MSc

2.º
CICLO

FCUP
ANO

Predicting the personalized duration of phlebotomy treatments in patients with Hereditary Hemochromatosis: a model based approach

Miguel Costa Carvalho Faria

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Predicting the personalized duration of phlebotomy treatments in patients with Hereditary Hemochromatosis: a model based approach

Miguel Costa Carvalho Faria

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional
2019

Predicting the personalized duration of phlebotomy treatments in patients with Hereditary Hemochromatosis:
a model based approach

Miguel Costa Carvalho Faria

Mestrado em Bioinformática e Biologia Computacional
2019

Orientador
Maria Eduarda da Rocha Pinto Augusto Silva, Professor Associado, Faculdade de Economia da Universidade do Porto

Coorientador
Fernando Manuel dos Santos Tavares, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

Supervisores da Oehoe Data Science
Teresa Maria Alves da Mota, Data Science, Web & App Development Consultant, Oehoe Data Science Portugal Lda.

Micha Bouts, Founder and Managing Director, Oehoe Data Science Portugal Lda.

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

MSc

Abstract

Background: Phlebotomy is the mainstay of treatment for Hereditary Hemochromatosis (HH) associated with mutations in the Hereditary Hemochromatosis Protein (HFE), a genetic disorder that leads to an iron accumulation in the tissues. However, the duration of the initial iron depletion phase of the treatment, commonly referred as Depletion Phase (DP), is highly variable among individuals. In this study, by analyzing data related to 384 patients with HFE HH, we aim at understanding the underlying factors affecting the duration of the DP and, subsequently, predict the personalized duration of this treatment phase for newly diagnosed patients with HFE-related HH.

Results: This study confirmed that the serum ferritin (SF) levels at diagnosis and the homozygous genotype (C282Y/C282Y) are associated with longer durations of the DP. Moreover, exponential and linear approximations of the rates of depletion of SF and the frequency of phlebotomy therapy also seemed to influence the duration of this treatment phase. Homozygous patients were more associated with higher initial SF concentrations, as well as male individuals and older patients. As patient specific values of exponential and linear approximations of the decay of SF during the DP are unknown at the time of the diagnosis, median values of these parameters, based on statistically significant differences between groups of other patients' factors, were calculated. The approach here presented relies on assigning these values to newly diagnosed patients and building regression models with these parameters as explanatory variables separately. Ultimately, the model with the best predictive accuracy for newly diagnosed patients with HFE-related HH was a Linear Regression with Box-Cox transformation, using data without influential data points and using the exponential approximation of the decay of SF as a predictor (MAPE = 46.8%, with 10-fold CV).

Conclusion: In sum, this thesis helped identifying factors affecting the duration of the DP, such as the initial SF level, the genotype, the frequency of phlebotomy therapy and the rate of depletion of SF. Although the regression models may not give, at this point, sufficiently accurate personalized predictions of the duration of the DP, the models assessed provide prediction intervals that may aid on the physicians' decision making, by at least giving an estimate of the minimum and maximum duration of the DP. Indeed, this study seemed to be a helpful first attempt to predict the duration of the DP for HFE-related HH patients.

Keywords: HFE-related Hereditary Hemochromatosis, phlebotomy, depletion phase, serum ferritin decay, nonlinear least squares, Linear Regression, Generalized Linear Models, K-fold Cross Validation.

Resumo

Contexto e objetivo: A flebotomia é amplamente utilizada como tratamento para a Hemocromatose Hereditária (HH) associada a mutações na Proteína da Hemocromatose Hereditária (HFE), um distúrbio genético que leva à acumulação de ferro nos tecidos. No entanto, a duração da fase inicial de depleção de ferro do tratamento, usualmente designada de Fase de Depleção (DP), é consideravalmente variável entre indivíduos. Neste estudo, analisando dados de 384 pacientes com HFE HH, tentámos compreender os fatores subjacentes que afetam a duração da DP e, posteriormente, prever a duração personalizada dessa fase de tratamento para pacientes recém-diagnosticados com HH relacionada com a HFE.

Resultados: Este estudo confirmou que os níveis séricos de ferritina (SF) no diagnóstico e o genótipo homozigótico (C282Y/C282Y) estão associados a durações mais longas da DP. Além disso, aproximações exponenciais e lineares das taxas de depleção de SF e a frequência da terapia de flebotomia também parecem influenciar a duração dessa fase de tratamento. Pacientes homozigóticos pareceram mais associados a concentrações iniciais de SF superiores, bem como indivíduos do sexo masculino e pacientes mais velhos. Como os valores específicos dos pacientes das aproximações exponenciais e lineares do decaimento do SF durante a DP são desconhecidos no momento do diagnóstico, foram calculados valores medianos desses parâmetros, com base em diferenças estatisticamente significativas entre grupos de outros fatores ou características dos pacientes. A abordagem aqui apresentada baseia-se na atribuição desses valores a pacientes recém-diagnosticados e na construção de modelos de regressão com esses parâmetros como variáveis explicativas separadamente. Por fim, o modelo com a melhor precisão preditiva para pacientes recém-diagnosticados com HH associada à HFE foi uma regressão linear com transformação Box-Cox, utilizando dados sem observações influentes e usando a aproximação exponencial do decaimento da SF como preditor (MAPE = 46,8%, com 10-fold Cross Validation).

Conclusão: Resumidamente, esta dissertação ajudou a identificar fatores que afetam a duração da DP, como o nível inicial de SF, o genótipo, a frequência da terapia de flebotomia e a taxa de depleção de SF. Embora os modelos de regressão possam não fornecer, neste momento, previsões personalizadas suficientemente precisas da duração da DP, os modelos avaliados providenciam intervalos de previsão que podem auxiliar a tomada de decisão dos médicos, dando pelo menos uma estimativa da duração mínima e máxima da DP. De facto, este estudo apresenta uma tentativa preliminar útil para prever a duração da DP de pacientes com HFE HH.

Palavras-chave: Hemocromatose Hereditária, flebotomoia, fase de depleção, decaimento da ferritina sérica, nonlinear least squares, Regressão Linear, Modelos Lineares Generalizados, K-fold Cross Validation.

Preface

This thesis' work was developed under a joint collaboration between the Faculty of Sciences of the University of Porto, Oehoe Data Science Portugal Lda. and the Department of Metabolic Diseases of the Leuven University Hospital (UZ Leuven, Campus Gasthuisberg) in Belgium. The data used and analyzed on this study was provided by the Leuven University Hospital. Besides the thesis supervisors, Professor Dr. Maria Eduarda Silva from the Faculty of Economics of the University of Porto, Professor Dr. Fernando Tavares from the Faculty of Sciences of the University of Porto, Ms. Teresa Mota, MSc, and Mr. Micha Bouts, MSc, from Oehoe Data Science Portugal Lda., Dr. Annick Vanclooster and Professor Dr. David Cassiman from the Leuven University Hospital actively cooperated on this work by giving clinical information or advises regarding the data used.

Acknowledgments

Desejo exprimir o meu agradecimento a todos aqueles que, direta ou indiretamente, me ajudaram nesta etapa académica que chega agora ao fim.

Em primeiro lugar, gostaria de agradecer ao Professor Doutor Fernando Tavares, meu orientador e diretor do mestrado em Bioinformática e Biologia Computacional, por ter acreditado neste projeto e por ter feito todos os esforços possíveis para que fosse realizado, mesmo sendo um processo moroso envolvendo várias entidades. Muito obrigado pelo acompanhamento ao longo do ano e pela disponibilidade e preocupação demonstradas.

À Professora Doutora Maria Eduarda Silva, queria deixar o meu agradecimento por ter aceite, prontamente, fazer parte deste estudo. Agradeço toda a disponibilidade, o conhecimento e as ideias que me transmitiu. Obrigado por acreditar em mim e neste projeto.

Um enorme obrigado à Teresa Mota e ao Micha Bouts, da Oehoe Data Science Portugal Lda., por tão bem me integrarem na vossa empresa, num projeto com tanta relevância, acreditando em mim desde o início. Obrigado pelo apoio constante, disponibilidade e preocupação demonstradas ao longo deste ano. Obrigado pela compreensão que sempre mostraram, pela confiança que tiveram em mim e pela excelente oportunidade de realizar um trabalho com tamanha importância clínica.

Gostaria de agradecer à Dr. Annick Vanclooster e ao Professor Doutor David Cassiman do Hospital Universitário de Leuven por disponibilizarem os dados utilizados neste trabalho. Para além disso, queria agradecer o acompanhamento dado pela Dr. Annick Vanclooster, que sempre se mostrou disponível para esclarecer qualquer questão relacionada com os dados e o seu significado clínico.

Não poderia deixar de agradecer à Doutora Goreti Carneiro e a toda a equipa da Unidade de Comunicação, Imagem e Cooperação da Faculdade de Ciências da Universidade do Porto pelo esforço e dedicação para efetivar todos os protocolos de cooperação, de forma a possibilitar a realização deste trabalho.

Obrigado aos meus amigos de mestrado, ao Luís, ao Filipe, ao Miguel e ao Pedro. Esta aventura foi incrivelmente melhor ao vosso lado.

Obrigado ao Luís e ao Nuno, amigos de faculdade, por estes dois anos.

À minha mãe, ao meu pai e à minha irmã, um enormíssimo obrigado. Nunca faria sentido sem vocês, sem o vosso apoio incansável, sem a vossa vontade em me ver alcançar o que ambiciono. A vocês, nunca conseguirei agradecer o suficiente.

Mariana, o meu obrigado não cabe em nenhuma dissertação do mundo. Acabo esta etapa a teu lado e tu começas uma ainda maior a meu lado. Obrigado por acreditares sempre em mim, por estares sempre comigo e, principalmente, me conseguires aturar sempre que começo a falar de estatística.

Um obrigado especial à minha avó pela companhia nestes últimos meses. A tua boa disposição constante animou os meus dias e deu-me força para continuar.

A toda a minha familia, tios, primos e, especialmente, aos meus avós, muito obrigado por estarem sempre comigo.

Aos meus amigos de sempre, obrigado por todos os momentos juntos, são esses que ficam e serão sempre lembrados.

Muito obrigado a todos!

Contents

Abstract **i**

Resumo **iii**

Preface **v**

Acknowledgments **vii**

Contents **xi**

List of Tables **xvi**

List of Figures **xxi**

Acronyms **xxiii**

1 Introduction **1**

1.1 Iron-overload disease 1

1.2 HFE-related Hereditary Hemochromatosis 2

1.2.1 Epidemiology 3

1.2.2 Iron metabolism 4

1.2.3 Pathophysiology 5

1.2.4 Clinical manifestations 6

1.2.5 Diagnosis and screening 7

1.2.6 Treatment and management 9

1.3 Motivation 10
1.4 Aims 11
1.5 Thesis outline 11

2 Materials and Methods 13

2.1 Data description 13
2.2 R Programming Language 14

3 Modeling the Depletion Phase 15

3.1 Data processing 15
3.1.1 Grouping of patients based on initial SF value 15
3.1.2 Selection of SF time-stamped values from the DP 15
3.1.3 Retrieval of the initial SF values 17
3.1.4 Treatment interruptions during the DP 18
3.1.5 The behavior of the decay of SF over time 19
3.1.6 Nonlinear Least Squares analysis 21
3.1.7 Retrieval of k parameters 23
3.1.8 The outcome variable: duration of the DP in weeks 24
3.1.9 Variables selection 25
3.1.10 Feature extraction 28
3.2 Exploratory Data Analysis 30
3.2.1 Variables' Correlations 30
3.2.2 DurationWeeks variable 32
3.2.3 Ferritin variable 40
3.2.4 kParameter variable 44
3.2.5 Gradient variable 50
3.2.6 AgeDiagnosis variable 56
3.2.7 Variables not included in the main analysis 57
3.3 Data processing after EDA: dealing with unknown variables at diagnosis 58

3.3.1 The MonthlyPhlebotomies variable 59

3.3.2 The kParameter and Gradient variables 59

3.4 Predictive models 63

3.4.1 Transformation of the outcome variable 63

3.4.2 kParameter vs Gradient 64

3.5 Linear Regressions 65

3.5.1 Outlier detection 71

3.6 Generalized Linear Models 74

3.6.1 Poisson Regressions 74

3.6.2 Negative Binomial Regressions 78

3.7 Model evaluation and comparison 81

4 Discussion **87**

5 Conclusion **93**

A Appendix **95**

Bibliography **125**

List of Tables

3.1 Appearance of the data set after pre-processing related to the retrieval of initial SF values, the duration of the DP and the k parameters from NLS analysis . . . 25

3.2 Number of missing values and number of patients of each level of the LiverDisease variable. Level 0 refers to patients with no known liver condition, level 1 refers to patients with Fibrosis and level 2 refers to patients with Cirrhosis 26

3.3 Number of missing values and number of patients of each level of the Alcohol intake variable. Level 0 refers to patients that do not consume alcohol or consume only sporadically and level 1 refers to patients that consume alcohol frequently . 26

3.4 Appearance of the data set after pre-processing steps and variable selection . . . 30

3.5 Table depicting correlation values between the target variable, DurationWeeks, and all the other variables included on the data set 32

3.6 Basic statistics for the DurationWeeks variable 33

3.7 Sample sizes for each Group of initial SF level 35

3.8 Games-Howell test for comparisons of DurationWeeks values between groups of initial SF level 36

3.9 Sample sizes for each Genotype 36

3.10 Sample sizes for each level of MonthlyPhlebotomies 38

3.11 Basic statistics for the Ferritin variable 40

3.12 Basic statistics for the kParameter variable 44

3.13 Basic statistics for the Gradient variable 50

3.14 Games-Howell test for comparisons of Gradient values between groups of initial SF level 52

3.15 Basic statistics for the AgeDiagnosis variable 56

3.16 Kruskal-Wallis tests for comparisons between groups, where AgeDiagnosis is grouped based on the categorical variables 57

3.17 kParameter values assigned to a newly diagnosed patient based on Genotype and MonthlyPhlebotomies. The attributed value corresponds to the median of each group 60

3.18 Gradient values assigned to a newly diagnosed patient based on Group of initial SF, Sex and MonthlyPhlebotomies. The attributed value corresponds to the median of each group 60

3.19 ANOVA table where: model 1) Ferritin, model 2) Ferritin + kParameter, model 3) Ferritin + kParameter + MonthlyPhlebotomies and model 4) Ferritin + kParameter + MonthlyPhlebotomies + AgeDiagnosis 67

3.20 ANOVA table where: model 1) Ferritin, model 2) Ferritin + Gradient, model 3) Ferritin + Gradient * MonthlyPhlebotomies, model 4) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis, model 5) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis + Genotype and model 6) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis + Genotype + Sex 69

3.21 Summary of linear regression models, one with kParameter and another with Gradient, on data without respective influential data points 73

3.22 Summary of Poisson regression models, one with kParameter and another with Gradient, on data with all the patients 76

3.23 Summary of Poisson regression models, one with kParameter and another with Gradient, on data without influential data points 77

3.24 Summary of NB regression models, one with kParameter and another with Gradient, on data with all the patients 79

3.25 Summary of NB regression models, one with kParameter and another with Gradient, on data without influential data points 80

3.26 10-fold cross-validation results using the MAPE metric for all the final models considered, in which the test set has the real patient specific values for the variables that describe the decay of SF over time (kParameter and Gradient) 82

3.27 10-fold cross-validation results using the MAPE metric for all the final models considered, in which the test set has the assigned k and gradient values (kParameter2 and Gradient2) 82

3.28 Patients from test set, consisting of 10% of the data set without influential data points and with assigned kParameter2 and Gradient2 values, with predictions (fit) and prediction intervals (lower and upper bounds) using the linear regression with data transformation with the kParameter as an explanatory variable 85

A.1 Time-stamped SF values during the DP for a patient from Group 0 (initial SF value < 500 μg/L) 95

A.2 Variables on the data set 1 that has information regarding time-stamped biochemical parameters 96

A.3 Variables on the data set 2 that has information regarding time-stamped biochemical parameters 96

A.4 Variables on the data set 3 that has information regarding static parameters . . . 101

A.5 Median values of the numerical variables based on Group of initial SF level . . . 101

A.6 Median values of the numerical variables based on Genotype 102

A.7 Median values of the numerical variables based on Sex 102

A.8 Sample sizes of each Sex 103

A.9 Median values of the numerical variables based on MonthlyPhlebotomies 103

A.10 Games-Howell test for comparisons of AgeDiagnosis values between groups of initial SF level 106

A.11 kParameter values assigned to a newly diagnosed patient based on Genotype, TransfGroup and MonthlyPhlebotomies. The attributed value corresponds to the median of each group 111

A.12 Summary of second variable addition, on a linear regression model with Ferritin as explanatory variable 113

A.13 Summary of third variable addition, on a linear regression model with Ferritin + kParameter as explanatory variables 114

A.14 VIFs of model with Ferritin and an interaction between kParameter and MonthlyPhlebotomies 115

A.15 VIFs of model with Ferritin, kParameter and MonthlyPhlebotomies 115

A.16 Summary of fourth variable addition, on a linear regression model with Ferritin + kParameter + MonthlyPhlebotomies as explanatory variables 116

A.17 AIC values where: model 1) Ferritin, model 2) Ferritin + kParameter, model 3) Ferritin + kParameter + MonthlyPhlebotomies and model 4) Ferritin + kParameter + MonthlyPhlebotomies + AgeDiagnosis 116

A.18 Summary of third variable addition, on a linear regression model with Ferritin + Gradient as explanatory variables 117

A.19 VIFs of model with Ferritin and an interaction between Gradient and MonthlyPhlebotomies 118

A.20 Summary of fourth variable addition, on a linear regression model with Ferritin + Gradient * MonthlyPhlebotomies as explanatory variables 118

A.21 AIC values where: model 1) Ferritin, model 2) Ferritin + Gradient, model 3) Ferritin + Gradient * MonthlyPhlebotomies and model 4) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis 119

A.22 Influential data points discovered on linear regression model with the kParameter variable 119

A.23 Influential data points discovered on linear regression model with the Gradient variable 119

List of Figures

1.1 Types of HH. Adapted from Kris V. Kowdley et al. [60] 2

1.2 Hepcidin regulation. Adapted from Pierre Brissot et al. [20] 5

1.3 Algorithm for HFE HH diagnosis and screening. Adapted from Kris V. Kowdley et al. [60] 8

3.1 Plot depicting the SF values for a patient from Group 0 (initial SF value < 500 μg/L) before removing data points associated with the MP (a) and after (b). On plot (a), black vertical line indicates approximate end point of the DP. On both plots, blue horizontal dashed line indicates threshold of SF concentration (100 μg/L) 16

3.2 Plot depicting the SF values for a patient from Group 500 (initial SF value ≥ 500 μg/L and ≤ 1000 μg/L) before removing data points associated with the MP (a) and after (b). On plot (a), black vertical line indicates approximate end point of the DP. On both plots, blue horizontal dashed line indicates threshold of SF concentration (100 μg/L) 17

3.3 Plot depicting the SF values for a patient from Group 1000 (initial SF value $>$ 1000 μg/L) before removing data points associated with the MP (a) and after (b). On plot (a), black vertical line indicates approximate end point of the DP. On both plots, blue horizontal dashed line indicates threshold of SF concentration (100 μg/L) 17

3.4 Plot depicting the SF values during the DP for a patient from Group 0 (initial SF value < 500 μg/L) before treatment interruptions > 100 days removal (a) and after (b). Red vertical line indicates cut-off point. 18

3.5 Plot depicting the SF values during the DP for a patient from Group 500 (initial SF value ≥ 500 μg/L and ≤ 1000 μg/L) before treatment interruptions > 100 days removal (a) and after (b). Red vertical line indicates cut-off point. 19

3.6 Plot depicting the SF values during the DP for a patient from Group 1000 (initial SF value > 1000 μg/L) before treatment interruptions > 100 days removal (a) and after (b). Red vertical line indicates cut-off point. 19

3.7 Plots depicting the SF values during the DP for 6 patients 20

3.8 Examples of monoexponential decays with the equation specified on the NLS analysis varying only the k parameter value. For all the lines, A = 1000 and Days = 1000 22

3.9 Plots with NLS fitted curves for three patients, one of each initial SF group. Patient number 154 is from Group 0 (k = -0.0202), patient number 177 is from Group 500 (k = -0.00216) and patient number 1 is from Group 1000 (k = -0.0102). Plots on the left are scale-free. Plots on the right have the same scale. 23

3.10 Flowchart of pre-processing steps until the retrieval of the k parameters 24

3.11 Correlation plot with coloured scale where red implies negative correlation and blue colors imply positive correlations. 31

3.12 Histogram of DurationWeeks values 33

3.13 Histogram of DurationWeeks values based on Group of initial SF level 34

3.14 Box plot of DurationWeeks values based on Group of initial SF level 35

3.15 Histogram of DurationWeeks values based on Genotype 37

3.16 Box plot of DurationWeeks values based on Genotype 37

3.17 Histogram of DurationWeeks values based on MonthlyPhlebotomies 39

3.18 Box plot of DurationWeeks values based on MonthlyPhlebotomies 39

3.19 Histogram of Ferritin values 41

3.20 Histogram of Ferritin values based on Genotype 42

3.21 Box plot of Ferritin values based on Genotype 42

3.22 Histogram of Ferritin values based on Sex 43

3.23 Box plot of Ferritin values based on Sex 44

3.24 Histogram of kParameter values 45

3.25 Histogram of kParameter values based on Group 46

3.26 Box plot of kParameter values based on Group 46

3.27 Histogram of kParameter values based on Genotype 47

3.28 Box plot of kParameter values based on Genotype 48

3.29 Histogram of kParameter values based on MonthlyPhlebotomies 49

3.30 Box plot of kParameter values based on MonthlyPhlebotomies 49

3.31 Histogram of Gradient values 50

3.32 Histogram of Gradient values based on Group 51

3.33 Box plot of Gradient values based on Group 52

3.34 Histogram of Gradient values based on Sex 53

3.35 Box plot of Gradient values based on Sex 54

3.36 Histogram of Gradient values based on MonthlyPhlebotomies 55

3.37 Box plot of Gradient values based on MonthlyPhlebotomies 55

3.38 Histogram of AgeDiagnosis values 56

3.39 Plot of decays of SF with the assigned kParameter2 values 61

3.40 Plot of decays of SF with the assigned Gradient2 values 61

3.41 DurationWeeks values and associated Ferritin values with a regression line before (a) and after Box-Cox transformation (b) 64

