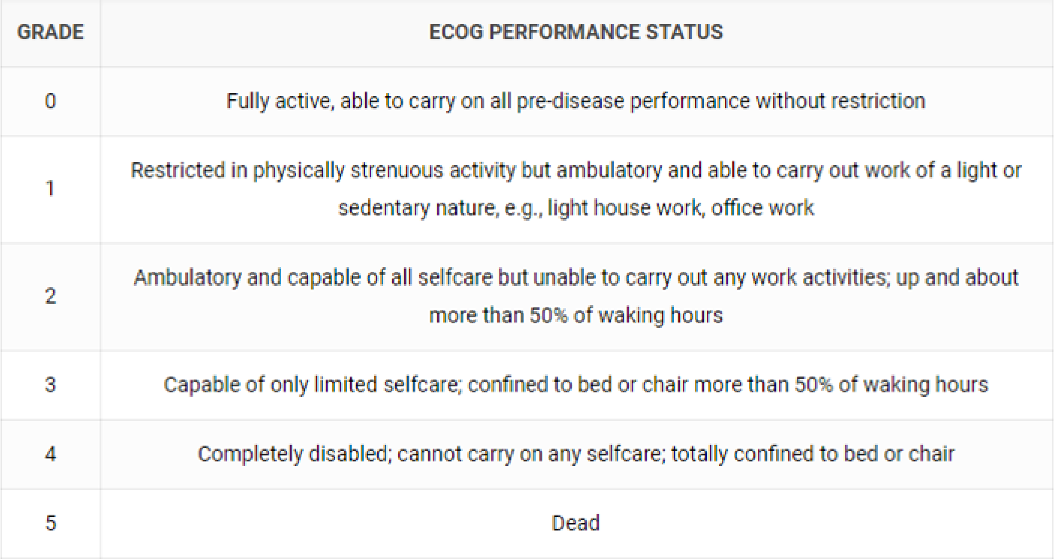
PHASE III

IS THE TREATMENT BETTER?

- to confirm effectiveness
- to monitor side effects
- to compare to other treatments

YOUR EXPERT FOR PHASE II-IV TRIALS - diagramresearch.com/trials

Figure E.1: Four phases of clinical trials [110].



GRADE	ECOG PERFORMANCE STATUS
0	Fully active, able to carry on all pre-disease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work
2	Ambulatory and capable of all selfcare but unable to carry out any work activities; up and about more than 50% of waking hours
3	Capable of only limited selfcare; confined to bed or chair more than 50% of waking hours
4	Completely disabled; cannot carry on any selfcare; totally confined to bed or chair
5	Dead

Figure E.2: ECOG Performance Status Scale [15].

E.1.4 White Blood Cells in detail

In section 2.1 a brief overview of the different types of leukocytes (white blood cells) is given. In this subsection a more in depth explanation of each cell type is provided.

E.1.4.1 Granulocytes

Granulocytes, the most prevalent type of white blood cells, play a crucial role in the immune system's response to infections, allergic reactions, and asthma episodes. These cells are derived from stem cells located in the bone marrow and exhibit a relatively short lifespan of only a few days. Granulocytes are characterized by the presence of small granules within their cytoplasm, which contain enzymes. When the body encounters an infection or inflammation, granulocytes swiftly migrate to the affected area and release their granules to combat the invading pathogens. The three distinct types of granulocytes are neutrophils, eosinophils, and basophils [111].

Function of Granulocytes
Granulocytes collaborate synergistically to eliminate infections and allergens from the body. Each type of granulocyte possesses a unique combination of chemicals and enzymes within their granules, resulting in distinct functional roles:

- **Neutrophils:** are the most abundant type of granulocytes that play a central role in combating bacterial infections. They have the ability to phagocytize and neutralize bacteria. Neutrophils make up 40%-60% of the total granulocyte population and serve as a crucial component of the immune system. They are responsible for defending against infections and aiding in the healing process. The absolute neutrophil count is used as an indicator of neutrophil levels in the body. In the immune system, neutrophils act as the first line of defense against invading pathogens, employing various strategies to capture and destroy them. They also facilitate tissue repair [112].
- **Eosinophils**: actively participate in various immune responses, particularly in the context of allergic reactions. Additionally, eosinophils are involved in combating parasitic infections.
- **Basophils**: primarily contribute to the modulation of allergic reactions. They release histamine, which facilitates the expulsion of allergens from the body, and heparin, a blood-thinning agent that inhibits clot formation.