3.42 Interaction plots in which (a) depicts an interaction between kParameter and MonthlyPhlebotomies and (b) depicts an interaction between Gradient and MonthlyPhlebotomies 65

3.43 Diagnostic plots of final linear regression model considered including the kParameter variable 68

3.44 Diagnostic plots of final linear regression model considered including the Gradient variable 70

3.45 Cook's distance plots to detect influential data points for (a) linear regression with the kParameter variable and (b) linear regression with the Gradient variable. Data points above red threshold line are considered influential points 71

3.46 Random simulation of a Poisson distribution (a) and the distribution of DurationWeeks (b) 74

3.47 Plot showing the predicted values (blue circles) and the true values (black circles) of the duration in weeks, in which arrows link the corresponding predictions to their true values. The model used was the linear regression with data transformation, with the kParameter as a predictor, trained on 90% of the data set without influential data points and with the assigned kParameter2 and Gradient2 values . 84

A.1 Correlation plot showing scatter plots between variables and associated r correlation values . . . 101

A.2 Histogram of DurationWeeks values based on Sex . . . 102

A.3 Box plot of DurationWeeks values based on Sex . . . 103

A.4 Histogram of kParameter values based on Sex . . . 104

A.5 Box plot of kParameter values based on Sex . . . 104

A.6 Histogram of Gradient values based on Genotype . . . 105

A.7 Box plot of Gradient values based on Genotype . . . 105

A.8 Histogram of AgeDiagnosis values based on Group . . . 106

A.9 Box plot of AgeDiagnosis values based on Group . . . 107

A.10 Histogram of AgeDiagnosis values based on Sex . . . 107

A.11 Box plot of AgeDiagnosis values based on Sex . . . 108

A.12 Correlation plot with coloured scale where red implies negative correlation and blue colors imply positive correlations. Data set with variables disregarded from the main analysis (n = 52) . . . 109

A.13 Decision Tree depicting the grouping of the kParameter values based on Genotype and MonthlyPhlebotomies. Gradient values correspond to the median of each group 109

A.14 Decision Tree depicting the grouping of the Gradient values based on Group of initial SF level, Sex and MonthlyPhlebotomies. Gradient values correspond to the median of each group . . . 110

A.15 Box plot of kParameter values based on TransfGroup . . . 111

A.16 Plot depicting the interval of best Lambdas found with 95% of confidence using the Box-Cox method . . . 112

A.17 Diagnostic plots of final linear regression model considered including the kParameter variable with the data without influential data points . . . 112

A.18 Diagnostic plots of final linear regression model considered including the Gradient variable with the data without influential data points . . . 115

A.19 Diagnostic plots of final poisson regression model considered including the kParameter variable with the data from all the patients . . . 120

A.20 Diagnostic plots of final poisson regression model considered including the Gradient variable with the data from all the patients . . . 120

A.21 Diagnostic plots of final poisson regression model considered including the kParameter variable with the data without influential data points 121

A.22 Diagnostic plots of final poisson regression model considered including the Gradient variable with the data from without influential data points 121

A.23 Diagnostic plots of final NB regression model considered including the kParameter variable with the data from all the patients 122

A.24 Diagnostic plots of final NB regression model considered including the Gradient variable with the data from all the patients 122

A.25 Diagnostic plots of final NB regression model considered including the kParameter variable with the data without influential data points 123

A.26 Diagnostic plots of final NB regression model considered including the Gradient variable with the data from without influential data points 123

Acronyms

AIC	akaike information criterion
ALT	alanine transaminase
ANOVA	analysis of variance
AST	aspartate transaminase
BMI	body mass index
BMP2	bone morphogenetic protein 2
BMP6	bone morphogenetic protein 6
BMPR	bone morphogenetic protein receptor
CV	cross validation
DP	depletion phase
EDA	exploratory data analysis
ERK/MAPK	mitogen-activated protein kinase
FPN	ferroportin
GLMs	generalized linear models
HAMP	hepcidin precursor
HB	hemoglobin
HCC	hepatocellular carcinoma
HCV	hepatitis C vírus
HFE	hereditary hemochromatosis protein
HH	hereditary hemochromatosis
HJV	hemojuvelin protein
HLA	human leukocyte antigen
MAPE	mean absolute percentage error
MP	maintenance phase
MRI	magnetic resonance imaging
NB	negative binomial
NLS	nonlinear least squares
NTBI	non-transferrin bound iron
ROS	reactive oxygen species
SF	serum ferritin
SMADs	small-mothers-against-decapentaplegic proteins
TfR1	transferrin receptor 1
TfR2	transferrin receptor 2
TS	transferrin saturation
VIF	variance inflation factor

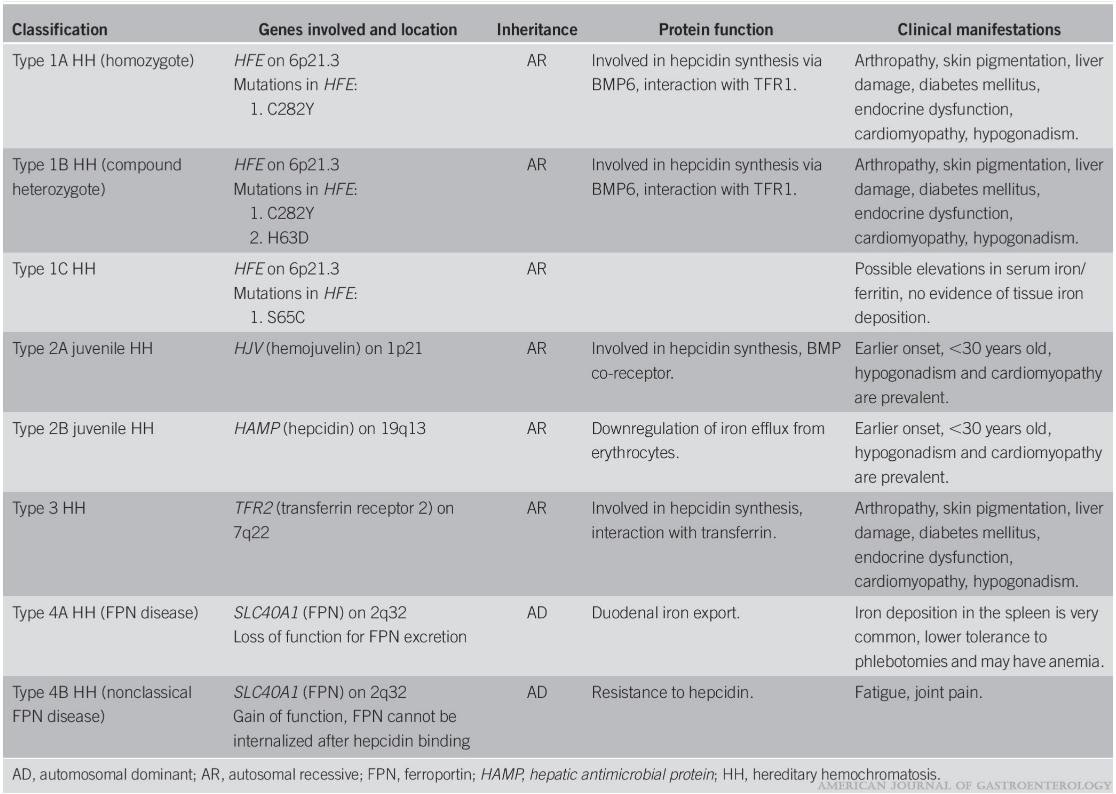
Chapter 1

Introduction

In this work, data from patients with a disease that deregulates the iron metabolism is analyzed. As such, the following sections serve as a contextualization of the disorder, its mechanism of action, diagnosis and treatment. The motivation and aims of this study are briefly explained on this Chapter. Furthermore, an outline of the thesis is presented on the last section.

1.1 Iron-overload disease

Iron is a crucial metal for hemoglobin (HB) synthesis of erythrocytes, cellular proliferation and oxidation-reduction reactions [58]. Most of the iron in the body is distributed between the hemoglobin of red cells, the liver, muscles and macrophages of the reticuloendothelial system [101]. However, excessive iron accumulation may lead to the production of reactive oxygen species (ROS), damaging tissues and organs [34, 58]. As such, iron homeostasis is fundamental in most organisms to guarantee a balance of iron for vital biological processes while avoiding its toxicity related to its excess [7, 58]. Indeed, iron-overload diseases can cause progressive and irreversible end-organ damage and are generally categorized in two different forms, primary or secondary [34, 83]. Secondary iron-overload disorders may be related to multiple transfusions, thalassemia major, cirrhosis or other factors [83, 101]. On the other hand, primary iron-overload is associated with Hereditary Hemochromatosis (HH) [58, 83, 101]. The majority of the cases of HH arise from alterations in genes that regulate hepcidin synthesis, including the most common mutation in the gene responsible for encoding hereditary hemochromatosis protein (HFE), comprising 80% of the cases, the transferrin receptor 2 (TfR2), the hemojuvelin protein (HJV), the hepcidin precursor (HAMP) or the ferroportin (FPN) [78, 79, 83]. The types of HH are presented in Figure 1.1. On the next sections, a comprehensive review of the HFE-related HH, also recognized as type-1 HH, will be performed, as this is the type of HH studied on this thesis.



Classification	Genes involved and location	Inheritance	Protein function	Clinical manifestations
Type 1A HH (homozygote)	*HFE* on 6p21.3 Mutations in *HFE*: 1. C282Y	AR	Involved in hepcidin synthesis via BMP6, interaction with TFR1.	Arthropathy, skin pigmentation, liver damage, diabetes mellitus, endocrine dysfunction, cardiomyopathy, hypogonadism.
Type 1B HH (compound heterozygote)	*HFE* on 6p21.3 Mutations in *HFE*: 1. C282Y 2. H63D	AR	Involved in hepcidin synthesis via BMP6, interaction with TFR1.	Arthropathy, skin pigmentation, liver damage, diabetes mellitus, endocrine dysfunction, cardiomyopathy, hypogonadism.
Type 1C HH	*HFE* on 6p21.3 Mutations in *HFE*: 1. S65C	AR		Possible elevations in serum iron/ferritin, no evidence of tissue iron deposition.
Type 2A juvenile HH	*HJV* (hemojuvelin) on 1p21	AR	Involved in hepcidin synthesis, BMP co-receptor.	Earlier onset, <30 years old, hypogonadism and cardiomyopathy are prevalent.
Type 2B juvenile HH	*HAMP* (hepcidin) on 19q13	AR	Downregulation of iron efflux from erythrocytes.	Earlier onset, <30 years old, hypogonadism and cardiomyopathy are prevalent.
Type 3 HH	*TFR2* (transferrin receptor 2) on 7q22	AR	Involved in hepcidin synthesis, interaction with transferrin.	Arthropathy, skin pigmentation, liver damage, diabetes mellitus, endocrine dysfunction, cardiomyopathy, hypogonadism.
Type 4A HH (FPN disease)	*SLC40A1* (FPN) on 2q32 Loss of function for FPN excretion	AD	Duodenal iron export.	Iron deposition in the spleen is very common, lower tolerance to phlebotomies and may have anemia.
Type 4B HH (nonclassical FPN disease)	*SLC40A1* (FPN) on 2q32 Gain of function, FPN cannot be internalized after hepcidin binding	AD	Resistance to hepcidin.	Fatigue, joint pain.

AD, automosomal dominant; AR, autosomal recessive; FPN, ferroportin; *HAMP, hepatic antimicrobial protein*; HH, hereditary hemochromatosis.

AMERICAN JOURNAL OF GASTROENTEROLOGY

Figure 1.1: Types of HH. Adapted from Kris V. Kowdley et al. [60]

1.2 HFE-related Hereditary Hemochromatosis

The most frequent form of HH is associated with mutations in the HFE gene, resulting in decreased production of hepcidin [60, 83, 105]. Also, the most common mutation is a G to A transition at nucleotide 845 of the HFE gene, resulting in a cysteine to tyrosine substitution at amino acid 282, referred as p.C282Y (type 1a) [60, 83]. This mutation is found almost exclusively on white individuals [20]. HFE-associated HH leads to hepcidin deficiency and consequently elevated release of iron from splenic macrophages and cells of the small intestine into the plasma. This increase in iron plasma levels results in increased iron transport into parenchymal cells, especially hepatocytes, pancreatic cells and cardiomyocytes, and thus, hepatic, pancreatic and cardiac iron-overload [6, 20]. Ultimately, increased plasma iron and transferrin saturation (TS) are two phenotype features associated with HFE HH [6, 20, 60, 89]. TS can be defined as the ratio of the number of occupied iron binding sites to the total number of iron binding sites on plasma transferrin, while serum ferritin (SF) is the principal iron storage protein [20, 89]. Both TS and SF values are used for diagnosis and treatment monitoring [89]. While a large number of C282Y homozygous do not develop clinically significant iron overload, and consequently no symptoms, those who do have an inappropriate iron accumulation may have tissue complications, resulting in organ damage, diabetes mellitus, osteoporosis, hepatocellular carcinoma (HCC) or cirrhosis [20, 60]. Other symptomatic manifestations may include chronic fatigue, hepatic fibrosis, skin pigmentation and joint problems [20]. Also, patients diagnosed with HFE-associated HH may be asymptomatic for many years [20].

Besides the most prevalent homozygous genotype (C282Y/C282Y), HFE HH can be related to H63D mutations (type 1b) or S65C mutations (type 1c) [60]. Also called compound heterozygote, the C282Y/H63D is a genetic subtype associated with HFE HH much rarer than the homozygous type 1a. Even if prone to increased TS and SF levels, patients with the compound heterozygote genotype, or the H63D/H63D genotype, rarely develop clinically significant iron-overload, unless cofactors such as alcohol, liver disease or hepatitis C virus (HCV) are involved [20, 60]. Similarly, type 1c HH, related with the S65C mutation, does not substantially affect the phenotype [20, 60]. These genetic subtypes, type 1b and type 1c, are less prevalent than the type 1a and may not be sufficient to result in clinical manifestations related to hemochromatosis and have uncertain pathogenic relevance, according to recent studies [6, 20, 60, 78].

1.2.1 Epidemiology

The C282Y mutation is highly related to white individuals, especially northern European descendants, as they have a carrier frequency of 1 in 10 individuals and a prevalence of homozygosity of approximately 5 of every 1000 individuals [20]. Indeed, the frequency of this mutation decreases from the northwest to southwest Europe, in accordance with the settlements of ancient Celts [60, 105]. Some studies suggest that this mutation provided survival advantage as these populations had poor iron diet [105]. About 80%-90% of the northern European individuals clinically diagnosed with HH are homozygous for the C282Y mutation [6]. Moreover, studies have demonstrated an average prevalence of 0.4% for C282Y homozygosity and 9.2% for C282Y heterozygosity while assessing samples from European countries and similar prevalences in North America [20]. In Asian, African and Middle Eastern populations, C282Y heterozygosity was detected with prevalences between 0% and 0.5%, albeit no C282Y homozygosity was detected [20]. According to the same studies, the C282Y/H63D compound heterozygote and the H63D/H63D homozygote genotypes had prevalences of 2% in the European population and 2.5% and 2.1% in the Americas, respectively [20, 79].

The penetrance of the disease is relatively low. Only 1%-33% of the homozygotes develop clinical manifestations related to iron-overload and the penetrance of the disorder seems associated with gender - 28,4% of males and only 1,2% of females showed iron-related disease [6, 20, 89]. Although, 81.8% of male individuals and 55.4% of female individuals had increased SF levels, which suggests that biochemical penetrance is higher than clinical penetrance [20]. Also, genetic modifiers, environmental factors and lifestyle factors seem to influence somehow the penetrance of HFE-related HH [20]. Indeed, excess alcohol consumption, blood loss either due to menstruation or routine blood donations, or increased dietary iron intake may be cofactors that may increase the phenotypic expression of HH [20].

1.2.2 Iron metabolism

The human body contains approximately 3-5 g of iron, of which between 60%-70% is used within HB in circulating red blood cells [75]. Approximately 20%-30% of body iron is stored in hepatocytes and in reticuloendothelial macrophages, mainly within ferritin [75, 101]. The duodenum plays an important role on the absorption of dietary iron [75, 115]. A healthy individual absorbs, daily, between 1-2 mg of iron compensating for iron losses related to skin and intestine cells desquamation, menstruation or sporadic blood losses [75]. The absorbed iron can be stored in ferritin, in the duodenal enterocytes, or bound to plasma transferrin in circulation, forming holotransferrin, which may transport the iron for subsequent tissue intake [75, 115]. Thereafter, iron is used for many biological processes like erythropoiesis in the bone marrow, myoglobin synthesis in muscle and oxidative metabolism in respiring cells [115].

Maintaining normal levels of TS (between 20% and 45%) is a fundamental part of iron homeostasis to avoid disorders of iron metabolism, either its deficiency or excess accumulation [20]. Besides the dietary iron absorbed, the macrophages responsible for erythrophagocytosis, a process in which senescent erythrocytes are degraded and their iron recycled, are also a source of plasma iron [20, 115]. The transport of iron from enterocytes and macrophages into the plasma occurs through the FPN, expressed on the membranes of these two cells [20]. FPN is mainly regulated by hepcidin, an iron-regulated peptide secreted by hepatocytes [20]. When hepcidin binds to FPN, it induces its internalization and degradation and, hence, plasma hepcidin levels strongly affect plasma iron concentration [20, 101]. So, low hepcidin levels trigger increased iron absorption from the duodedum and iron release from enterocytes and macrophages, leading to increased plasma iron concentration and TS [101]. Conversely, elevated secretion of hepcidin by hepatocytes leads to decreased iron absorption and iron retention in reticuloendothelial macrophages [101]. Indeed, elevated holotransferrin levels lead to the secretion of hepcidin to plasma to control the iron export [20].

The main regulator of hepcidin expression is the HAMP, in which its transcription promotes increased expression of hepcidin (Figure 1.2) [20, 115]. Several proteins found on the hepatocellular membrane act as sensors of iron levels and regulate the hepcidin synthesis [115]. One of the pathways that controls the HAMP transcription involves the HJV and the bone morphogenetic protein receptor (BMPR), which forms a protein complex (HJV-BMPR), that is reactive to bone morphogenetic protein 6 (BMP6) and to bone morphogenetic protein 2 (BMP2) [20]. When BMP6 or BMP2, produced in sinusoidal cells, hepatic stellate cells and hepatocytes when cell iron stores are increased, bind to the HJV-BMPR complex, they activate it and induce the phosphorylation of cytosolic small-mothers-against-decapentaplegic proteins (SMADs) 1, 5 and 8 [115, 117]. Thereafter, these SMADs bind to SMAD4, forming a protein complex that enters the nucleus and binds to the HAMP promoter, resulting in its transcription and subsequent hepcidin synthesis [20, 21, 115]. Additionally, TS is also involved in hepcidin regulation [20]. Increased TS may induce a shift of the interaction between HFE, transferrin receptor 1 (TfR1) and TfR2 on the hepatocellular membrane, which leads to signalling to increase HAMP transcription via the extracellular signal-regulated kinase/mitogen-activated protein kinase (ERK/MAPK), resulting in hepcidin expression [20, 115]. Also, studies suggest that the HFE-TfR1-TfR2 complex may interact with the HJV-BMPR complex, hinting at a key role of the HFE on the regulation of hepcidin synthesis [38, 115].

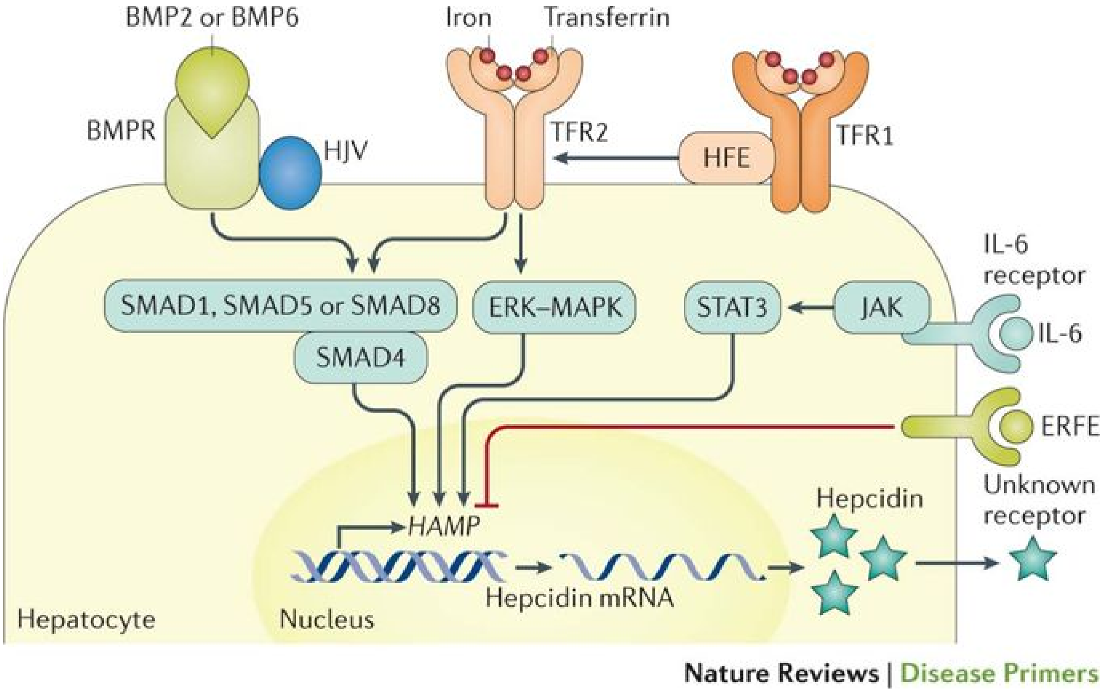


Figure 1.2: Hepcidin regulation. Adapted from Pierre Brissot et al. [20]

1.2.3 Pathophysiology

As previously mentioned, type 1 HH is associated with mutations in the HFE gene. HFE, a human leukocyte antigen (HLA) class I molecule, is expressed on cell membranes associated with β2-microglobulin [20, 78]. The C282Y mutation, the most common pathogenic mutation of HFE, is associated with the disruption of a disulfide bond in HFE, affecting its conformation and its ability to bind to β2-microglobulin [20, 30, 78]. Thus, interactions with TfR2, TfR1 and the HJV-BMPR complex are also affected, ultimately leading to decreased hepcidin synthesis [20]. Hepcidin deficiency is responsible for excessive expression of FPN at the cell surface, resulting in increased intestinal iron absorption and iron egress, which increase the plasma iron concentration and the TS, leading to the occurrence of non-transferrin bound iron (NTBI) [20, 115]. Indeed, the NTBI can form when TS is $> 45\%$, which is a common phenotypic manifestation in HFE-related HH, and is involved in the production of reactive oxygen species [20]. NTBI is taken up by hepatic, pancreatic, endocrine and cardiac cells, causing parenchymal iron excess, which may result in organ damage and other complications associated with HH [20, 84, 115].

1.2.4 Clinical manifestations

Manifestations of HFE HH usually occur in middle-aged patients and are diverse because the iron accumulation can occur in multiple tissues [20, 89]. In fact, manifestations may vary from only the genetic abnormalities (genotype), biochemical abnormalities related to increased TS and SF levels (biochemical phenotype) or organ damage (clinical phenotype) [20, 79, 89]. Classic features of HH, like cirrhosis, bronze-colored skin, diabetes, joint inflammation, heart disease or arthropathy are rarely found nowadays, as diagnosis for HH is possible at early stages due to enhanced screening techniques and a greater awareness of the disease among clinicians [20]. Common symptoms are now associated with chronic fatigue, joint and abdominal pain, malaise and hepatomegaly [20, 79, 89].

The TS is consistently increased among patients with HFE HH, along with increased SF levels, which indicate accumulation of iron in tissues [20, 79]. Studies have shown that 32% of male patients and 26% of female patients who were homozygous for HFE (C282Y/C282Y) had increased SF levels at diagnosis (> 300 μg/L for males and > 200 μg/L for females) [79]. Moreover, other studies suggested that 18% of man and 5% of women may have hepatic iron-overload, even if no clinical symptoms are present [60]. Indeed, symptoms seem to appear earlier for men, as clinical symptoms in women usually occur after postmenopause, due to iron loss during menstruation, pregnancy and lactation delaying the iron accumulation during this time [60, 79]. Other longitudinal studies, in which patients were followed up for more than 30 years, shown that 38% to 50% of C282Y homozygotes may develop iron-overload, while only between 10% to 33% may develop HH-associated morbidity [79, 108]. As stated before, according to another study, around 28% of males and only 1% of females develop iron-overload disease, suggesting that male C282Y homozygous are more prone to clinical manifestations related to iron-overload [6, 79].

The most commonly affected organ in type 1 HH is the liver, as iron is initially stored in this organ [60, 89]. However, the clinical presentation tends to vary, as it may be related to asymptomatic increases on serum aminotransferases, alanine transaminase (ALT) and aspartate transaminase (AST), or to liver disease [60, 79, 89]. In patients with HFE-related HH, SF levels > 1000 μg/L may indicate liver fibrosis and increases the risk of developing cirrhosis [60, 79]. After the development of cirrhosis, patients with HH are at increased risk of developing HCC [60]. Also, SF levels > 2000 μg/L seem to increase the risk of developing HCC [60]. Cirrhosis and HCC are the major causes of death among patients with HFE HH [89]. Indeed, HCC accounts for 45% of deaths in this population [60]. Studies demonstrated that elevated alcohol consumption and tobacco smoking may worsen iron-overload, in which alcohol intake > 60 g/d increases the risk of developing cirrhosis and > 80 g/d reduces survival, whereas the reduction of alcohol consumption over time led to a reduction in phenotypic expression of the disease [3, 35]. Although the concentration of iron stored in the liver can be used to predict the onset of cirrhosis, studies have shown that some HH patients develop this condition even at lower levels of iron-overload, whilst other patients with severe iron-overload do not have cirrhosis, suggesting that other factors besides iron-overload influence the development of cirrhosis and HCC [89].

Besides liver complications, iron-overload may affect other organs like the pancreas, heart or pituitary glands [60, 89]. As such, diabetes, cardiomyopathy and arrhythmias, hypogonadotropic hypogonadism, arthopathy and arthritis are clinical complications that may occur in HFE-associated HH patients [89]. Studies have shown that the prevalence of diabetes in type 1 HH patients is between 13% and 23% [76], while others shown that the prevalence of cardiomyopathy was 0.9%, in one study, and 3.1% on another [12, 29]. While conditions like cirrhosis, diabetes and cardiomyopathy are irreversible, others like weakness, fatigue, skin pigmentation and hepatic fibrosis may regress with appropriate treatment [60, 89]. Thus, screening, early diagnosis and treatment are essential to reduce the morbidity and mortality of type 1 HH patients [60, 79, 89].

1.2.5 Diagnosis and screening

Diagnosis of HH involves a sequential approach based on biochemical, clinical and imaging data assessment (Figure 1.3) [20, 57]. Common clinical manifestations should be taken into account to determine the possibility of iron-overload disease, although, as discussed before, symptoms can either be absent or very variable, which burdens the diagnosis of HFE-HH based solely on this analysis [20, 82]. As such, the main initial approach to diagnosis is through markers of iron stores, specifically TS and SF levels [10, 20, 60]. Increased TS ($> 45\%$) is usually the earliest biochemical manifestation observed in HFE HH, as it identifies 97.9%-100% of C282Y homozygotes [20, 60, 82]. Although, even if the cut-off TS value of 45% is often chosen, in some cases individuals with minor secondary iron-overload or C282Y/wild heterozygotes are identified as C282Y homozygotes, and, as such, further evaluation is needed [10]. Additionally, normal or low TS can be found even if iron-overload is present [20]. Furthermore, SF levels > 200 μg/L in premenopausal women and > 300 μg/L in men are also used to assess iron-overload, as they provide a valuable correlation with the degree of body iron stores [10, 60, 64]. Moreover, studies have shown the potential of SF as a marker to predict advanced fibrosis or cirrhosis in HFE-associated HH patients, in which SF levels > 1000 μg/L, along with increased ALT and ALT levels and reduced platelet count predicted the presence of cirrhosis in 80% of C282Y homozygotes [10, 60]. Nonetheless, SF suffers from low specificity as increased values can be due to other factors not related to iron-overload like inflammation, metabolic syndrome, diabetes mellitus, alcohol consumption or hepatocellular necrosis [20, 60, 102]. However, SF is a widely used biochemical parameter to unveil iron-overload related to HH and normal values of SF can be used to rule out iron-overload [64]. Actually, combined normal SF and TS values have a negative predictive value of 97% for excluding iron-overload occurrence [60]. Notably, a joint evaluation of SF levels and TS is usually necessary and is suggestive of HH [10, 89, 102]. A common strategy in HFE-related screening and diagnosis is to take into account serum iron markers (TS and SF) to target high-risk groups, including individuals with clinical manifestations associated with HH and/or with a family history of this genetic disorder [10].

Patients with consistently increased TS and SF, in the absence of hematological or inflammatory diseases, are suspected to have HH and should be referred for molecular genetic testing [60, 82, 89]. Indeed, to identify the genetic cause of the disease, testing for C282Y homozygosity is a suitable analysis and should be performed especially in white individuals of north European origin [20, 89]. If C282Y homozygosity is confirmed through genetic tests, first degree relatives should also be tested for the mutation and for biochemical phenotypic expression (TS and SF levels) [10, 89]. On the other hand, detecting C282Y heterozygosity in patients with iron-overload should prompt to investigate other genetic and/or acquired causes of iron accumulation, as compound heterozygotes do not usually develop iron-overload disease and only present mild iron-overload when other cofactors like alcoholism or metabolic syndrome are involved [20, 60].

Moreover, liver biopsy can be important to assess hepatic complications of HH, especially the stage of fibrosis in C282Y homozygous patients with SF levels > 1000 μg/L [20, 60, 89]. However, liver biopsy is no longer required to confirm the diagnosis of HH or to quantify iron accumulation, as it has been replaced by the evaluation of biological data and genetic testing [20, 64]. Also, magnetic resonance imaging (MRI) may aid visualizing and quantifying hepatic, pancreatic and splenic excess iron and differentiate between HH related to hepcidin deficiency , like the HFE HH, from FPN disease [20, 60]. Recent studies suggest that the measurement of hepcidin may be relevant to the diagnosis strategy of HFE-associated HH and that, in the future, hepcidin-sensitive tests might be available [20].

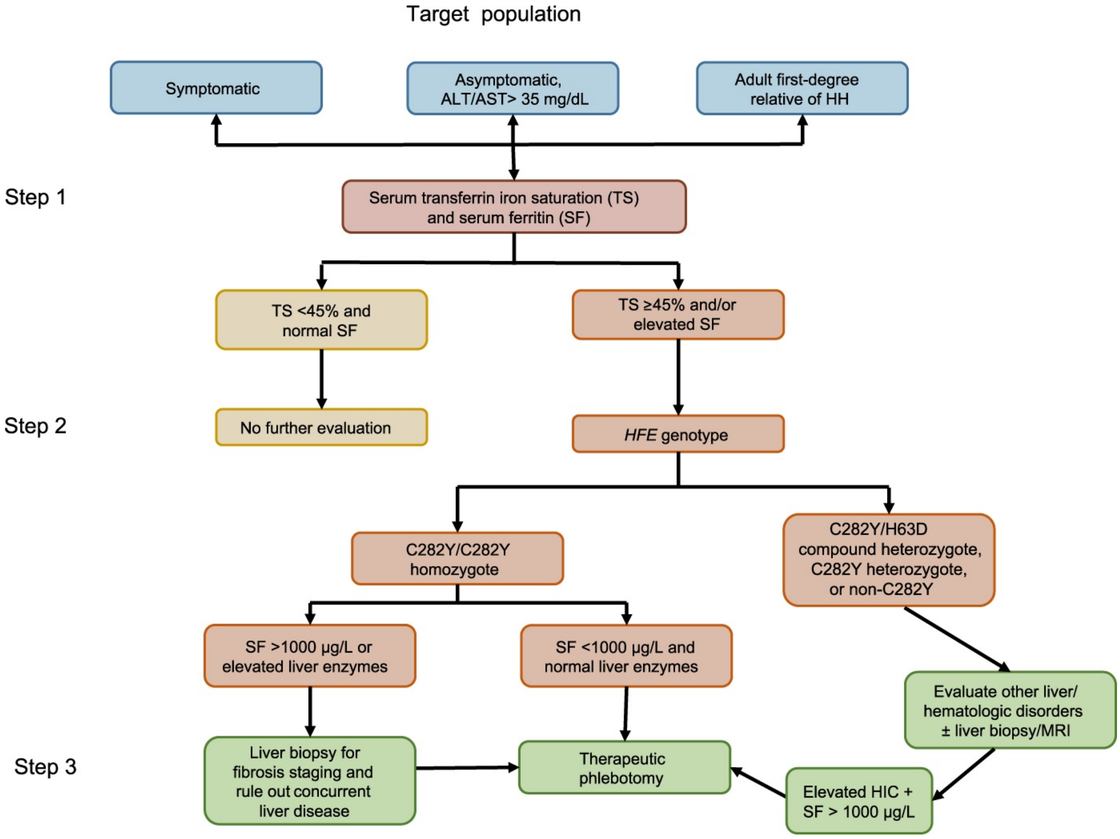

Target population
Symptomatic
Asymptomatic, ALT/AST> 35 mg/dL
Adult first-degree relative of HH
Step 1
Serum transferrin iron saturation (TS) and serum ferritin (SF)
TS <45% and normal SF
TS ≥45% and/or elevated SF
Step 2
No further evaluation
HFE genotype
C282Y/C282Y homozygote
C282Y/H63D compound heterozygote, C282Y heterozygote, or non-C282Y
SF >1000 µg/L or elevated liver enzymes
SF <1000 µg/L and normal liver enzymes
Evaluate other liver/ hematologic disorders ± liver biopsy/MRI
Step 3
Liver biopsy for fibrosis staging and rule out concurrent liver disease
Therapeutic phlebotomy
Elevated HIC + SF > 1000 µg/L

Figure 1.3: Algorithm for HFE HH diagnosis and screening. Adapted from Kris V. Kowdley et al. [60]

1.2.6 Treatment and management

Phlebotomy is the mainstay of treatment for HFE-related HH [20, 60, 84]. The goal of phlebotomy is to remove excess iron to prevent further tissue damage and complications related to iron-overload [20, 79]. Repeated removal of red blood cells reduces the iron in HB, subsequently stimulating erythropoiesis and thereby mobilising iron stored in organs, eliminating excess stored iron [89]. As no randomised controlled trials were assessed yet, venesection therapy is recommended based on clinical evidence that blood withdrawal and, consequently, iron removal before the development of cirrhosis and diabetes is associated with reduced morbidity and mortality [84]. In fact, most experts believe that this form of iron depletion can improve chronic fatigue, cardiac function, hepatic fibrosis and reduce skin pigmentation in patients with HH, although life-threatening conditions like cirrhosis or HCC continue to be a threat to survival even after adequate phlebotomy [10, 20, 84]. Life expectancy of HH patients on phlebotomy therapy is equal to that of the non-HH individuals when the disease is diagnosed before the onset of cirrhosis and diabetes [89]. Also, studies have shown the benefits of phlebotomy through the comparison of the iron depleted between groups of patients treated with phlebotomy, not treated with phlebotomy or inadequately treated with phlebotomy [64]. Yet, adverse effects of venesection therapy may occur in 37%-50% of the patients and include phlebitis, malaise and fatigue [20]. Systematic studies have never been endured to determine the starting point, frequency and ending point of therapeutic phlebotomy [64, 79, 84]. As such, there are no evidence based protocols for phlebotomy treatment [64]. Thus, treatment planning is established based on empirical recommendations for the timing and frequency of venesections, as well as to the level of iron burden at which therapy should be proposed [64, 84].

Still, current guidelines suggest that the treatment should be initiated in C282Y homozygotes with increased TS ($> 45\%$) and SF levels (> 200 μg/L for women and > 300 μg/L for men) [60, 84, 89]. Homozygous patients with SF within normal limits at diagnosis should be monitored by assessing their SF, ALT and AST values [60]. For compound heterozygotes or H63D homozygotes, in case of liver disease and SF > 1000 μg/L, iron depletion therapy may be considered [60]. The treatment consists of two phases, often called depletion phase (DP) and maintenance phase (MP) [89, 90]. While the DP has the goal of lowering SF to target values (50-100 μg/L), the MP aims at maintaining stable target SF values (≈ 50 μg/L) [60, 79, 89].

The initial stage of therapy, the DP, usually consists of weekly removal of 400-500 ml of blood, which contains approximately between 200 and 250 mg of iron and reflects an average of 30 μg/L of depleted SF per phlebotomy [60]. Although, therapy frequency and volume of blood extracted may vary depending on patients tolerability [60, 89]. Also, the periodicity or volume of venesections should be adapted to maintain HB concentrations above 11–12 g/dL [84]. The number of venesection procedures is highly variable and dependent on iron reserve status [89]. Hence, the duration of the DP can differ between patients, ranging from months to years [20, 89]. Indeed, studies have shown mean durations of 1.4 years, with durations ranging from 0.44 to 3.6 years [44, 89].

In the MP, the phlebotomy frequency is reduced to 2-6/year [60, 89]. Although, the periodicity of venesections is highly dependent on the rate of reaccumulation of iron, which varies among individuals [20, 89]. The SF levels should be maintained at $\approx 50\ \mu$g/L, albeit the TS may still be elevated for several patients and consequently NTBI may be present as well [20, 84].

More recently, erythrocytapheresis has been used on some patients [20, 89]. Erythrocytapheresis selectively removes red blood cells and returns the remaining components, such as plasma proteins, clotting factors, and platlets, to the patient [60, 89]. This alternative method allows to remove more red blood cells than phlebotomy (1000 ml vs 200-250 ml) and, also, studies have shown reductions in the number of procedures needed and in the duration of treatment when compared to phlebotomy therapy [89]. Furthermore, this technique can be individualized based on gender, weight, hematocrit and total blood volume [60]. Notwithstanding, erythrocytapheresis is more expensive, less available and a more complex technique than phlebotomy based therapeutics [19, 20].

1.3 Motivation

Starting point, frequency and end point of the first phlebotomy therapy phase, the DP, is recommended based on current guidelines and empirical expertise, as there are not any evidence-based data to support an established protocol [79, 84]. Consequently, the duration of this treatment stage is highly variable among HFE-associated HH patients [20, 89]. Also, the duration of the DP may be greatly related to the degree of iron-overload [20]. Moreover, side effects related to the DP seem to be present for several individuals, which may be another factor adding to the degree of variability of this treatment phase duration. Actually, a study shown that 52% of the patients reported negative experiences related to the treatment, whilst on another study, only 33% of the patients complied with weekly schedules and 43% with every two weeks venesections, during the DP [47, 90]. The same study referred that thirteen patients, out of 118, required more than the average number of phlebotomies to achieve iron depletion and took more than one year to end the DP, opposed to the other patients that took one year or less and a reduced number of venesections [47]. Therefore, having a better understanding of the underlying factors affecting the duration of the DP may be clinically relevant.

In this work, we propose a comprehensive study of some patients characteristics and their influence on the duration of the first stage of therapy of HFE-related patients (statistical inference) and, thereafter, attempt at estimating the duration of the DP for newly diagnosed HFE HH patients, based on previous found insights (prediction). Conceivably, this study may aid HFE-associated HH patients, medical practitioners and medical facilities. Indeed, patients may benefit from an estimation of the duration of the DP, allowing them to have a better perspective of the time and frequency of the phlebotomy therapies and, perhaps, adapt better to the treatment schedule and thus, increase treatment compliance rates overall. Furthermore, treating physicians may gain new insights regarding the factors affecting the duration of the DP, allowing for an enhanced and more personalized treatment planning. Besides, hospitals and medical facilities may indirectly benefit by reducing costs associated with the therapy, given that a more patient-specific treatment planning is possible. In reality, we believe that this study may be a relevant first step on the development of a model based tool to understand the influence of external factors on the initial treatment phase of HH and predict the duration of the DP of HFE-related HH patients.

1.4 Aims

The main goal of this work was to predict the duration of the initial phase of phlebotomy treatment, the DP, of patients newly diagnosed with HFE-related HH. As such, a sequential statistical study was performed to understand the underlying factors affecting the duration of the phlebotomy therapy during this iron depletion phase. Thereafter, by taking advantage of this statistical analysis, predictive models were built, evaluated and compared, aiming at obtaining a regression model to grant HFE-associated HH patients and their treating physicians a reasonably accurate estimation interval for the duration of the DP.

1.5 Thesis outline

A brief explanation of each Chapter of this thesis is provided in this section.

Introduction: Provides a general introduction of the disease studied on this thesis, its pathophysiology, epidemiology, diagnostic and treatment strategies. Besides, the main motivation and the aims of this study are explained.

Materials and Methods: Presents a brief description of the data used throughout the thesis and indicates the software used.

Modeling the Depletion Phase: Addresses the core results of this work, while concomitantly discussing them. Altogether, the topics discussed involve data pre-processing, statistical data analysis and correlation studies, subsequent data processing and finally, regression models building, evaluation and comparison.

Discussion: Serves as a wrap-up after presenting the results, where the main findings are highlighted and discussed. Future perspectives, recommendations and alternative approaches are also suggested.

Conclusion: Summarizes the major findings, providing the final remarks for this work.

Chapter 2

Materials and Methods

2.1 Data description

Data related to 384 HFE-related HH patients was provided by the Department of Metabolic Diseases of the Leuven University Hospital (UZ Leuven, Campus Gasthuisberg) in Belgium. Although patient data is anonymized, as each individual is identified by an arbitrary number, approval from the hospital's Ethical Committee had to be granted due to the data's personal and clinically sensitive information. Professor Dr. David Cassiman, from the referred department of the Leuven University Hospital, and his colleague Dr. Annick Vanclooster were involved in this project, supervising it and giving clinically relevant insights when needed.

The Department of Metabolic Diseases of the Leuven University Hospital supplied, for this master thesis, three data sets. Two of these data sets contain information regarding biochemical parameters over a certain period of time, depending on the individual. While some patients have data from the mid-90's until 2019, others only have data from this millennium. One of these data sets includes time-stamped values of SF (μg/L), TS (%) and AST (U/L), among others (Table A.2). The other data set contains time-stamped values for HB (g/dL) and ALT (U/L) (Table A.3). Every time-stamped observation corresponds to a phlebotomy performed by the patient. The two data sets with time-stamped values have 18 variables related to biochemical parameters. The third data set contains 177 variables that are time-invariant, contrary to the data set with time-stamped biochemical information (Table A.4). While the data sets seem very information-rich, some variables have numerous missing values. This data set has information regarding patients' gender, age when diagnosed, SF levels at diagnosis, body mass index (BMI), smoking habits, alcohol habits, HH genotype, HH-derived complications and other variables. On the data set with static variables, approximately 75% of the values are missing. In addition, time-stamped biochemical data is highly irregularly spaced in time.

2.2 R Programming Language

The R programming language was used on this thesis, as it is widely used for statistical programming (R version 3.5.3) [85]. The integrated development environment used was the R Studio version 1.1.463 [93]. The R packages used on this study are listed below.

- *readxl* [111]
- *ggplot2* [110]
- *knitr* [114]
- *dplyr* [112]
- *plyr* [109]
- *magrittr* [9]
- *rlist* [87]
- *janitor* [33]
- *xts* [95]
- *zoo* [116]
- *flextable* [42]
- *xtable* [26]
- *corrplot* [107]
- *GGally* [97]
- *ggsci* [113]
- *stargazer* [49]
- *VIM* [59]
- *car* [36]
- *FSA* [73]
- *userfriendlyscience* [77]
- *onewaytests* [25]
- *interactions* [65]
- *MASS* [104]
- *olsrr* [45]
- *caret* [37]
- *scorer* [46]

Chapter 3

Modeling the Depletion Phase

This Chapter presents the core results of this dissertation. Key procedures of this study include: i) data processing, in which the data is prepared for further analysis, ii) Exploratory Data Analysis (EDA), with emphasis on correlations between variables and differences between groups, iii) data processing after EDA, to prepare the data for modeling according to gained insights from the previous analysis and iv) predictive models building, comparison and evaluation.