E.1.4.2 Lymphocytes

Lymphocytes are essential components of the immune system and play a vital role in defending against diseases and infections. There are two main types of lymphocytes: T-cells and B-cells. T-cells are immune cells involved in eliminating infected cells and regulating the immune response. They differentiate into various subtypes, including cytotoxic T-cells that destroy abnormal cells, helper T-cells that assist other immune cells, and regulatory T-cells that help control the immune response. B cells play a crucial role in the immune response by producing specific antibodies that target antigens to 'neutralize' foreign invaders. They respond to antigens through a primary immune response, where they differentiate into memory cells and plasma cells, and a secondary immune response, where memory cells quickly multiply and produce the appropriate antibody upon encountering the same antigen again. This memory response is the basis for the effectiveness of vaccinations and the prevention of recurring infections [113].

E.1.4.3 Monocytes

Monocytes are a subset of leukocytes. They circulate in the blood and reside in various tissues, actively seeking out and eliminating germs such as viruses, bacteria, fungi, and protozoa. Additionally, monocytes play a role in coordinating immune responses to injuries and preventing infections. When an invading germ or bacteria enters the body, monocytes can transform into macrophages or dendritic cells. These specialized cells have the ability to either destroy the invader themselves or signal other blood cells to assist in its elimination, thus preventing infection. Dendritic cells capture antigens and release cytokines to coordinate the immune response against infections, primarily in superficial tissues. Contrastingly, macrophages are frontline defenders that engulf and neutralize a wide range of pathogens, contributing to tissue clearance and immune defense [114].

E.1.5 Cytokines

Cytokines are small proteins that play a crucial role in controlling the growth and activity of immune system cells and blood cells. They act as signaling molecules, instructing the immune system to perform its functions. Cytokines have a broad impact on the growth and development of various blood cells and other cells involved in the body's immune and inflammatory responses. They are involved in processes such as tissue repair, cancer development and progression, regulation of cell replication and apoptosis, and modulation of immune reactions and sensitization. One specific type of cytokine is called a chemokine, which is responsible for guiding immune cells toward specific targets. Chemokines encompass various subclasses, including interleukins, interferons, tumor necrosis factors, and growth factors.

Interleukins
Interleukins are a group of cytokines that act as chemical signals between white blood cells. They facilitate communication and coordination among immune cells. One notable interleukin is interleukin-2 (IL-2), which promotes the growth and division of immune system cells. A synthetic version of IL-2 has been approved for the treatment of advanced kidney cancer and metastatic melanoma. IL-2 can be used as a standalone treatment for these cancers or combined with chemotherapy or other cytokines like interferon-alfa. Other interleukins, such as IL-7, IL-12, and IL-21, are still under investigation for their potential use against cancer. They are being studied both as adjuvants, which enhance the effects of other treatments, and as standalone agents.

Interferons
Interferons represent a group of chemical substances that facilitate the body's defense against viral infections and cancers. The three primary types of interferon, denoted by the first three letters of the Greek alphabet, are: Interferon-alfa (IFN-alfa), Interferon-beta (IFN-beta) and Interferon-gamma (IFN-gamma). Among these, IFN-alfa is the sole interferon used in cancer treatment. It enhances the capability of specific immune cells to target and eliminate cancer cells. Additionally, it may directly impede the proliferation of cancer cells, as well as curtail the formation of blood vessels crucial for tumor growth.

E.1.6 Other important terms

Overall Response Rate (ORR)
The overall response rate (ORR) is a metric used to evaluate the efficacy of a treatment in a study or treatment group. It represents the percentage of individuals who exhibit either a partial or complete response to the treatment within a specified timeframe. A partial response refers to a reduction in tumor size or cancer burden, while a complete response indicates the complete disappearance of all cancer manifestations in the body. Assessing the overall response rate is an important approach in clinical trials to gauge the effectiveness of a novel treatment. It serves as a measure of treatment success, indicating the proportion of patients who experience a positive response to the therapy.

Overall Survival Rate (OS)
The overall survival rate refers to the percentage of individuals who remain alive at a specified time point after their cancer diagnosis. It represents the cumulative probability of survival for the entire population affected by a particular cancer type. The overall survival rate includes all individuals with the specific cancer type, regardless of other factors or treatments they may have received. Survival rates can be calculated for different durations of time, but researchers commonly report the 5-year relative survival rate in cancer statistics. This rate provides an estimate of the proportion of individuals who survive at least 5 years following their diagnosis, taking into account the expected survival of individuals in the general population.

Progression-Free Survival (PFS)
Progression-free survival (PFS) is a measure that assesses the length of time a patient lives with a disease, such as cancer, without experiencing disease progression. In the field of oncology, PFS typically applies to situations where there is evidence of a tumor, as determined through laboratory tests, radiological imaging, or clinical evaluation. Alternatively, "disease-free survival" refers to the duration after treatment when patients have no detectable signs of the disease. PFS is specifically defined as the period from randomization until objective tumor progression or death. The precise criteria for determining tumor progression are of utmost importance and are thoroughly outlined in the study protocol. Furthermore, PFS1 evaluates the progression-free survival for first-line therapy, while PFS2 focuses on second-line therapy.