3.1 Data processing

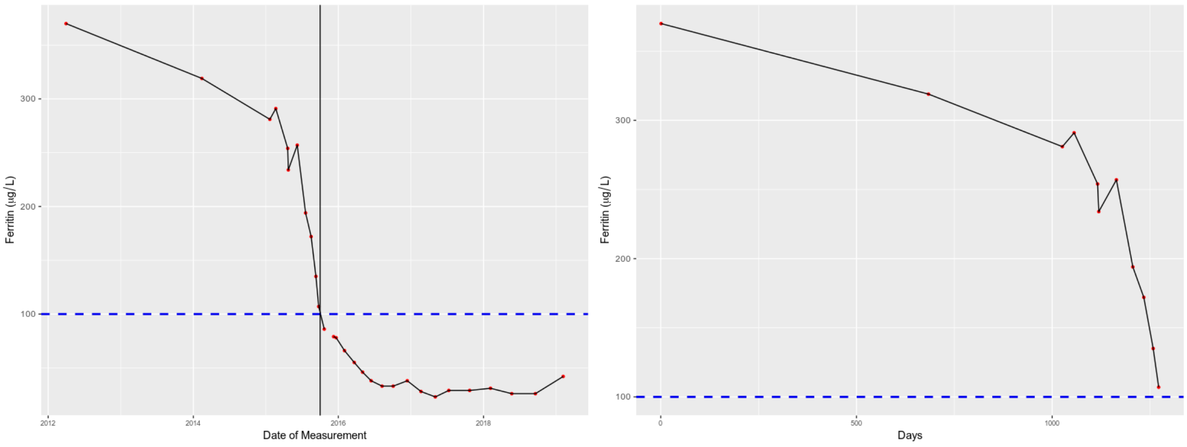
Some data pre-processing steps had to be performed before EDA and modeling. The data sets provided did not include a clear indication of the time the DP took for each patient. As this information is fundamental to the main goal of this work, some processing techniques were performed to create a variable for this effect. Further, understanding and characterizing the behavior of the SF concentration over time allowed to extract some knowledge regarding the DP.

3.1.1 Grouping of patients based on initial SF value

HFE-related HH patients were grouped based on the initial SF value, creating a new variable named **Group**. Three groups were created: Group 0 includes patients with initial SF value $<$ 500 μg/L, Group 500 includes patients with initial SF value $\geq$ 500 μg/L and $\leq$ 1000 μg/L and Group 1000 includes patients with initial SF value $>$ 1000 μg/L. As the SF is used as a marker to assess treatment planning and is directly associated with the duration of the treatment phases, this variable seemed relevant to characterize the patients.

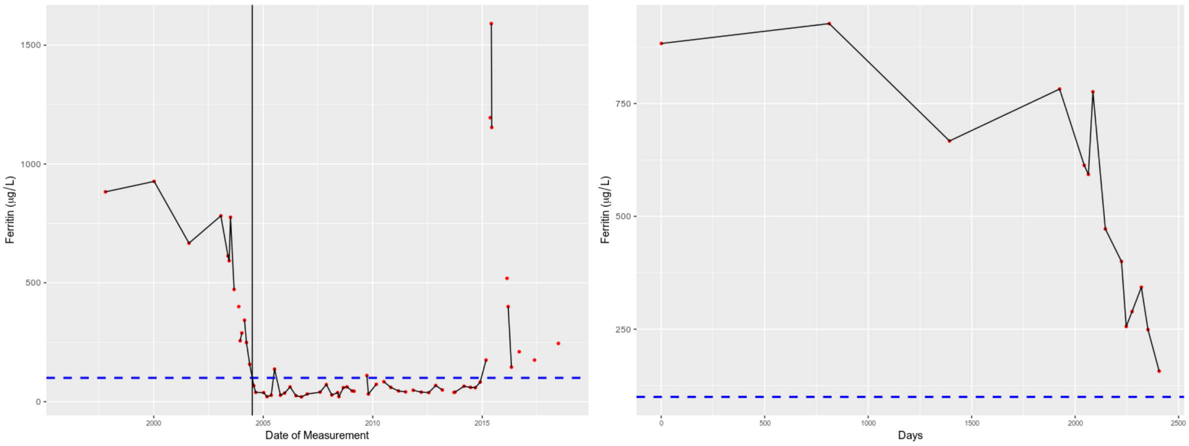
3.1.2 Selection of SF time-stamped values from the DP

The data sets provided did not have information dictating the duration of the DP, its starting and ending date, or whether a time-stamped SF value is associated with that treatment phase or with the MP. So, the SF values that belong to the DP were selected under some assumptions: i) patients with initial SF values < 100 μg/L do not have data regarding the DP and have to be disregarded, ii) patients with initial SF values > 100 μg/L have data regarding the DP, iii) the next time-stamped SF values were assessed iteratively and were assumed as part of the DP if their concentration was > 100 μg/L and iv) the cut-off point to identify the end of the DP is when the SF value being checked is < 100 μg/L, so all the previous time-stamped SF values are included on the DP. Similar assumptions were considered on a previous study [91]. Also, dates (YY/MM/DD) were converted to number of days, being the day 1 the first day of the DP. Figures 3.1, 3.2 and 3.3 show plots with the time-stamped SF values for three arbitrary patients, one of each group of initial SF level, for both the DP and the MP, before any processing, and for the DP only, after the processing explained before. A table with SF values and respective day for one of these patients (from Group 0) can be seen in the Appendix, as illustration (Table A.1).



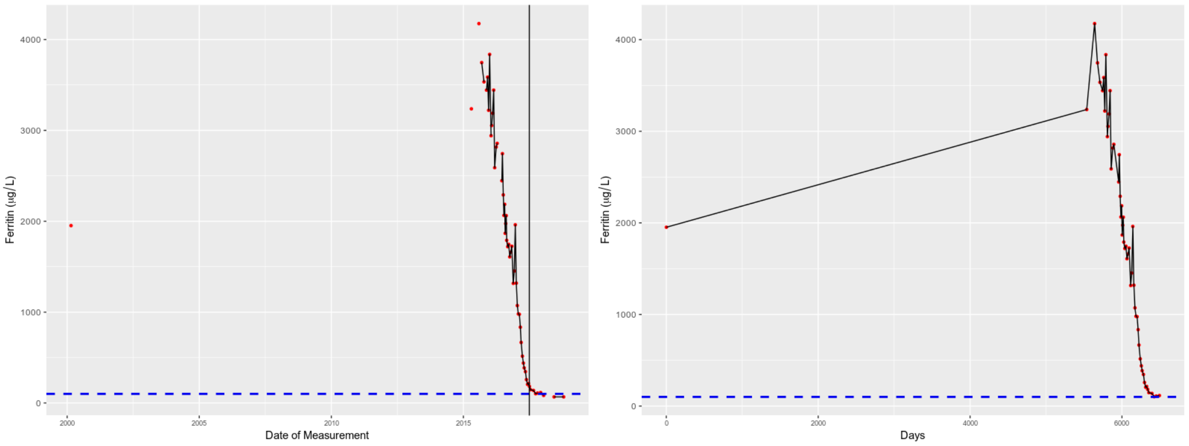
(a) (b)

Figure 3.1: Plot depicting the SF values for a patient from Group 0 (initial SF value < 500 μg/L) before removing data points associated with the MP (a) and after (b). On plot (a), black vertical line indicates approximate end point of the DP. On both plots, blue horizontal dashed line indicates threshold of SF concentration (100 μg/L)



(a) (b)

Figure 3.2: Plot depicting the SF values for a patient from Group 500 (initial SF value ≥ 500 μg/L and ≤ 1000 μg/L) before removing data points associated with the MP (a) and after (b). On plot (a), black vertical line indicates approximate end point of the DP. On both plots, blue horizontal dashed line indicates threshold of SF concentration (100 μg/L)



(a) (b)

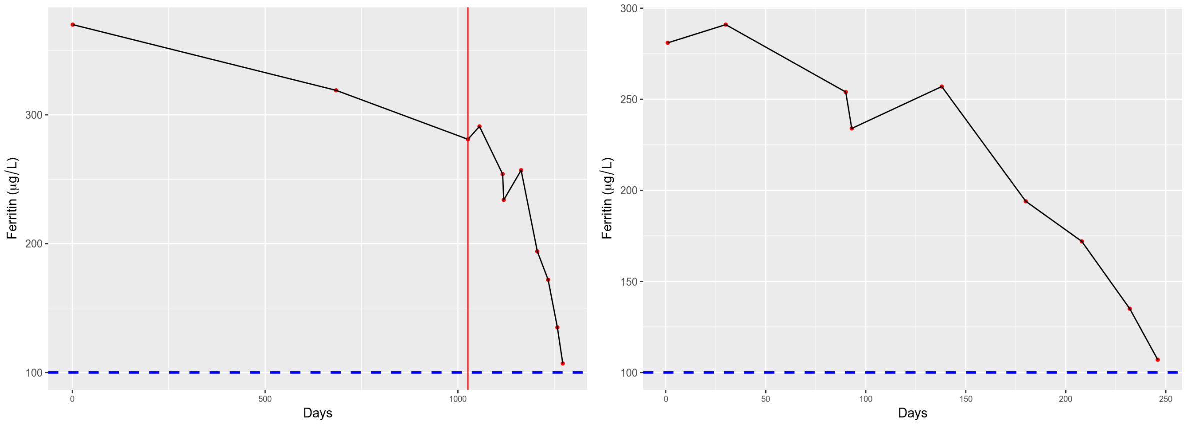
Figure 3.3: Plot depicting the SF values for a patient from Group 1000 (initial SF value > 1000 μg/L) before removing data points associated with the MP (a) and after (b). On plot (a), black vertical line indicates approximate end point of the DP. On both plots, blue horizontal dashed line indicates threshold of SF concentration (100 μg/L)

3.1.3 Retrieval of the initial SF values

The initial SF values were retrieved using a pre-existing static variable originally called **Ferritine (diagnosis)** and renamed to **Ferritin** for simplification. Indeed, **Ferritin** as a static variable representing the initial SF value at diagnosis seemed relevant for further analysis. As this variable has 5 missing values, the data set now contains information regarding 379 patients. Also, patients with SF $<$ 100 μg/L are discarded under the assumption explained before. Eleven patients did not have data concerning the DP according to this assumption, which reduced the data set to 368 patients. Besides, to estimate the duration of the DP, two time-stamped SF values are needed to define a starting and end point. Hence, patients with $<$ 2 time-stamped SF values were removed. This processing step resulted in a major data loss as 77 patients did not have at least two time-stamped observations. At this point, the data set included information regarding 294 patients.

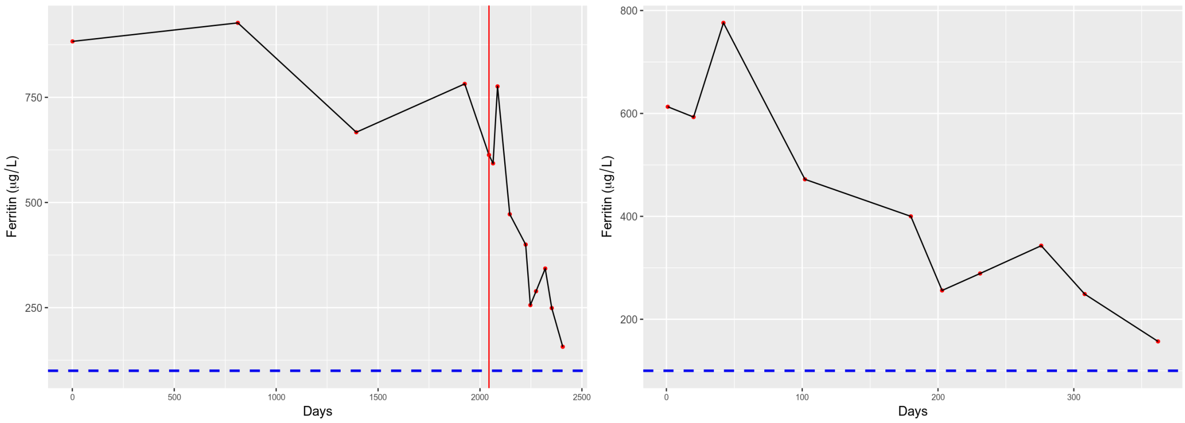
3.1.4 Treatment interruptions during the DP

Large treatment interruptions during the DP may be an indicator that during that time period, a patient did not endure the treatment as planned by the physician. So, to try to approximate the SF time-stamped data to a more ideal behavior in which the patients followed treatment more according to the recommendations given by the physician, treatment interruptions $>$ 100 days were excluded. To achieve this, each patient's time-stamped SF values were checked and if the difference between one observation and the next was $>$ 100 days, all the previous data points were disregarded. Also, the initial SF values (**Ferritin** variable) were updated accordingly. Plots before and after interruptions $>$ 100 days removal, for the same patients as before, can be observed on Figures 3.4, 3.5 and 3.6. After this procedure, patients with $<$ 2 time-stamped SF values were removed. 33 patients had less than 2 observations, which lead to a data set with 261 patients.



(a) (b)

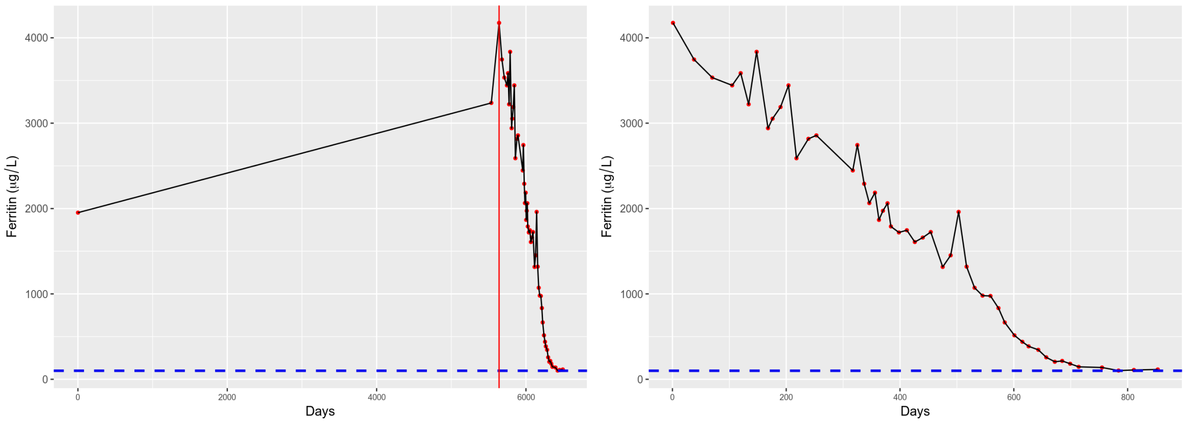
Figure 3.4: Plot depicting the SF values during the DP for a patient from Group 0 (initial SF value $<$ 500 μg/L) before treatment interruptions $>$ 100 days removal (a) and after (b). Red vertical line indicates cut-off point.



(a)

(b)

Figure 3.5: Plot depicting the SF values during the DP for a patient from Group 500 (initial SF value $\geq$ 500 μg/L and $\leq$ 1000 μg/L) before treatment interruptions > 100 days removal (a) and after (b). Red vertical line indicates cut-off point.



(a)

(b)

Figure 3.6: Plot depicting the SF values during the DP for a patient from Group 1000 (initial SF value > 1000 μg/L) before treatment interruptions > 100 days removal (a) and after (b). Red vertical line indicates cut-off point.

3.1.5 The behavior of the decay of SF over time

Considering that the duration of the DP depends on the SF concentration over time, under the assumption explained above, characterizing its behavior may be important to discover any patterns on the data at hand. As the data set has time-stamped values of SF for a considerable number of patients (n = 261), it was not possible to plot and study every single patient individually. Although, by observing plots of SF values during the DP for several random patients, it seemed that: i) the SF decay over time is not linear, ii) the SF decay over time resembles an exponential decay, iii) there is some degree of variability from patient to patient, iv) the phlebotomies do not seem evenly spaced in time, meaning that the treatment plans are not performed as recommended by the medical practitioners and v) there are some fluctuations on the SF values over time, as there are cases where the SF concentration increases after the last phlebotomy (previous observation). In fact, the SF values during the DP seem to have a stochastic behavior in which physiological indeterminacy may play a major role, as other biological processes may influence the measurement of interest. Figure 3.7 shows the curves of SF values during the DP for six randomly chosen patients. On this section, the SF concentration time-series is handled and analyzed as an exponential decay curve.

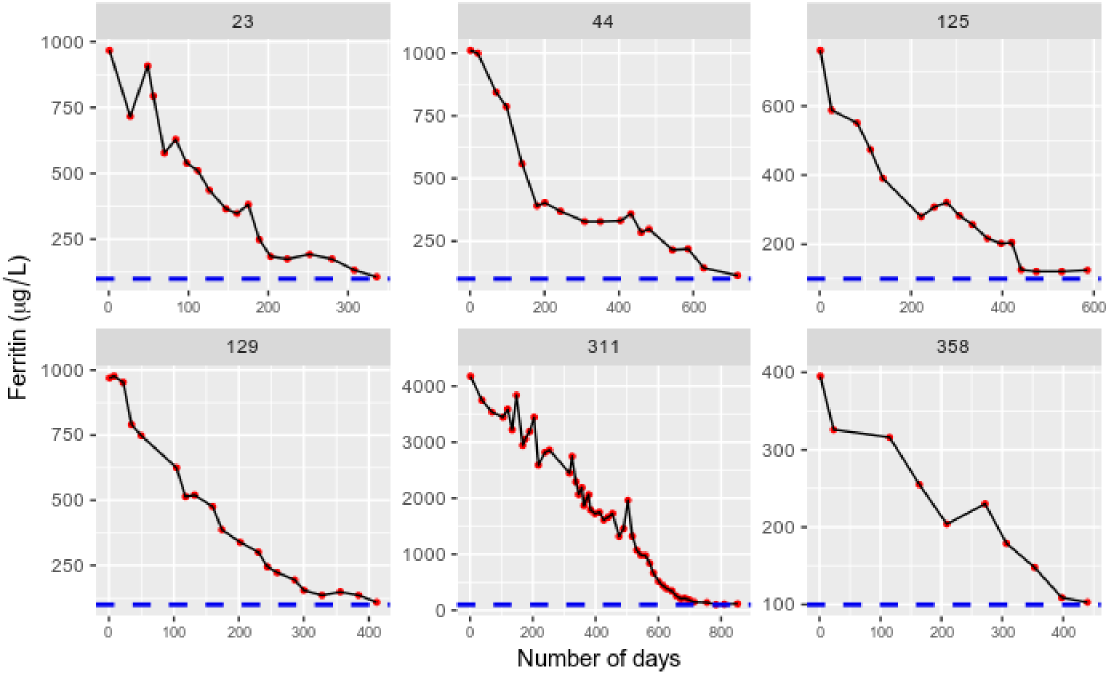


Figure 3.7: Plots depicting the SF values during the DP for 6 patients

3.1.6 Nonlinear Least Squares analysis

To further explore the hypothesis that the SF values over time, during the DP, have an exponential decay, a method called Nonlinear Least Squares (NLS) was used to fit curves to each patient time-series. Characterizing the underlying patterns behind the depletion of SF during the DP may enhance the understanding of this first phase of treatment of HFE-related HH [27]. The NLS method tries to find the optimal parameters for an equation that describes a set of data points, X_i and Y_i. The algorithm is based on iteratively finding the best guess for the parameters, given initial guesses [51]. As stated above, the NLS method requires the specification of an equation. The formula chosen is based on First Order Kinetics, which describe a monoexponential decay [11, 94]. So, the equation specified on this NLS analysis is:

$$SF = A \cdot e^{k \cdot Days} + T$$

where:

- **SF** refers to the SF concentration at a specified day
- **A** refers to the initial SF value
- **k** refers to a parameter to be estimated by NLS
- **Days** refers to the number of days
- **T** refers to the last SF value, approximated to 100 μg/L for all the patients

On the proposed equation, **A** and **Days** are known parameters specific to each patient and **T** is constant to all the individuals, taking into account the assumption that the DP ends when the SF concentration is close to 100 μg/L. The parameter **k** is the only value to be estimated by the NLS method and may help characterize and differentiate each patient's SF curve behavior during the DP. As the **k** parameter is closer to zero, the decay speed is slower and more negative values are associated with higher speed decays. Three exponential decays with this equation, each with a different **k** parameter value, are illustrated in Figure 3.8. Also, on Figure 3.9, it is possible to observe the result of the NLS method on three patients from each group of initial SF value. The plots with fixed scales allow to visually conclude that more negative **k** parameters are associated with higher speed decays.

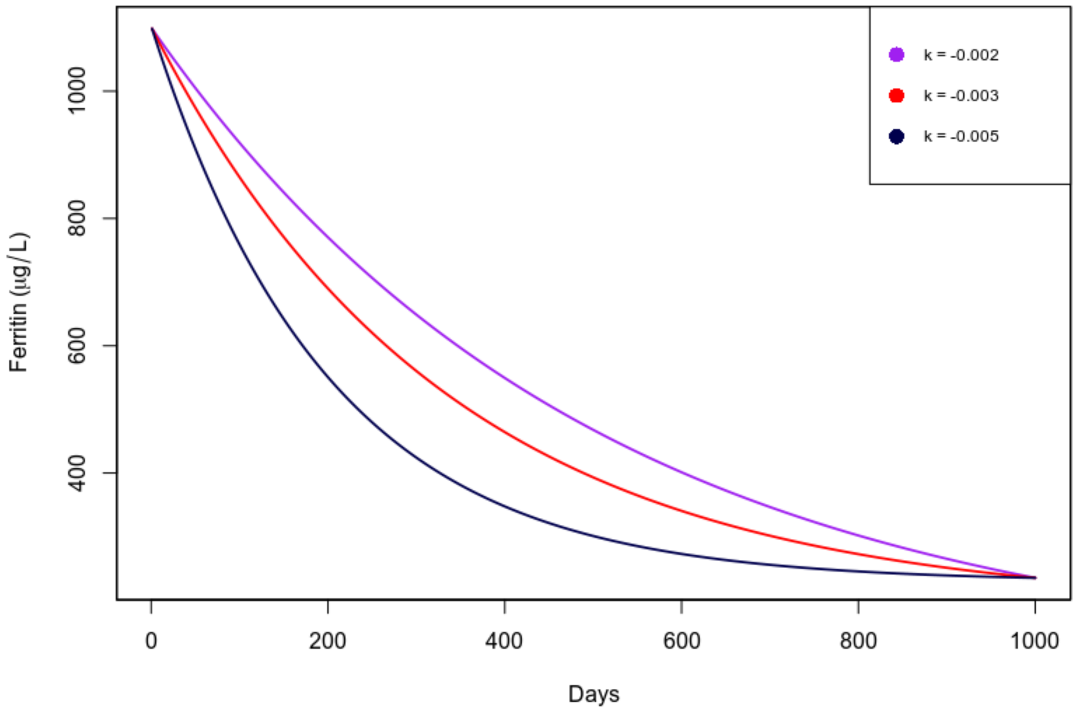


Figure 3.8: Examples of monoexponential decays with the equation specified on the NLS analysis varying only the k parameter value. For all the lines, A = 1000 and Days = 1000

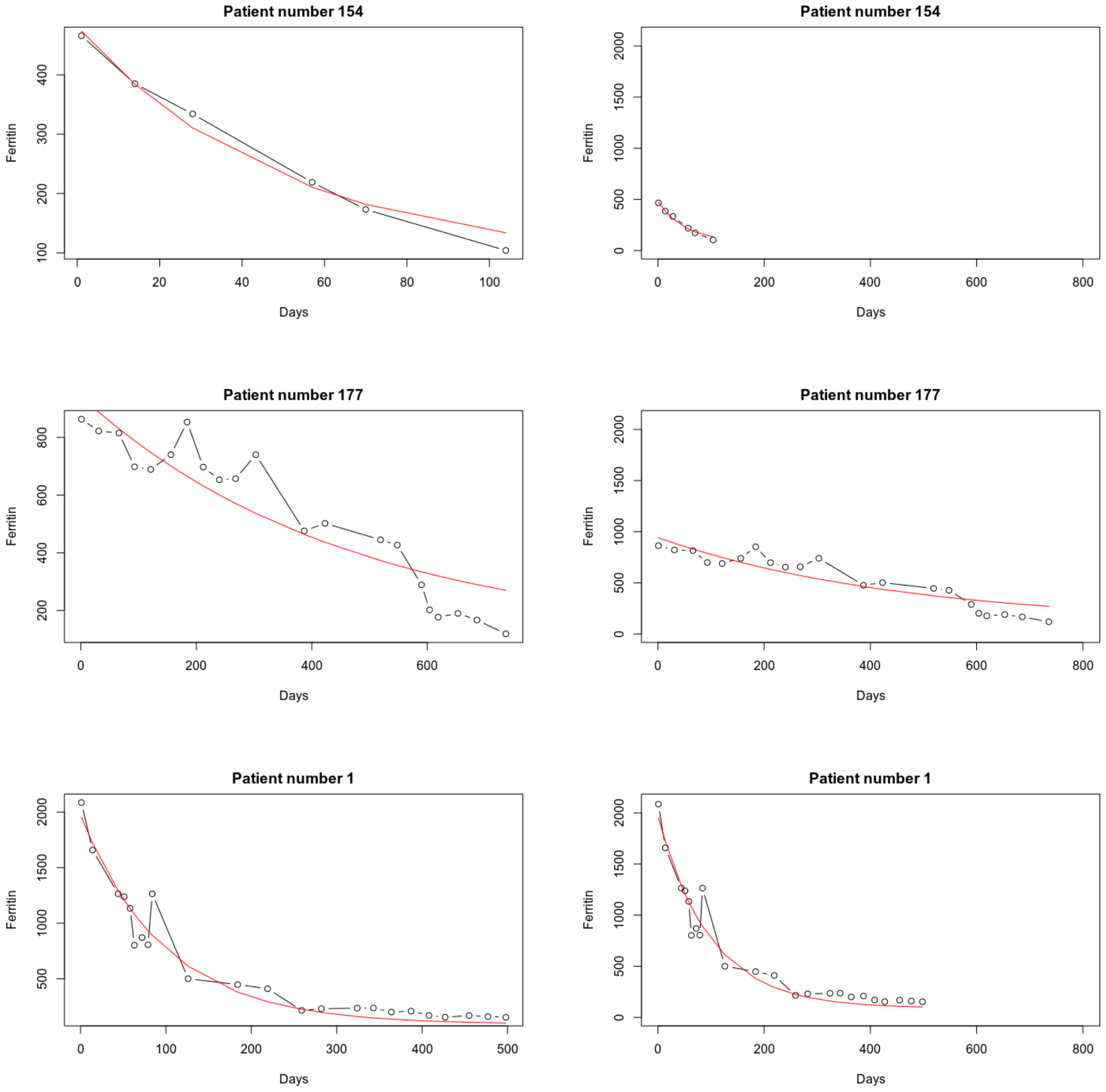


Figure 3.9: Plots with NLS fitted curves for three patients, one of each initial SF group. Patient number 154 is from Group 0 (k = -0.0202), patient number 177 is from Group 500 (k = -0.00216) and patient number 1 is from Group 1000 (k = -0.0102). Plots on the left are scale-free. Plots on the right have the same scale.

3.1.7 Retrieval of k parameters

After assessing the **k** parameter potential to characterize the speed of decay of SF for a given patient, it seemed relevant to retrieve these parameters for all the patients on the data set. Thus, a variable was created with the values of these parameters for all the patients. Although, to fit an exponential curve and estimate the **k** parameter, a patient needs at least three time-stamped SF values. Hence, patients with less than three observations during the DP were excluded (n = 15), which reduced the data set to 246 patients. Figure 3.10 shows a flowchart with the pre-processing steps discussed until now and the number of patients discarded in each step.

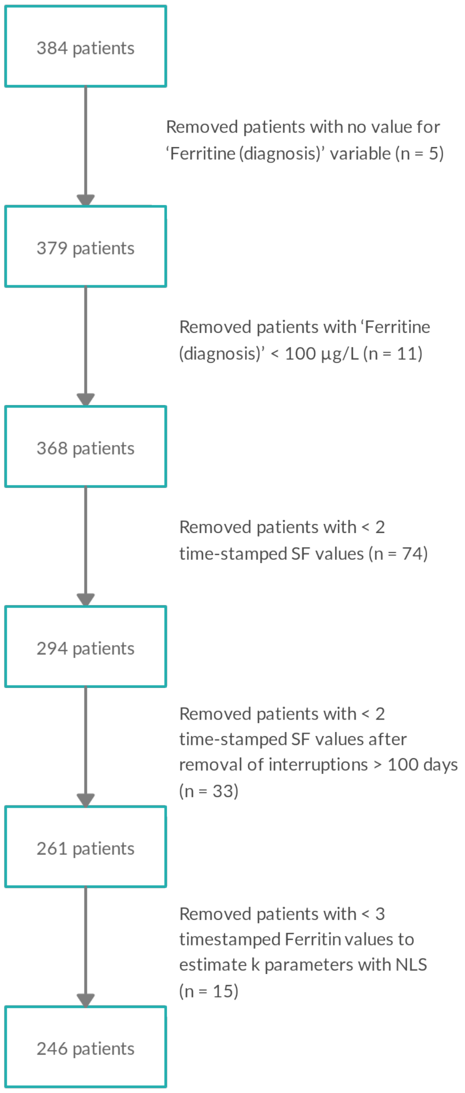

384 patients
Removed patients with no value for 'Ferritine (diagnosis)' variable (n = 5)
379 patients
Removed patients with 'Ferritine (diagnosis)' < 100 µg/L (n = 11)
368 patients
Removed patients with < 2 time-stamped SF values (n = 74)
294 patients
Removed patients with < 2 time-stamped SF values after removal of interruptions > 100 days (n = 33)
261 patients
Removed patients with < 3 timestamped Ferritin values to estimate k parameters with NLS (n = 15)
246 patients

Figure 3.10: Flowchart of pre-processing steps until the retrieval of the k parameters

3.1.8 The outcome variable: duration of the DP in weeks

Recommended frequencies to perform phlebotomies during the DP are usually weekly, every two weeks or monthly depending on the physicians' decision [1, 60, 79]. Therefore, the duration of the DP of all the patients was converted from number of days to number of weeks. This new variable, named **DurationWeeks**, is considered the outcome variable in the ensuing analysis. Table 3.1 shows the appearance of the data set after the pre-processing steps carried until this point, where **Number** refers to the patient number, **Ferritin** refers to the initial SF level, **Group** refers to the group of initial SF value, **DurationWeeks** corresponds to the duration of the DP in weeks and **kParameter** to the parameter k estimated with the NLS analysis.

Table 3.1: Appearance of the data set after pre-processing related to the retrieval of initial SF values, the duration of the DP and the k parameters from NLS analysis

Number	Ferritin	Group	DurationWeeks	kParameter
1	2085	Group 1000	71	-0.0102
2	1125	Group 1000	31	-0.0088
3	956	Group 500	18	-0.0097

3.1.9 Variables selection

The three data sets have, combined, a large number of variables. Thus, two main criteria were established to identify and select the variables of interest for further analysis: i) variables with clinical relevance and ii) variables with few missing values. Also, as the main goal is to predict the duration of the DP for patients recently diagnosed with HFE-related HH, variables with values known before the start of the treatment were prioritized. Moreover, biochemical parameters have time-stamped data which difficult the analysis of these values either to: i) not knowing the behaviour of these values during the DP and ii) having a large number of missing values unlike the SF values. In fact, to fully understand the potential relevance of these variables to the study at hand, a similar analysis to the one performed to the SF levels over time had to be assessed. This section covers some preliminary analysis done on variables that seemed clinically relevant. Although, some were not taken into account for the main analysis of this work as they do not satisfy some of the criteria mentioned before. Chosen variables were added to the current data set (Table 3.1).

3.1.9.1 Liver Disease and Alcohol Intake

Fibrosis and cirrhosis are two potential complications associated with HFE-related HH patients, especially the ones with higher SF levels of more than 1000 μg/L at diagnosis [60]. Besides, alcohol consumption seems to be associated with an increasing risk of developing cirrhosis [15]. In addition, alcohol may worsen iron overload [60, 92]. Both these factors may contribute to a prolonged DP as they seem associated with more elevated SF levels. So, these variables were selected for a preliminary analysis to verify if they fulfill the requirements discussed previously. The **LiverDisease** variable was created from two pre-existing variables called **Fibrosis** and **Cirrhosis**. Patients that do not have neither **Fibrosis** or **Cirrhosis** were given the level **0**, patients with **Fibrosis** were encoded to **1** and patients with **Cirrhosis** to **2** (Table 3.2). The number of missing values (n = 80) and the sample size inequality were two decisive factors to dismiss this variable for further analysis. In reality, the number of patients with cirrhosis (n = 14) seemed too low to assess potential associations with either higher initial SF levels or longer DP durations.

Table 3.2: Number of missing values and number of patients of each level of the LiverDisease variable. Level 0 refers to patients with no known liver condition, level 1 refers to patients with Fibrosis and level 2 refers to patients with Cirrhosis

LiverDisease	n
NA	80
0	110
1	42
2	14

The **Alcohol intake** variable had, initially, thirteen different levels. Besides having a large number of factors, most of them were represented by few patients. This excessive categorization may be associated with wrongfully collected data and is an obstacle for future analysis. As such, a simplification was made where the patients were encoded either to the value **0** if they do not drink or drink sporadically or **1** if they drink frequently (at least weekly). Although the two encoded levels have similar sample sizes, they already endured a processing step that may bias the analysis negatively (Table 3.3). Furthermore, the number of missing values (n = 117) seemed excessively high to consider including this variable on the data set for the next analysis.

Table 3.3: Number of missing values and number of patients of each level of the Alcohol intake variable. Level 0 refers to patients that do not consume alcohol or consume only sporadically and level 1 refers to patients that consume alcohol frequently

Alcohol	n
NA	117
0	53
1	76

3.1.9.2 BMI and Age at diagnosis

Previous research has shown that the **BMI** is associated with increased SF levels in adolescents, especially among obese individuals [99]. Moreover, higher **BMI** seems to be associated with increased risk of diabetes in individuals with HFE HH [13] and higher SF levels seem to be a reliable marker of inflammation among obese individuals [53]. Although the **BMI** seems clinically relevant, it was not included at this point on the data set due to the number of missing values (n = 64). Age at diagnosis, encoded as **AgeDiagnosis**, was another variable considered for a preliminary analysis as it seemed clinically pertinent. Studies have shown that greater age at diagnosis is associated with increased risk of developing cirrhosis and HCC [15, 71]. Additionally, higher age at diagnosis seems to be an indicator for the duration of exposure to iron overload as the progressive accumulation of iron is not counteracted by iron depletion therapies during longer periods compared to inferior ages at diagnosis [71]. Indeed, age seems to be positively associated with increased SF levels [61]. This variable was added to the data set for further analysis as it does not have any missing values and has an apparent clinical importance.

3.1.9.3 Sex

The sex of a patient was another factor assessed. Past research has shown that male p.C282Y homozygous HH patients seem to be more associated with increased SF levels [61]. For female p.C282Y homozygous HH patients, a study suggested that they have higher SF levels after menopause when compared to pre-menopausal females [106]. The **Sex** variable was incorporated on the data set at hand due to its possible influence on the SF levels of HFE-related HH patients and due to the fact that it does not have any missing value.

3.1.9.4 Genotype

SF levels were found to be more associated with the C282Y homozygous genotype in recent studies [5, 80]. Also, a meta-analysis based study identified that homozygous C282Y/C282Y genotype was 60-times stronger associated with TS values greater than 55% and elevated SF levels than compound heterozygous genotype and 100-times stronger than H63D/H63D genotype [70]. Thus, understanding the influence of the genotype on iron overload and the duration of the DP may be of utmost relevance to the present study. The initial **Genotype** variable contained three levels when accounting for the data set after the pre-processing explained on the previous section (n = 246): i) C282Y/C282Y (n = 195), ii) C282Y/H63D (n = 45) and iii) C282Y/S65C (n = 6). As the homozygous C282Y genotype is highly overrepresented and the heterozygous C282/S65C has few cases, this variables was encoded to either being *Homozygous* (n = 195) or *Heterozygous* (n = 51). Even after this encoding, the homozygous genotype is highly more represented than the heterozygous group. Nevertheless, the variable **Genotype** was included on the data set due to its clinical relevance and the fact that it does not contain missing values.

3.1.9.5 Time-stamped biochemical parameters

Besides the SF levels studied before, some time-stamped biochemical parameters were taken into account as a first preliminary analysis. The **TS** levels were deemed of interest as they are a marker, along with the SF concentration, to assess iron stores and disease severity. Indeed, TS greater than 45% identifies more than 97.9% of C282Y homozygotes and may be used as a metric to determine the start of the treatment [60]. **AST** and **ALT** were also considered at first since they are traditional markers for liver disease or excessive alcohol consumption [23, 72, 98]. In fact, serum levels of these two biomarkers are increased, to some extent, in case of liver disease [43]. Additionally, the **HB** variable was also recognized as a potential candidate to include in the data set as it is regularly used to assess anemia in patients undergoing frequent phlebotomy therapy [18, 60]. Regardless, these variables were not included on the data set for the main analysis due to: i) having missing values which would reduce the number of patients and ii) the added complexity of using time-stamped variables.

As the main goal of this work is to study which parameters may be associated with the duration of the DP from data available when a HFE-related HH patient is diagnosed, the inclusion of these four variables was attempted by creating variables that contain the initial value, at diagnosis, of each biochemical parameter, similarly to what was done with the initial SF value. Adding these 4 variables reduced the data set from 246 to 175 patients, meaning that 71 patients have at least one missing value on the initial values for these parameters (at diagnosis). Individually, **AST** had 51 missing values, **ALT** had 49, **HB** had 39 and **TS** had 17 missing values. Plus, these parameters' time-stamped values during the DP were not studied like the SF values over time. Identifying and characterizing the behavior of these time-stamped values could be essential to understand their impact on the duration of the therapy. For these two reasons, these time-stamped values were not incorporated on the main analysis of this work.

3.1.10 Feature extraction

This subsection covers some feature extraction techniques regarding two newly created variables named **Gradient** and **MonthlyPhlebomoties**. While **Gradient** is a variable associated with the SF values over time, like the **k** parameter discussed before, the **MonthlyPhlebotomies** is a variable related to the number of phlebotomies performed by each patient per month.

3.1.10.1 Gradient

Opposed to the **k** parameter extracted before, that tries to characterize a monoexponential decay, the **Gradient** is a linear approximation of the same phenomenon. The idea behind this variable was to model the decay of SF during the DP in a linear and thus, simplified way. The equation used to derive this linear approximation is given by:

$$Gradient = \frac{T - A}{DurationWeeks}$$

Where:

- **Gradient** refers to the gradient of a given patient
- **A** refers to the initial SF value
- **T** refers to the last SF value, approximated to 100 μg/L for all the patients
- **DurationWeeks** refers to the number of weeks of the DP

This equation was used to retrieve the **Gradient** values for all the patients currently included on the data set (n = 246). The output of this equation, **Gradient**, can be considered an approximation of the mean rate of depletion of SF per week for a given patient. As an example, an arbitrary patient with **A** = 1100 μg/L and **DurationWeeks** = 20 would deplete, on average, per week, approximately 50 μg/L of SF. More negative **Gradient** values reflect higher speeds of decay of SF. Differentiating patients based on their ability to deplete greater SF per week during the DP may bring relevant insights when trying to predict the duration of this treatment phase.

3.1.10.2 Monthly phlebotomies

Another variable created, and named **MonthlyPhlebotomies**, is related to the frequency of phlebotomies performed. The duration of DP may be directly associated with the number of phlebotomies performed by a patient, under the assumption that more phlebotomies are related to more iron expelled. As such, the variable created refers to the average number of phlebotomies performed by a patient per month. The formula used to extract this information for all the patients is presented below:

$$MonthlyPhlebotomies = \frac{NumberCases \cdot 30}{Duration}$$

where:

- **MonthlyPhlebotomies** refers to the number of phlebotomies performed per month
- **NumberCases** refers to the total number of phlebotomies performed
- **Duration** refers to the number of days of the DP

Initially, the rounded results of the equation above separated the patients in four groups based on the number of phlebotomies done per month: i) one phlebotomy (n = 135), ii) two phlebotomies (n = 90), iii) three phlebotomies (n = 18) and iv) four phlebotomies (n = 3). As the sample sizes were very distinct and could burden future analysis, this variable was encoded in two levels: i) patients that performed approximately one phlebotomy (n = 135) and ii) patients that performed two or more phlebotomies (n = 114). This simplification may bias further analysis but has the advantage of dividing the patients in two samples of approximate sizes.

3.2 Exploratory Data Analysis

After the pre-processing steps endured and the selection of variables of interest, the data set has information regarding 246 patients and its appearance is shown in Table 3.4. This section focuses on studying the distributions and correlations of the variables considered on this data set, with an increased emphasis on the target variable, **DurationWeeks**. Variables were studied, one by one, in case they were deemed as relevant after initially assessing correlations between the variables. Note that a confidence of 95% ($\alpha = 0.05$) was chosen for all the statistical tests performed.

Table 3.4: Appearance of the data set after pre-processing steps and variable selection

Number	Ferritin	Group	kParameter	Sex	AgeDiagnosis	Genotype	MonthlyPhlebotomies	DurationWeeks	Gradient
1	2085	Group 1000	-0.0102	m	39	Homozygous	1	71	-27.90
2	1125	Group 1000	-0.0088	m	46	Homozygous	>1	31	-32.91
3	956	Group 500	-0.0097	m	52	Homozygous	>1	18	-47.18

3.2.1 Variables' Correlations

Correlation plots were built to study possible associations between variables. A correlation plot with a color gradient scale is presented in Figure 3.11 and Figure A.1, in the Appendix, shows scatter plots and associated correlation values. Categorical variables were encoded to numerical to be able to plot these correlograms. **Sex** values were encoded to 0 for *male* patients and 1 for *females*. **Group** of initial SF level was encoded to 0 if the patient is from *Group 0*, 1 if the patient is from *Group 500* and 2 if the patient is from *Group 1000*. For the **Genotype** variable, *Homozygous* patients were encoded to the value 0 and *Heterozygous* to the value 1. Further, the **MonthlyPhlebotomies** values were encoded to 0 if the patients did one phlebotomy per month or encoded to 1 if the patients performed more than one monthly phlebotomy.

The correlation values between each variable and the target variable, **DurationWeeks**, are summarized in Table 3.5. It is possible to observe that the **kParameter** variable seems to be the one more positively associated with the outcome variable (r = 0.533), meaning that values of **k** closer to zero may be more associated with longer durations of the DP. This correlation value seems to be in line with what was expected taking into account that more negative **k** values describe higher speeds of decay of SF. The **Gradient**, as it is another approximation of the decay of SF over time, has also a positive correlation value (r = 0.325) but does not seem as strongly correlated. Similarly to the **k** parameter exponential approximation, closer to zero **Gradient** values seem to be more associated with longer durations of the DP. The **Ferritin** variable, that has the initial SF level (at diagnosis), and the **Group** of initial SF value variable, also seem positively correlated with **DurationWeeks** (r = 0.497 and r = 0.417 respectively), meaning that more elevated SF concentrations at diagnosis may be more associated with longer DP's durations. Also, a mild positive correlation was found between the **AgeDiagnosis** variable and the target variable (r = 0.169). **MonthlyPhlebotomies** is negatively associated with the duration in weeks of the DP (r = -0.348), which may mean that patients that only perform one phlebotomy per month are more associated with longer treatment periods. Regarding the **Genotype** variable, *Homozygous* HH patients seem to be more associated with longer durations of the DP due to the negative correlation value found between the variables (r = -0.268). Essentially, no associations were detected for the **Sex** variable in respect to the duration of the DP (r = -0.0123).