Disease Control Rate (DCR) and Clinical Benefit Rate (CBR)
Disease Control Rate (DCR) and Clinical Benefit Rate (CBR) are metrics used to assess the effectiveness of therapeutic interventions in clinical trials involving patients with advanced or metastatic cancer. These rates represent the percentage of patients who demonstrate a favorable response to the treatment, which includes achieving complete response, partial response, or stable disease. DCR and CBR provide valuable information on the overall disease control and clinical benefits experienced by patients receiving anticancer agents in the context of clinical trials.

Disease-free survival (DFS)
Disease-Free Survival (DFS) is a measure of the length of time between the initiation of treatment and the occurrence of disease relapse. This endpoint is typically used in the context of treatments such as surgery, chemotherapy, or targeted therapy, where patients show no detectable signs of disease following treatment. DFS provides valuable information on the duration of disease remission or absence of disease progression after the completion of treatment. It is commonly used as an important outcome measure in clinical trials to assess the efficacy of interventions in maintaining disease-free status in

patients.

Duration of Response (DoR)
Duration of Response (DoR) is defined as the period from the initial documentation of tumor response, either complete response or partial response, to the subsequent occurrence of disease progression. A complete response refers to the complete disappearance of the cancer, while a partial response indicates a significant reduction in tumor size by at least one-third. Stable disease is characterized by the absence of tumor growth or shrinkage. Relapse or recurrence refers to the reappearance of cancer after a period of remission, while progression signifies the resumption of tumor growth in a previously responsive cancer. The duration of response corresponds to the time interval between the initial response and the occurrence of relapse or progression. It provides insights into the sustainability and durability of the treatment response achieved in patients with cancer.

Palliative Chemotherapy
Palliative chemotherapy in PDAC refers to the use of chemotherapy when curative treatments are not possible due to advanced disease. Its primary goals are symptom alleviation, improved quality of life, and potentially extended overall survival. PDAC is often diagnosed at an advanced stage, making palliative chemotherapy the main treatment approach. Palliative chemotherapy regimens target tumor cells, inhibiting their growth and spread, leading to tumor shrinkage, symptom reduction, pain relief, and improved well-being. While not curative, palliative chemotherapy plays a crucial role in managing PDAC, improving patient outcomes, and enhancing their quality of life.

(Neo)adjuvant treatment
Neoadjuvant therapy involves administering treatment before the primary treatment, typically surgery, to shrink the tumor or eliminate cancer cells. It is used when the tumor is initially inoperable or to improve surgical outcomes. Neoadjuvant therapy can include chemotherapy, radiation therapy, or hormone therapy. The choice of treatment depends on factors such as tumor characteristics, disease stage, and patient's health. Neoadjuvant therapy aims to downsize the tumor for easier surgical removal, especially for complex or challenging tumors. Adjuvant therapy, on the other hand, is given after the primary treatment to eliminate any remaining cancer cells and reduce the risk of recurrence. It can involve additional chemotherapy, radiation therapy, or targeted therapies. The goal of adjuvant therapy is to target residual cancer cells and improve long-term outcomes. The selection and sequencing of these therapies depend on cancer type, stage, and individual patient factors. Together, neoadjuvant and adjuvant therapies form a comprehensive treatment strategy to optimize outcomes and achieve long-term disease control in cancer patients.

(Tumor) Angiogenesis
Angiogenesis is a natural process involved in the formation of new blood vessels from pre-existing ones. It plays a crucial role in various physiological processes, such as wound healing and embryonic development. Tumor angiogenesis is the process by which a tumor develops a network of blood vessels to support its growth and progression. It involves the formation of new blood vessels from existing ones, driven by specific signaling molecules released by tumor cells. This process ensures the supply of nutrients and oxygen to the growing tumor. Dysregulated angiogenesis can contribute to cancer development and spread. Angiogenic factors play a critical role in stimulating the growth of blood vessels and promoting tumor angiogenesis [115].

Apoptosis and Necrosis
Cell death can occur through two main processes: apoptosis and necrosis. Apoptosis is a programmed and regulated form of cell self-destruction that does not elicit an inflammatory response. It is a normal physiological process that serves protective functions in the body. In contrast, necrosis is an accidental and uncontrolled form of cell death caused by external factors or disease. It leads to the release of inflammatory cellular contents and is considered an abnormal cell death process. While apoptosis is a well-studied and beneficial process, necrosis is typically associated with pathological conditions.

Hemoglobin and Glycated Hemoglobin
Hemoglobin is a protein molecule in red blood cells that transports oxygen throughout the body. Glycated hemoglobin (HbA1c) is a form of hemoglobin with glucose attached to it. It reflects average blood glucose levels over a period of time and is used to diagnose and monitor diabetes. HbA1c provides information about long-term glycemic control and helps assess the effectiveness of diabetes management [116] [68].