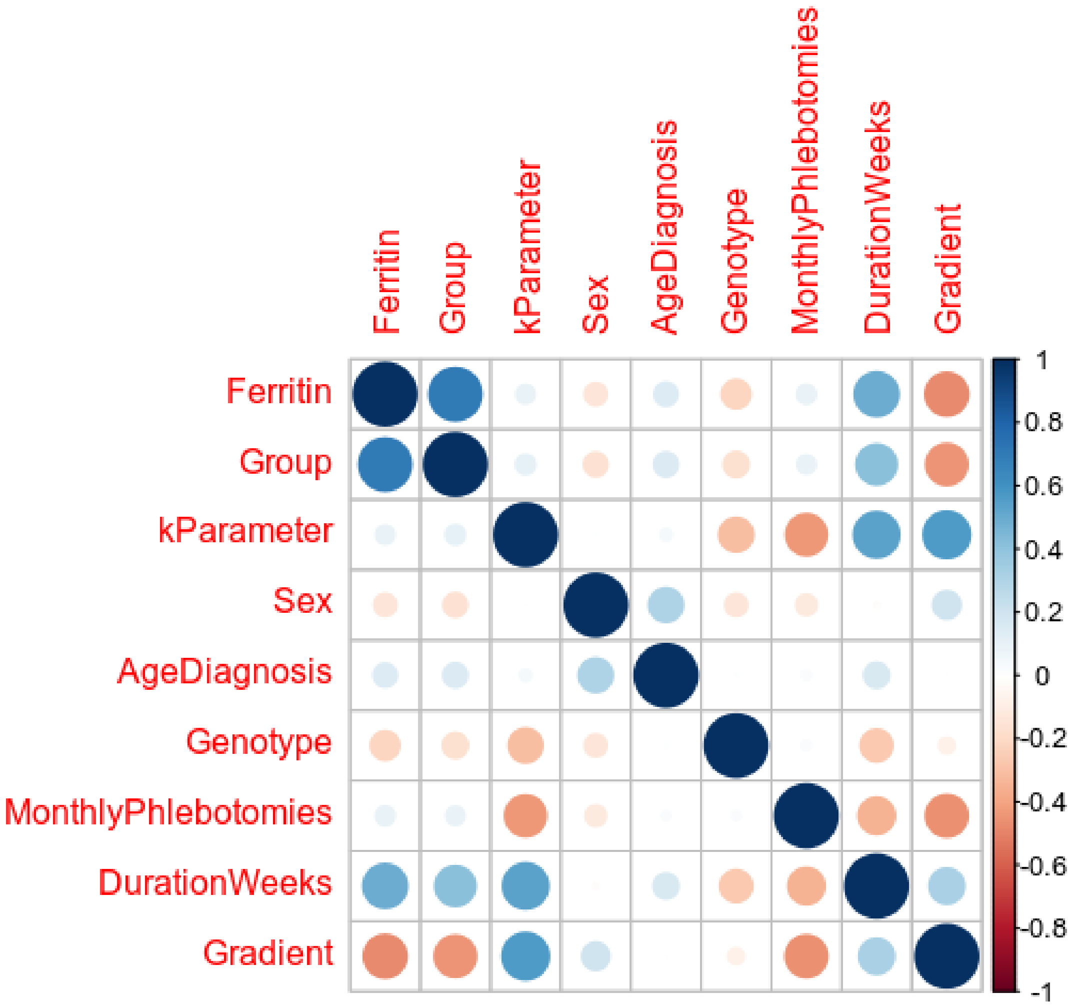


Figure 3.11: Correlation plot with coloured scale where red implies negative correlation and blue colors imply positive correlations.

Table 3.5: Table depicting correlation values between the target variable, DurationWeeks, and all the other variables included on the data set

Positive	Negative
kParameter (0.533)	MonthlyPhlebotomies (-0.348)
Ferritin (0.497)	Genotype (-0.268)
Group (0.417)	Sex (-0.0123)
Gradient (0.325)	-
AgeDiagnosis (0.169)	-

Other correlations were found between the other variables. **Ferritin** seems to have a negative correlation with **Gradient** (r = -0.474), but only a slight correlation with the **kParameter** (r = 0.0935). **Ferritin** and **Genotype** were also found to be negatively correlated (r = -0.218). Correlations between **Ferritin** and the other variables seem very mild or non-existent, given their correlation values. As expected, the **Group** variable has a similar pattern of correlations to the ones found with **Ferritin**. For the case of **Sex**, the major correlation found was with **AgeDiagnosis** (r = 0.302) followed by a slightly weaker correlation with **Gradient** (0.201). So, *female* patients seem to be associated with greater ages at diagnosis and with closer to 0 **Gradient** values. No other significant correlations seem to be found for **Sex** and **AgeDiagnosis**. Furthermore, a negative correlation was found between **Genotype** and **kParameter** (r = -0.301), meaning that *Homozygous* patients may be more associated with more closer to 0 **k** values and consequently slower speed decays of SF. On the other hand, no correlation was found between **Genotype** and **Gradient**. Negative correlations were found between **MonthlyPhlebotomies** and **Gradient** (r = -0.453) and between **MonthlyPhlebotomies** and **kParamater** (r = -0.436), which may imply that patients that only do one monthly phlebotomy are more associated with slower decays of SF. Finally, the variables **Gradient** and **kParameter** have a positive correlation (r = 0.561), which was expected as both try to approximate the same physical phenomenon. Whilst the majority of the correlations found do not seem to be very strong, they may serve as a starting point to understand the underlying associations of these variables.

3.2.2 DurationWeeks variable

Firstly, it was essential to study the distribution, calculate basic statistics and determine associations with other variables of the outcome variable. Table 3.6 provides some statistics regarding this variable. The range of the values (204) seems to be very high when compared to the mean (44.17) and median values (36.50), which may lead to suspect that the distribution has a positive tail. In addition, the skew (1.58) and kurtosis (4.09) values, which measure the degree of asymmetry of probability distribution, seem to indicate some degree of positive skewness (right tail). At this point, these may be indicators that the distribution of the **DurationWeeks** is deviated from a normal one, to some extent.

Table 3.6: Basic statistics for the DurationWeeks variable

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
DurationWeeks	1.00	246.00	44.17	30.05	36.50	40.33	25.20	6.00	210.00	204.00	1.58	4.09	1.92

Likewise, the histogram presented on Figure 3.12 seems to corroborate the idea that the **DurationWeeks** has a non-normal distribution with a right tail (positive skewness). A Shapiro-Wilk normality test, in which the null hypothesis rejection indicates that the data may derive from a non-Gaussian distribution, was also used to help confirm data non-normality [86]. This test also seems to prove that the data is not normally distributed (p-value = 5.47×10^{-13}).

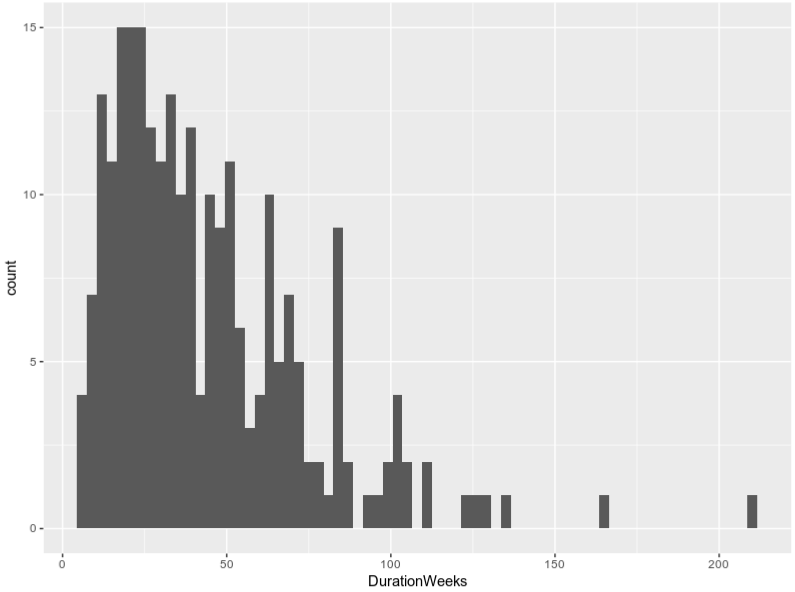


Figure 3.12: Histogram of DurationWeeks values

3.2.2.1 Association with Group of initial SF level

After initial correlation studies where a positive association between **DurationWeeks**, **Ferritin** and **Group** of initial SF value was found, further analysis were performed to check this association. The **Group** categorical variable was used to test associations with initial SF levels. Table A.5 shows the median values for all the numerical variables based on the **Group** of initial SF level. The histogram of the **DurationWeeks** values based on **Group**, on Figure 3.13, seem to show slightly different distributions of the duration of the DP depending on the initial SF concentration. Also, the histogram seems to show that each **Group** has different sizes. Table 3.7 displays the sample sizes, where it is possible to observe that **Group 500** has almost twice the number of patients of **Group 0**. The box plot on Figure 3.14 also seems to show differences on the distribution depending on the **Group**, with increasing medians for increasing initial SF value and more extreme values, associated with the tail of the distribution, with the same pattern.

Data heteroscedasticity was also assessed using the Levene's test for variance homogeneity.

Heteroscedasticity occurs when sub-groups of the data have different variances depending on one or more exploratory variables, time or spacial ordering [24]. The Levene's test is appropriate to check data's variance homogeneity as its null hypothesis is that every sample assessed has the same variance [41]. Results from this test seem to show that some degree of data variance heterogeneity is present (p-value $= 5.12 \times 10^{-4}$).

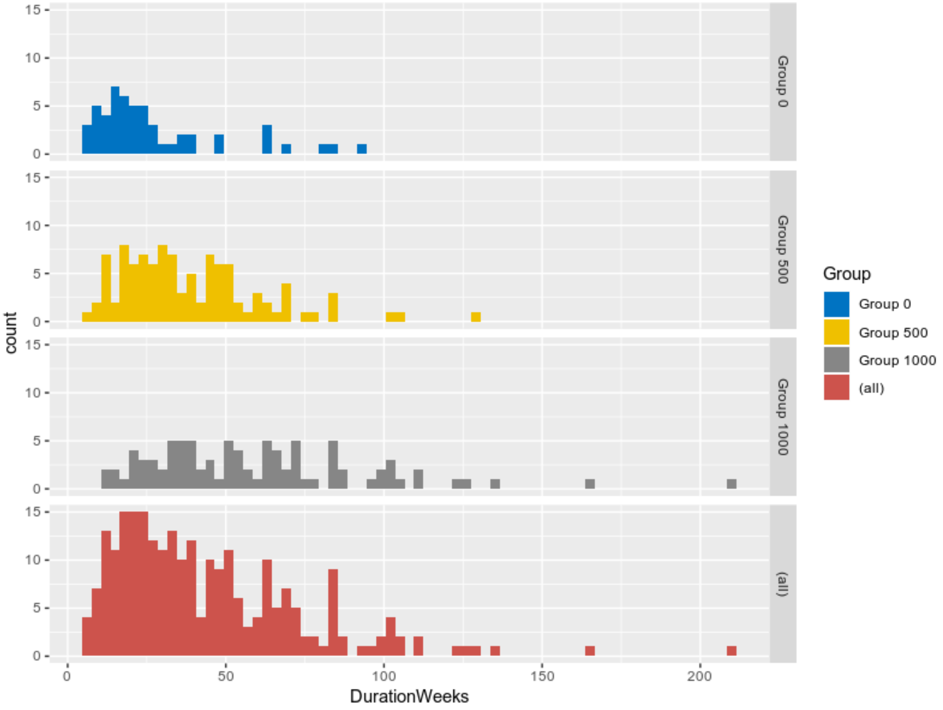


Figure 3.13: Histogram of DurationWeeks values based on Group of initial SF level

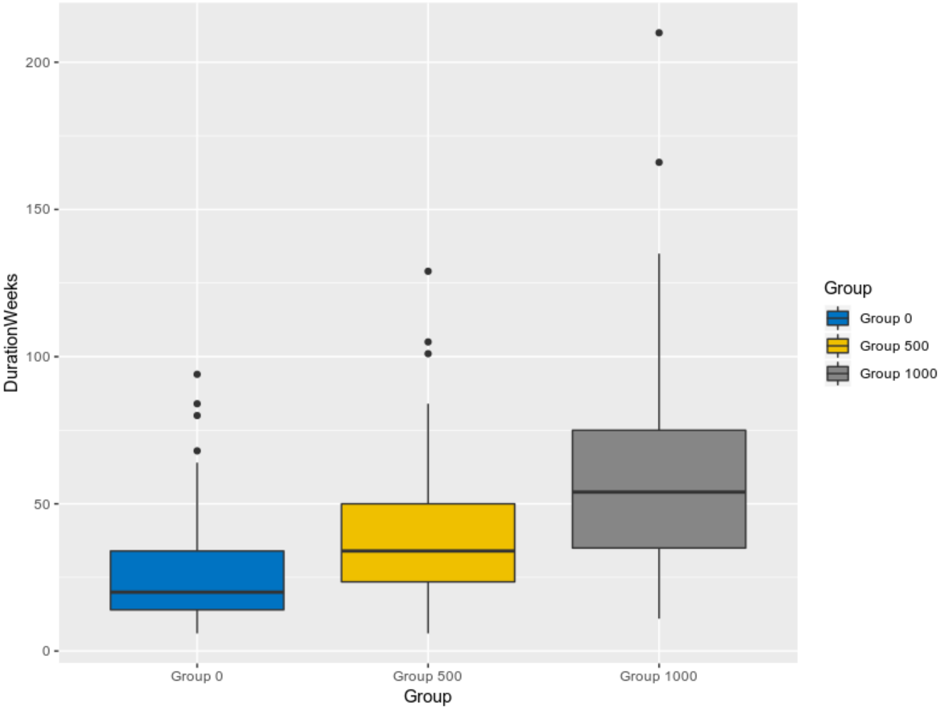


Figure 3.14: Box plot of DurationWeeks values based on Group of initial SF level

Table 3.7: Sample sizes for each Group of initial SF level

Group	n
Group 0	53
Group 500	104
Group 1000	89

Common statistical tests to compare differences between groups assume data normality, variance homogeneity and similar samples size [63, 67]. As the data at hand seems to violate, to some degree, these assumptions, non-parametric (distribution-free) tests were taken into account. As such, the Kruskal-Wallis non-parametric test was favored [31, 63]. Generally, this test' null hypothesis stipulates that there are no differences among the samples [22]. These differences are assessed by comparing each sample curve to the form of the population curve [22]. After using the Kruskal-Wallis test on **DurationWeeks** values based on **Group**, the null hypothesis was rejected at the chosen significance level of 5 % (p-value = 7.36×10^{-12}) and thus, it seems that there are differences on the distribution of the duration of the DP depending on the **Group** of initial SF concentration. To further verify if all the groups were different from each other, a post-hoc test was performed. The Games-Howell test was preferred as it seems to be robust to unequal sample sizes and variance and that data non-normality is not problematic [50, 100]. This test's results support the idea that differences between each **Group** are statistically significant (Table 3.8).

Table 3.8: Games-Howell test for comparisons of DurationWeeks values between groups of initial SF level

Comparison	p-value
Group 500-Group 0	4.73×10^{-3}
Group 1000-Group 0	4.27×10^{-10}
Group 1000-Group 500	9.75×10^{-6}

3.2.2.2 Association with Genotype

As previously discussed, **DurationWeeks** and **Genotype** seem to be negatively correlated. To understand and confirm this association, further analysis involving these two variables were performed. Sample sizes are presented in Table 3.9 and seem unbalanced, being the *Homozygous* group overrepresented. Medians of the numerical variables grouped by **Genotype** can be seen in Table A.6. The medians of **DurationWeeks** depending on **Genotype** seem different.

Table 3.9: Sample sizes for each Genotype

Genotype	n
Heterozygous	51
Homozygous	195

The histogram on Figure 3.15 seems to show that there are some differences on the distribution of the **DurationWeeks** depending on the **Genotype** of the HH patients. Additionally, the box plot on Figure 3.16 seems to support the same rationale. It seems that greater durations, of above (at least) 75 weeks, are all associated with the *Homozygous* genotype.

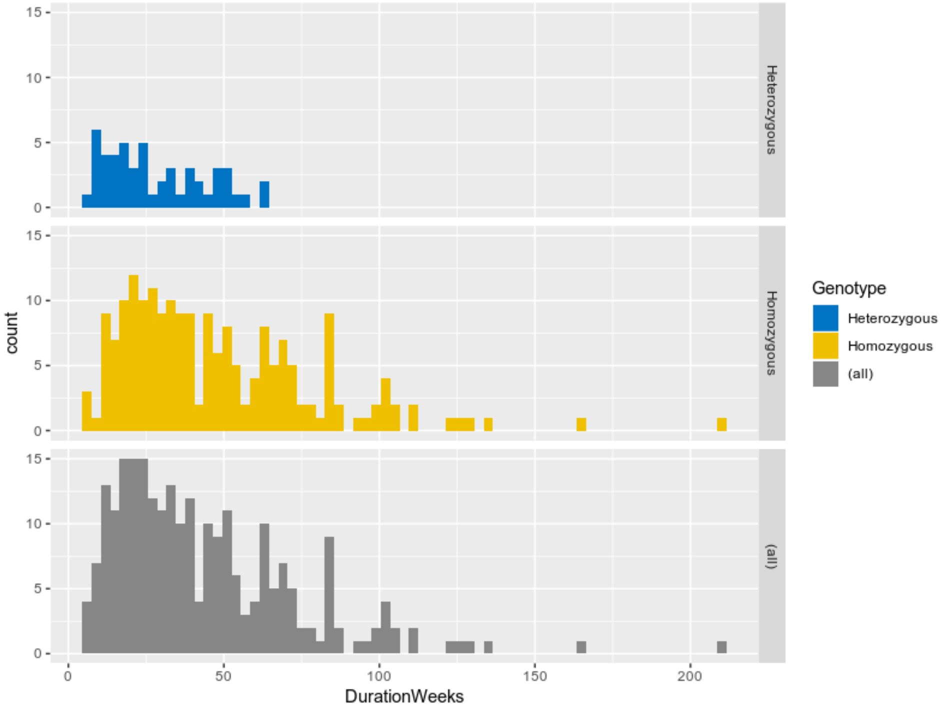


Figure 3.15: Histogram of DurationWeeks values based on Genotype

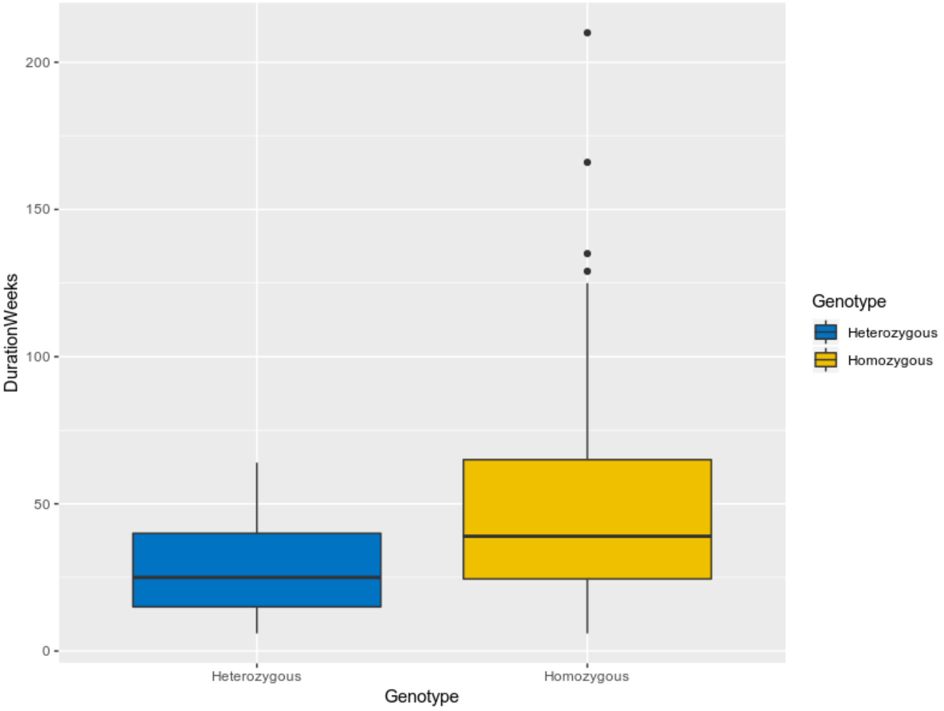


Figure 3.16: Box plot of DurationWeeks values based on Genotype

Hence, a similar statistical analysis was performed to verify differences between *Homozygous* and *Heterozygous* patients on the duration of the DP. Levene's test' null hypothesis was rejected with the confidence established earlier (p-value = 1.6×10^{-3}), leading to believe that the variance of the data is not homogeneous. The Kruskall-Wallis test allowed to indulge that there are significant differences between the two genotypes, as the null hypothesis was rejected (p-value = 7.5×10^{-6}). The initial correlation value between the two variables, the visual analysis through the histogram and box plot and the statistical tests performed allowed to prove the hypothesis that the **Genotype** is associated with the **DurationWeeks**, specifically that the *Homozygous* patients are associated with longer durations of the DP.

3.2.2.3 Association with Sex

Albeit only a mild correlation was found between **DurationWeeks** and **Sex**, this relationship was still studied on a similar fashion to confirm this earlier assessment. In the Appendix, Table A.7 shows the medians for the numerical variables depending on the **Sex** of the patients, Table A.8 shows the sample size for each group, Figure A.2 shows an histogram of the values of **DurationWeeks** per **Sex** and Figure A.3 a box plot involving the same variables. It seems that the medians of the duration of the DP in weeks are very similar for each patient **Sex** and that the *male* patients are overrepresented on the data. Visually, from the histogram and box plot, it is not clear that there are any differences in the distribution of the values of **DurationWeeks** depending on **Sex**. Results from the Kruskall-Wallis test corroborate this idea (p-value = 0.95). These results do not allow to infer any difference between the **Sex** of a patient and the **DurationWeeks**, meaning that the gender of a patient does not seem to influence the duration of the DP of type-1 HH patients.

3.2.2.4 Association with Monthly Phlebotomies

On the correlation analysis endured previously, **DurationWeeks** and **MonthlyPhlebotomies** seemed to be negatively correlated, which seemed to indicate that patients that perform only one phlebotomy per month are more associated with longer treatment periods. Again, sample sizes seem similar and can be seen in Table 3.10. Medians of **DurationWeeks** are different depending on the **MonthlyPhlebotomies** (Table A.9).

Table 3.10: Sample sizes for each level of MonthlyPhlebotomies

MonthlyPhlebotomies	n
>1	114
1	132

Both the histogram (Figure 3.17) and the box plot (Figure 3.18) of **DurationWeeks** based on the number of phlebotomies performed per month seem to confirm that the distributions are different, whereas a single monthly phlebotomy seems responsible for longer durations of the DP on the right tail of the data distribution.

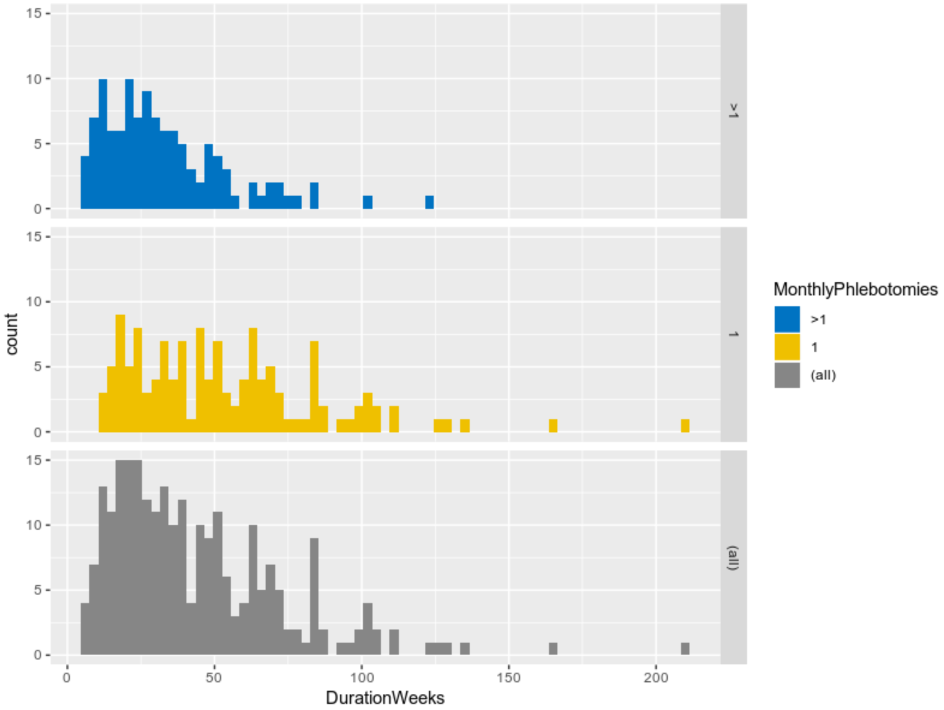


Figure 3.17: Histogram of DurationWeeks values based on MonthlyPhlebotomies

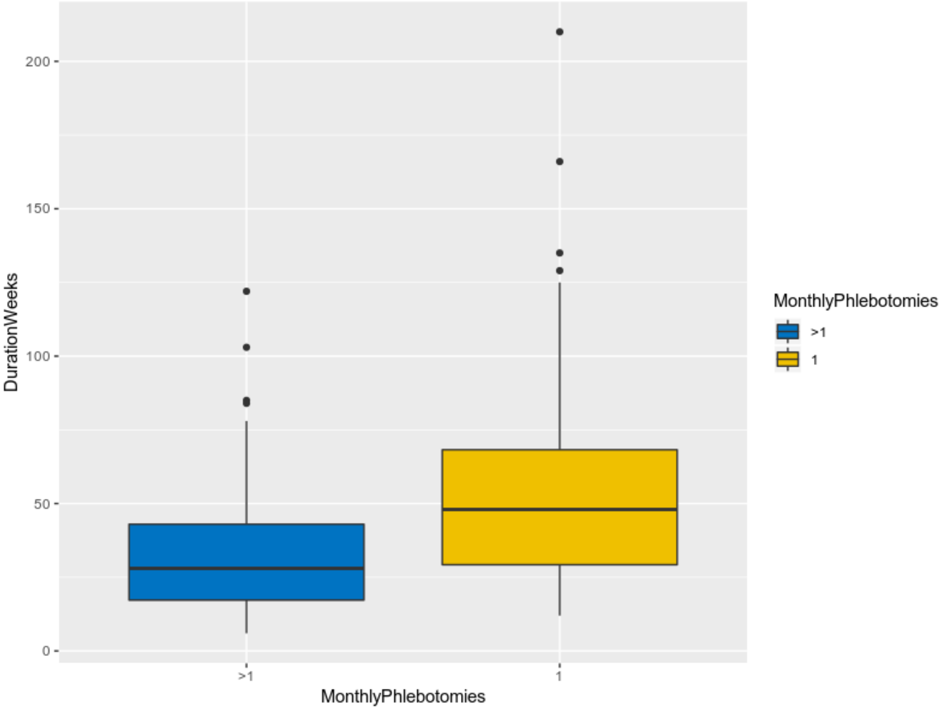


Figure 3.18: Box plot of DurationWeeks values based on MonthlyPhlebotomies

The Levene's test confirmed some degree of data heteroscedasticity (p-value $= 4.9 \times 10^{-4}$). Differences between groups were also confirmed with the Kruskall-Wallis test, where it was possible to reject the null hypothesis (p-value $= 1.4 \times 10^{-8}$). This analysis seems to confirm that the frequency of phlebotomies influences the duration of the DP. Indeed, patients that only perform phlebotomies on a monthly basis seem to be more associated with longer durations of this phase of treatment.

3.2.3 Ferritin variable

The **Ferritin** variable, that contains initial SF levels, was another numerical variable chosen to further investigate relationships with other categorical variables. Actually, at this point, this variable seems very relevant for this study as: i) SF values are taken into account when planning patient-specific therapy, ii) SF concentration at diagnosis is an upfront available biochemical marker and iii) previous analysis demonstrated that higher SF values at diagnosis seem associated with longer DPs. Similarly to what was found with **DurationWeeks**, the range of **Ferritin** values (4767) seems high compared to the mean (1012) and median (840). Moreover, elevated skew (2.39) and kurtosis (7.37) values seem to display positive data skewness.

Table 3.11: Basic statistics for the Ferritin variable

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Ferritin	1.00	246.00	1012.02	765.01	840.00	890.41	480.36	110.00	4877.00	4767.00	2.39	7.37	48.78

Visual analysis of the histogram presented on Figure 3.19 seems to go along with previous findings that the data is right-tailed and consequently deviates from normality. The Shapiro-Wilk normality test also confirms this theory (p-value $= 2.2 \times 10^{-16}$).

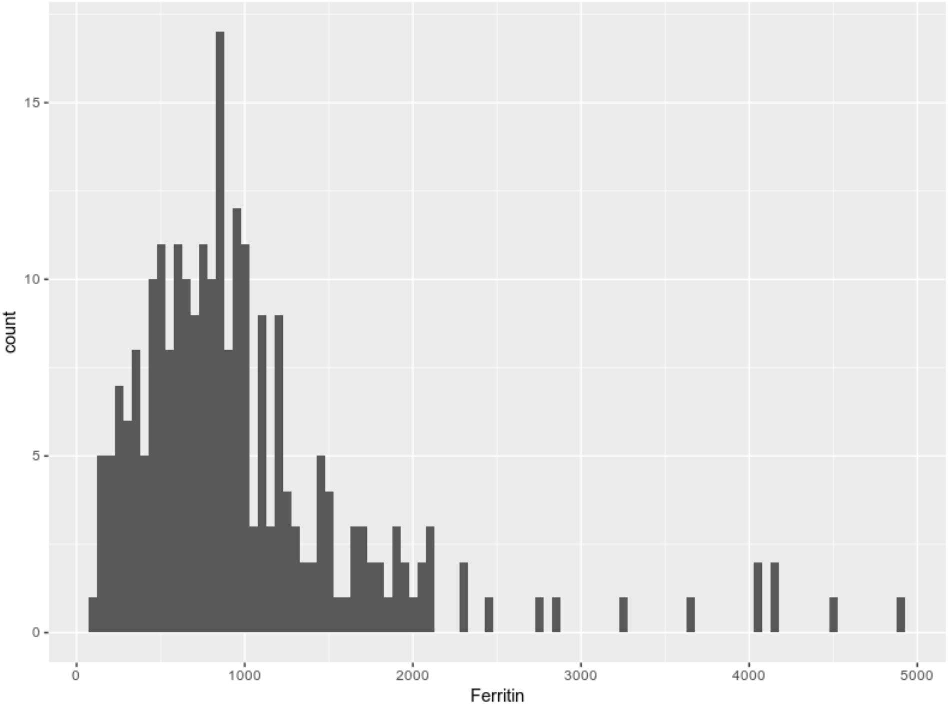


Figure 3.19: Histogram of Ferritin values

3.2.3.1 Association with Genotype

Previous correlation studies showed a mild negative correlation between **Ferritin** and **Genotype**, indicating a possible slight association between *Homozygous* patients and higher SF concentrations at diagnosis. Medians of **Ferritin** values are presented on Table A.6 and seem to differ depending on the **Genotype**. Besides, both the histogram (Figure 3.20) and the box plot (Figure 3.21) of **Ferritin** values based on the **Genotype** seem to reveal some level of discrepancy between the distributions of SF values depending on the **Genotype**. Higher SF levels, located on the tail of the distribution, seem associated with the *Homozygous* genotype. Some data variance heterogeneity was found while performing the Levene's test (p-value = 1.1×10^{-3}) and the Kruskal-Wallis unveiled differences between groups (p-value = 1.7×10^{-4}). Accordingly, it seems that *Homozygous* patients are more prone to have higher SF levels at diagnosis than *Heterozygous* HFE-related HH patients.

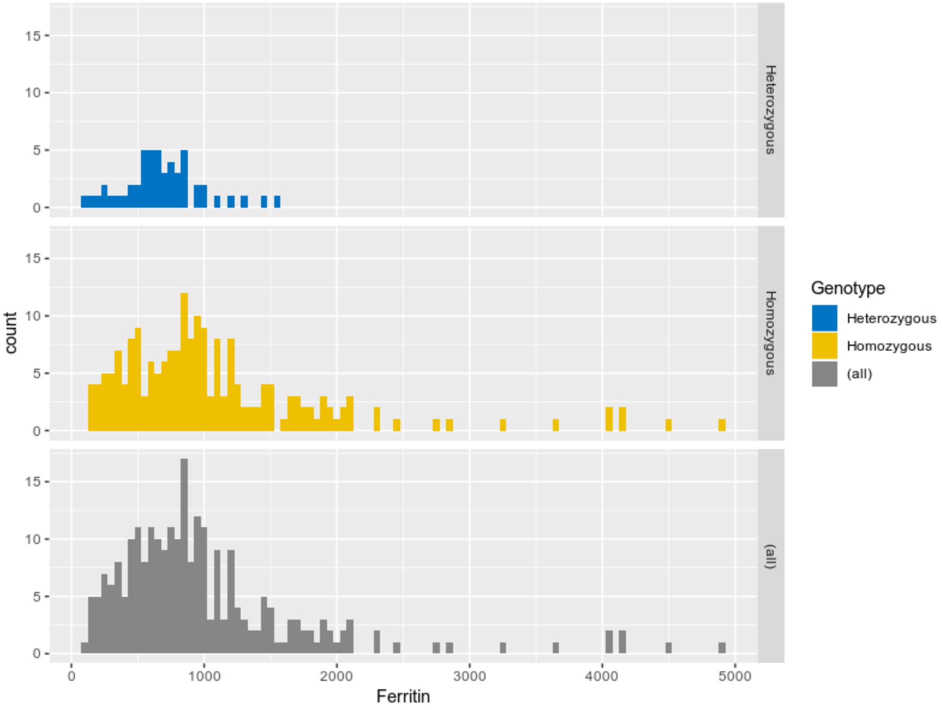


Figure 3.20: Histogram of Ferritin values based on Genotype

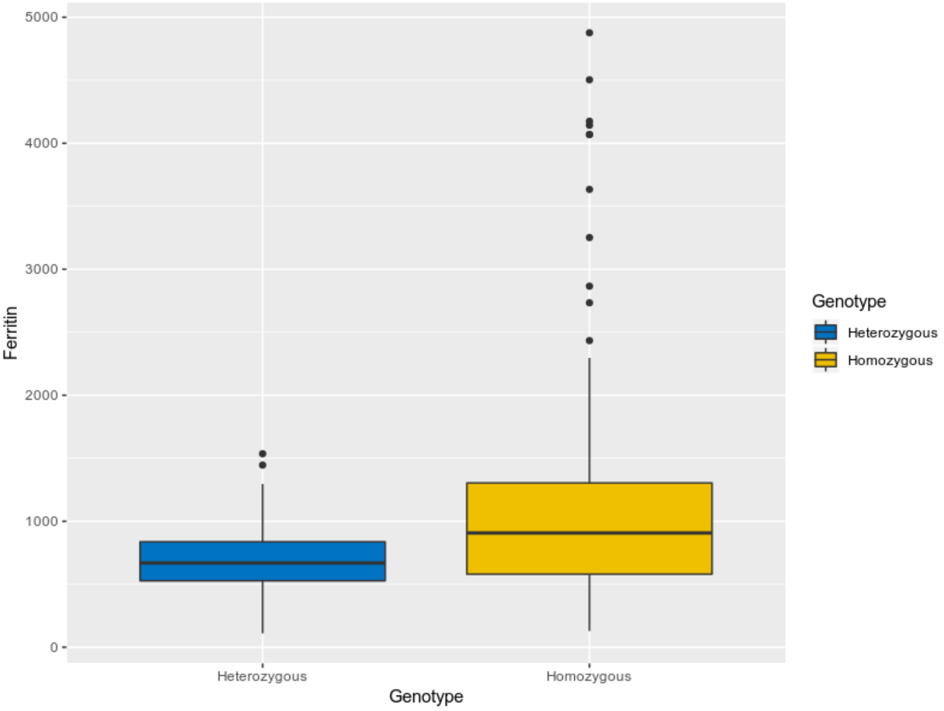


Figure 3.21: Box plot of Ferritin values based on Genotype

3.2.3.2 Association with Sex

While no major correlation was found between **Ferritin** and **Sex**, it seemed important to confirm whether the gender of a patient has influence on the initial SF value. A small difference seem to exist on the medians of **Ferritin** for *male* and *female* patients (Table A.7). Also, by looking at the histogram (Figure 3.22) and box plot (Figure 3.23) of **Ferritin** values by **Sex**, it seems that more extreme (higher) SF values are more associated with *male* patients, but visual analysis seems inconclusive as of now. Thus, after confirming data variance homoscedasticity with the Levene's test (p-value = 0.60), the Kruskal-Wallis test was still selected to assess differences between the genders, as the data seems to violate the assumption of normality. The null hypothesis of this test was rejected with 95% of confidence (p-value $= 3.2 \times 10^{-3}$). These findings hint at a possible association between **Sex** and **Ferritin**, in which *male* type-1 HH patients seem more related to higher SF levels at diagnosis.

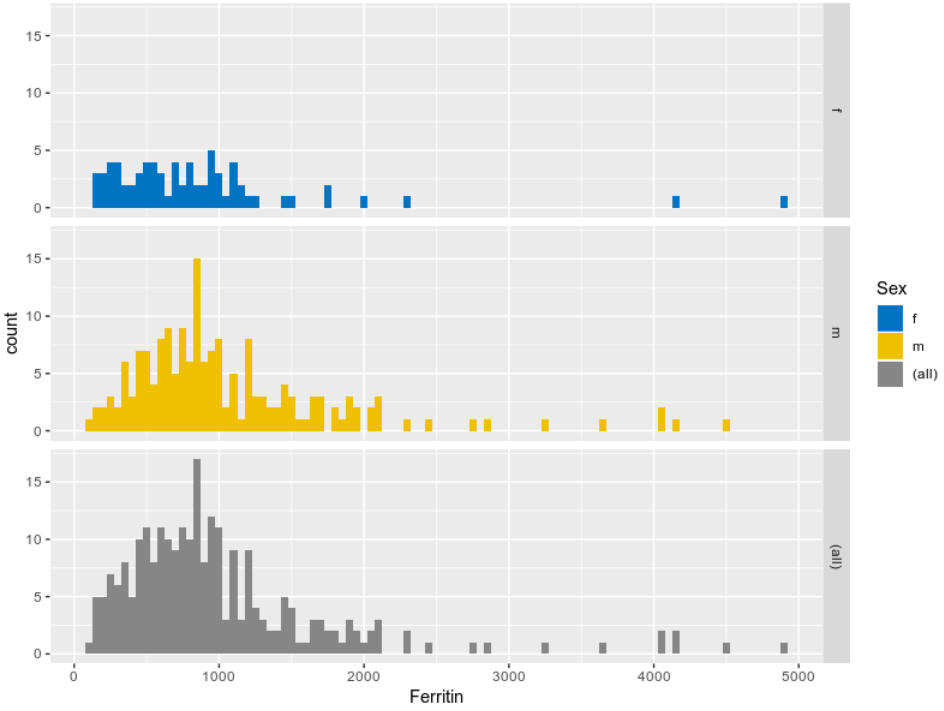


Figure 3.22: Histogram of Ferritin values based on Sex

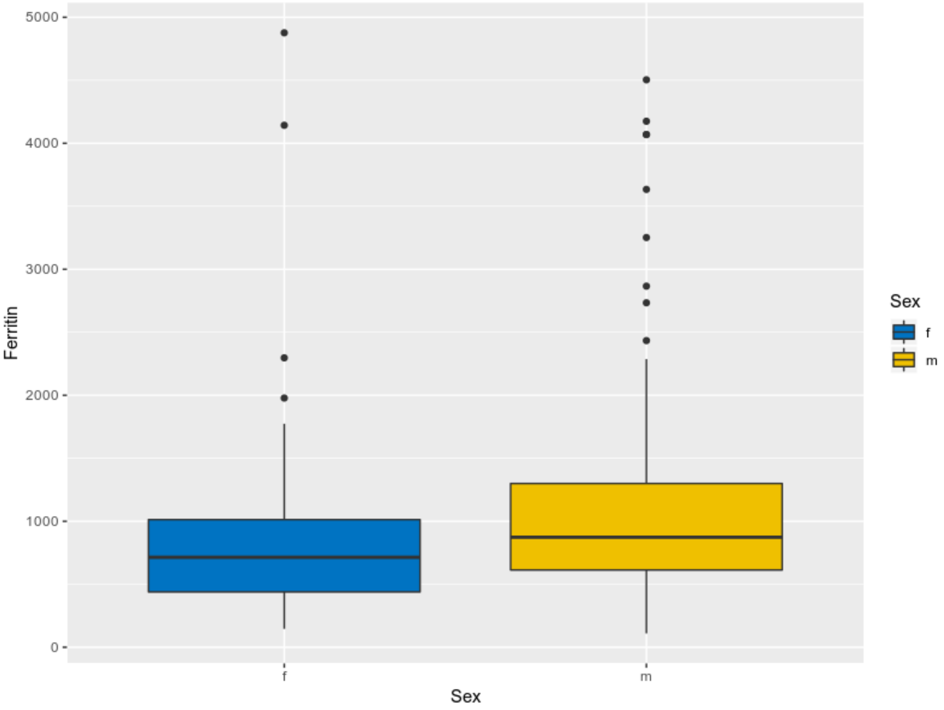


Figure 3.23: Box plot of Ferritin values based on Sex

3.2.4 kParameter variable

Exponential curves found with NLS, for each patient, try to describe the behavior of the SF levels in the course of this treatment stage. The **k** parameters, estimated with NLS, give insights regarding the speed of decay of SF over time during the DP. As seen before, the **kParameter** variable seems to be positively correlated with **DurationWeeks**. This preliminary discovery means that more negative **k** values are more related to shorter durations of the DP and values closer to zero to longer therapy phases. Hereby, this variable seems to add value to this study and it becomes essential to perform a comprehensive analysis of these estimated parameters. Table 3.12 depicts some statistics regarding this variable. The skew value (-1.74) leads to believe that, to some extent, the data is negatively skewed (left tail).

Table 3.12: Basic statistics for the kParameter variable

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
kParamater	1	246	-0.0089	0.0062	-0.0071	-0.0079	0.0041	-0.0361	0.0003	0.0363	-1.74	3.39	0.0004

The histogram on Figure 3.24 also seems to reflect that the data is left skewed. Additionally, results from the Shapiro-Wilk test suggest that the data is not normally distributed (p-value = 2.2×10^{-15}).

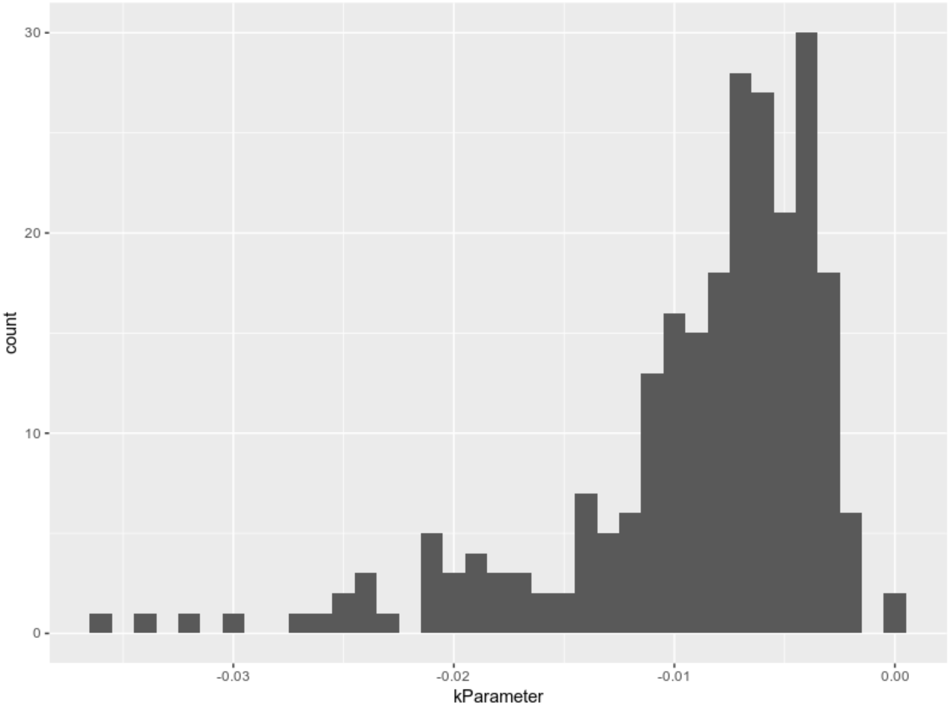


Figure 3.24: Histogram of kParameter values

3.2.4.1 Association with Group of initial SF level

Even if the correlation found between **kParameter** and **Group** of initial SF value seemed weak, further analysis was performed as these two variables seem highly associated with the duration of the DP. At a first glance at Table A.5, medians between each group seem very similar. The histogram plotted (Figure 3.25) and the box plot (Figure 3.26) do not seem to show noticeable differences on the **k** parameters distribution based on **Group**. The Levene's test results seem to indicate some degree of data variance heterogeneity with 95% of confidence (p-value = 0.043). The null hypothesis of the Kruskal-Wallis test was not rejected (p-value = 0.22), meaning that there are no differences on the **k** parameters depending on the **Group** of SF levels at diagnosis. As this test' null hypothesis was not rejected, no post-hoc test for assessing differences of groups was done. Thereafter, the verdict from this statistical test was that the grouping of initial SF concentration does not seem to have influence on the speed decay of SF during the DP.

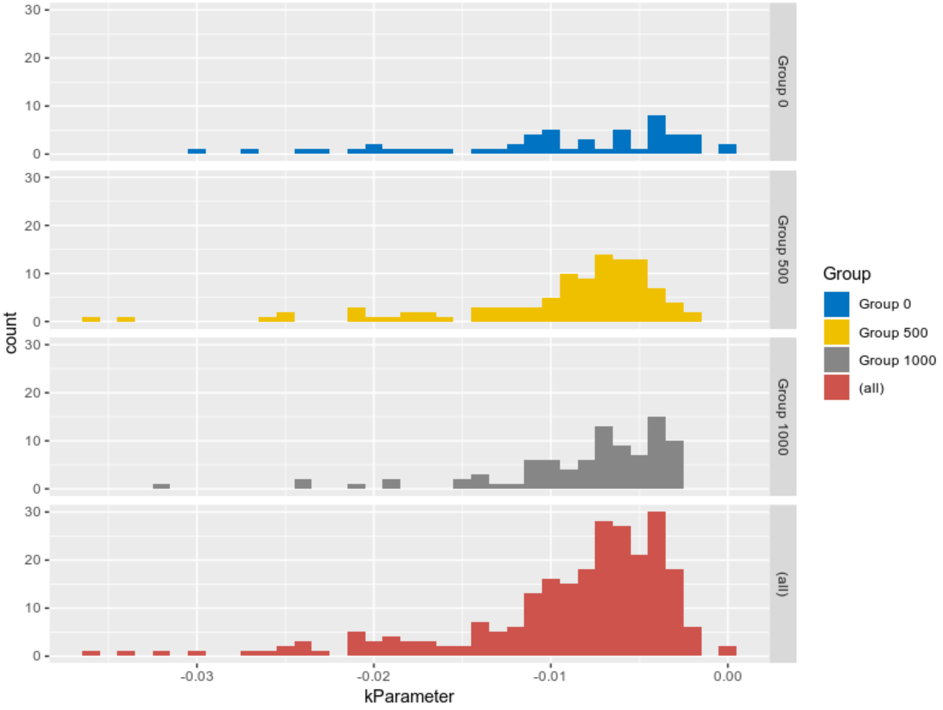


Figure 3.25: Histogram of kParameter values based on Group

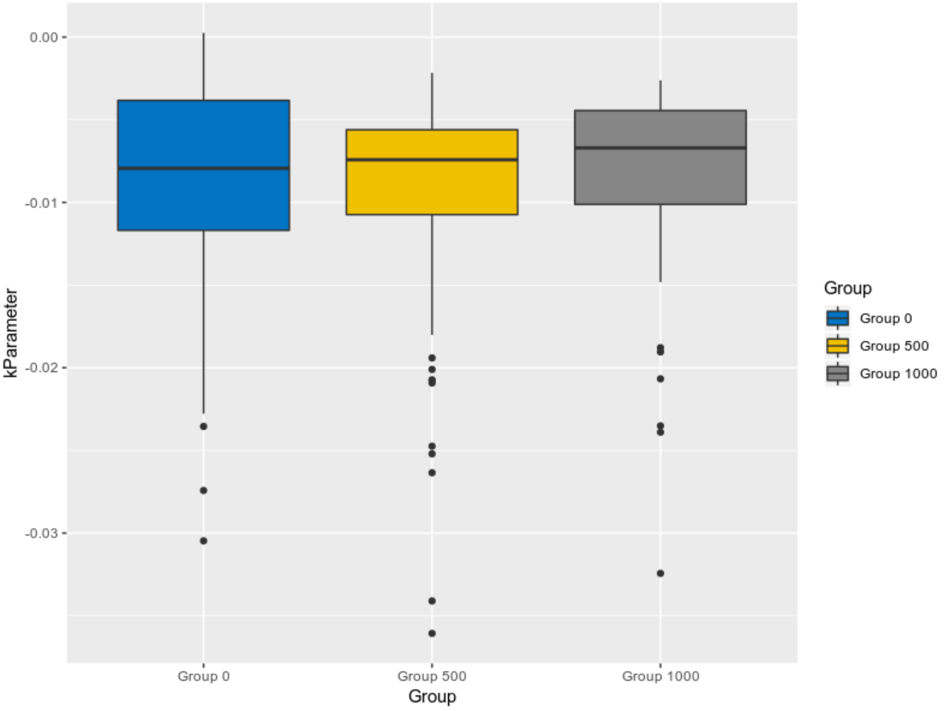


Figure 3.26: Box plot of kParameter values based on Group

3.2.4.2 Association with Genotype

Further, possible associations between the **kParameter** and **Genotype** were assessed. A negative correlation was previously found between these two variables. Medians of the **k** values seem to vary depending on the **Genotype**, as *Heterozygous* patients have a higher median (Table A.6). More than that, both the histogram (Figure 3.27) and box plot (Figure 3.28) seem to show slightly different distributions, where *Heterozygous* patients seem more spread along more negative values. Data heteroscedasticity was assessed with the Levene's test and seems to be present (p-value = 3.4×10^{-4}). Statistically significant difference between groups was checked with the Kruskal-Wallis and the null hypothesis was rejected for the selected significance level (p-value = 2.1×10^{-5}). Consequently, it seems that the **kParameter** values are associated with the **Genotype**, in that *Heterozygous* HH patients tend to be related to more negative **k** values and the *Homozygous* individuals to closer to zero values of this parameter.

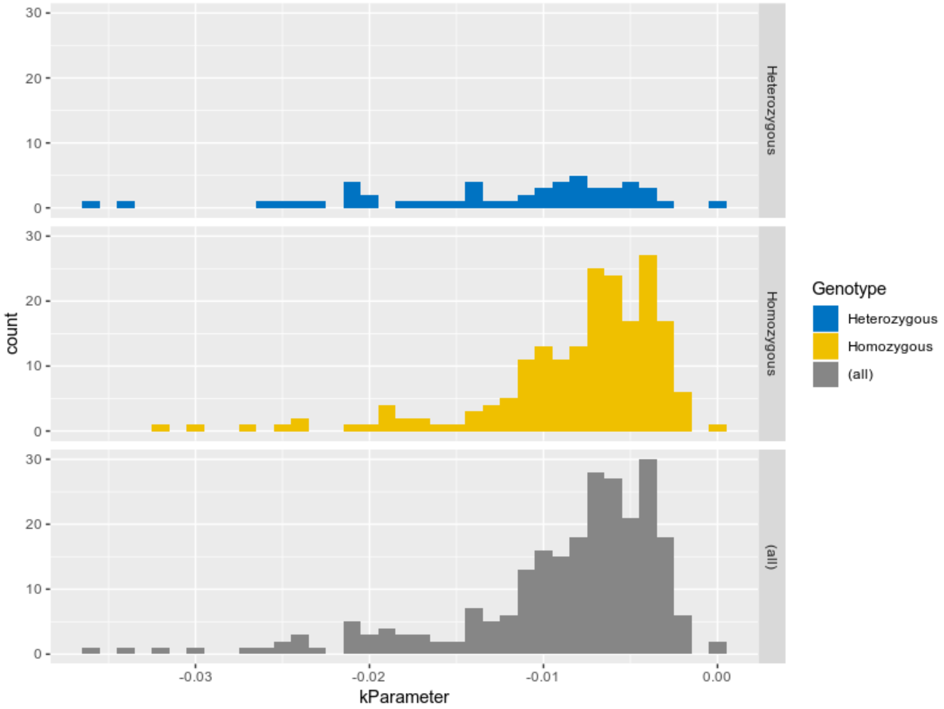


Figure 3.27: Histogram of kParameter values based on Genotype

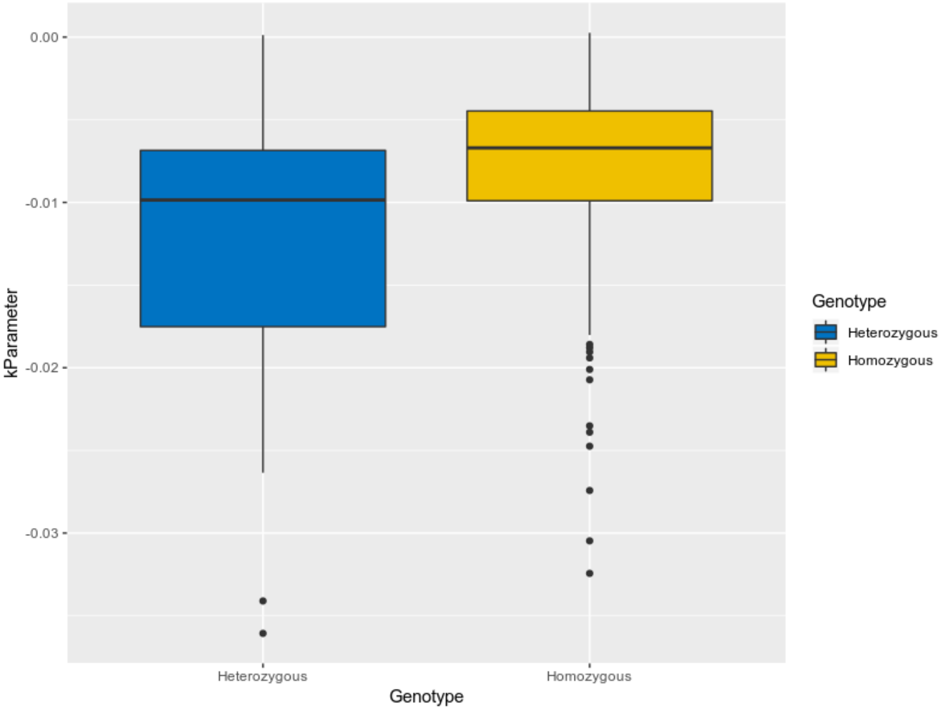


Figure 3.28: Box plot of kParameter values based on Genotype

3.2.4.3 Association with Sex

No relationship was discovered from the previous correlation analysis involving **kParameter** and **Sex**. Also, this was confirmed visually with a histogram (Figure A.4) and a box plot (Figure A.5) and through the Kruskal-Wallis test, in which the null hypothesis was not rejected (p-value = 0.74). In sum, there does not seem to exist any association between the speed of decay of the SF concentration, **k** parameters, with the **Sex** of a type-1 HH patient.

3.2.4.4 Association with Monthly Phlebotomies

In line with the negative correlation found between **kParameter** and **MonthlyPhlebotomies**, an extended analysis was performed to validate this relationship. Disparity of the medians of the **k** values depending on the number of phlebotomies done per month was a first identified sign that corroborates this preceding correlation study (Table A.9). On top of that, a histogram (Figure 3.29) and box plot (Figure 3.30) drawn seem to show varying distributions of **k** parameters values based on **MonthlyPhlebotomies**, in which patients that do more than one phlebotomy per month seem more related to more negative values of this exponential decay parameter. This data seems to possess variance heterogeneity, according to the Levene's test (p-value = 7.8×10^{-7}). Finally, the Kruskal-Wallis test assisted on establishing differences between groups (p-value = 1.84×10^{-13}). Indeed, it seems that patients that only perform one phlebotomy per month are related to slower decays of SF while the patients that do more than one therapeutic session per month are more likely to have higher speeds of decays of SF levels.

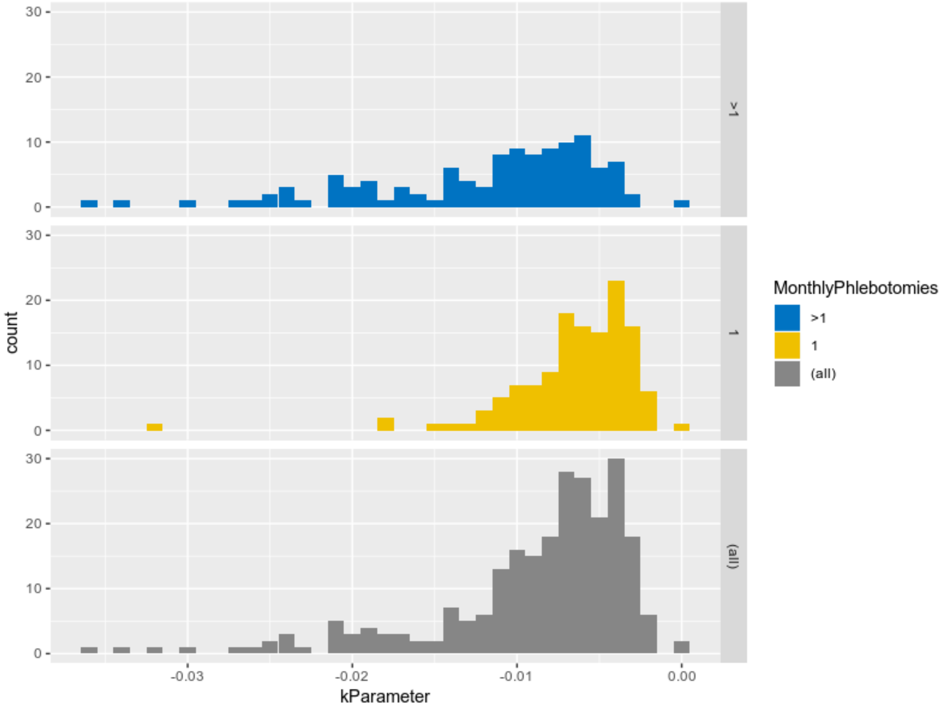


Figure 3.29: Histogram of kParameter values based on MonthlyPhlebotomies

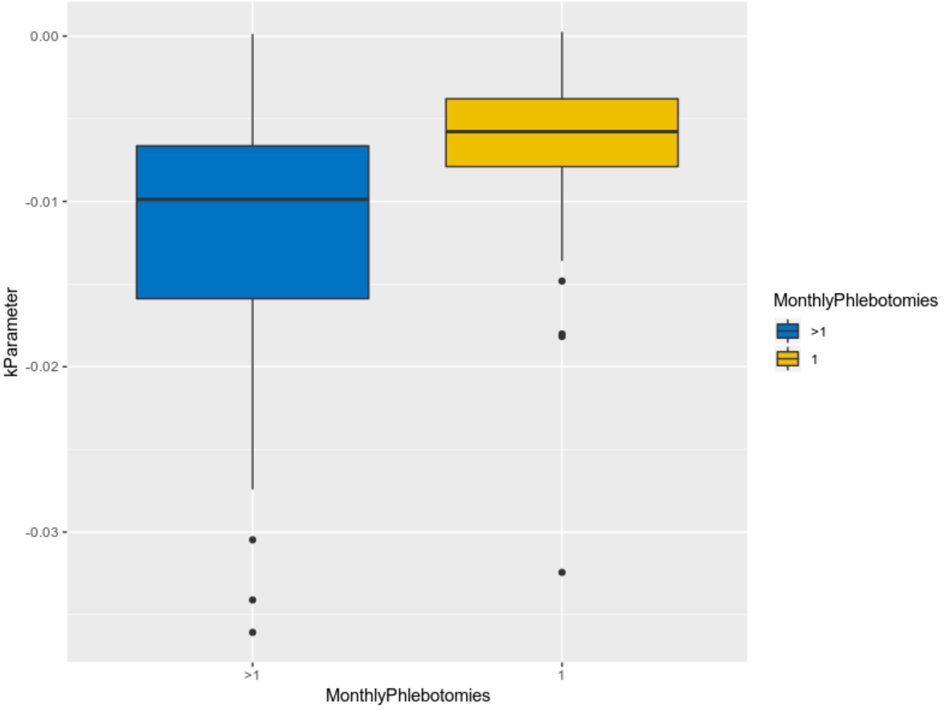


Figure 3.30: Box plot of kParameter values based on MonthlyPhlebotomies

3.2.5 Gradient variable

Contrary to the **kParameter** estimated with NLS that describes a monoexponential curve, the **Gradient** variable describes a linear approximation of the SF decay during the DP. It can be interpreted as an approximation of the average rate of depletion of SF per week. Past correlation analysis pointed to a positive association between **Gradient** and the target variable **DurationWeeks**, meaning that patients with more negative **Gradient** values are more related to shorter durations of the DP. As such, this variable was taken as greatly informative and other possible relationships were assessed. Analogously to the **k** parameter values, the **Gradient** comprises negative values. The negative skewness value obtained (-1.49) suggests non-normal data with a left tail (Table 3.13).

Table 3.13: Basic statistics for the Gradient variable

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Gradient	1.00	246.00	-24.62	18.05	-18.92	-22.05	13.33	-100.58	-0.21	100.37	-1.49	2.38	1.15

By examining the histogram on Figure 3.31, it was possible to visually identify a left tail, associated with negative skewness. Lastly, data non-normality was validated with the Shapiro-Wilk test, where the null hypothesis was rejected (p-value = 1.3×10^{-13}).

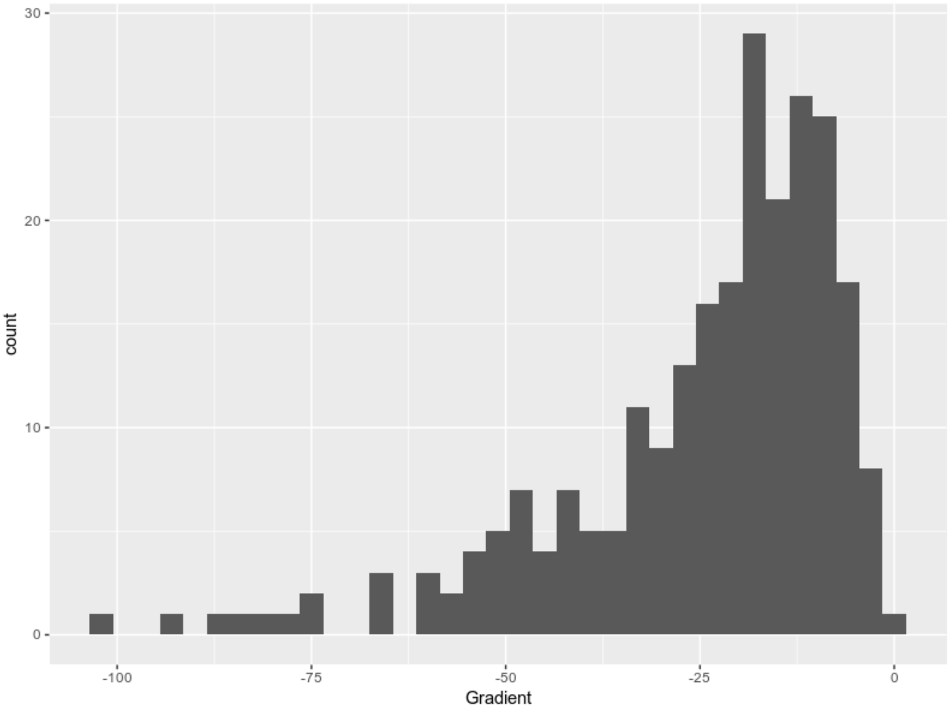


Figure 3.31: Histogram of Gradient values

3.2.5.1 Association with Group of initial SF level

While no differences between **kParameter** and **Group** of SF at diagnosis seemed to occur, initial correlation studies indicated some degree of positive relatedness between **Gradient** and **Group** of initial SF levels. First, medians of **Gradient** values between each **Group** seem to differ (Table A.5). Furthermore, by analyzing the histogram in Figure 3.32 and box plot in Figure 3.33, it seems that the distributions of **Gradient** values diverge depending on the **Group** of SF concentration at diagnosis. Indeed, increases on the threshold of initial SF level (*Group 0* < *Group 500* < *Group 1000*) seem to be associated with extreme, more negative, **Gradient** values, related to the left tail of the distribution of all the patients. Next, data heteroscedasticity was tested and confirmed with the Levene's test (p-value $= 1.3 \times 10^{-4}$) and difference between groups was also verified with the Kruskal-Wallis test (p-value $= 3.8 \times 10^{-14}$). To compare each **Group** with each other, the Games-Howell post-hoc test was used. It was possible to confirm that the **Gradient** values are statistically different between each **Group** of SF at diagnosis (Table 3.14). These insights suggest that: i) HFE-related HH patients with SF values at diagnosis > 1000 μg/L are more associated with more rapid depletion rates than patients with initial SF ≤ 1000 μg/L and ii) patients with SF at diagnosis ≤ 1000 μg/L and ≥ 500 μg/L are related to higher depletion rates of SF than patients with initial SF concentration < 500 μg/L.

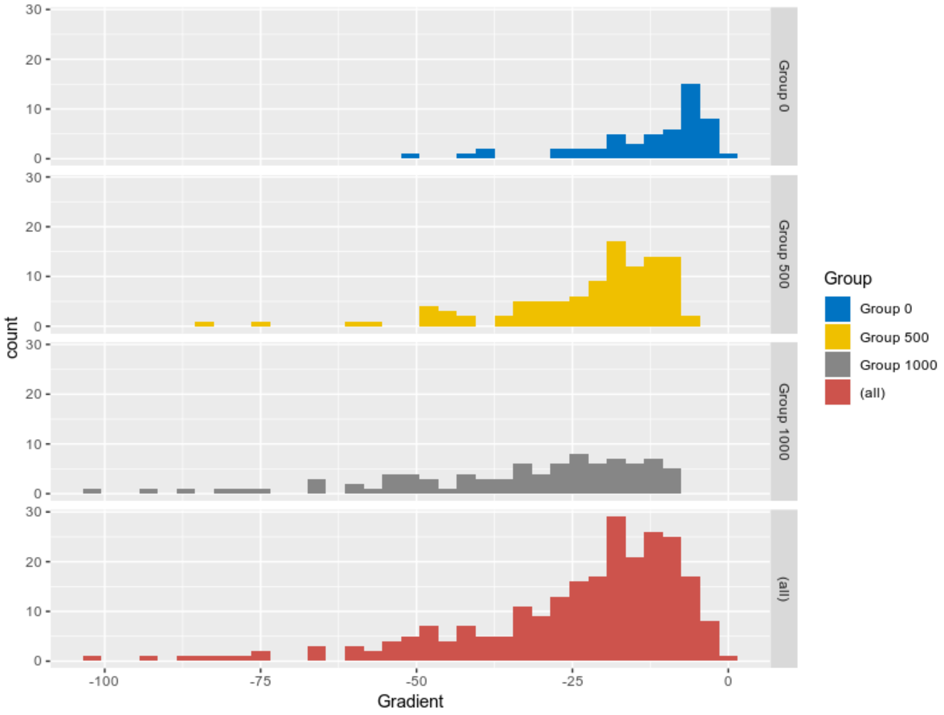


Figure 3.32: Histogram of Gradient values based on Group

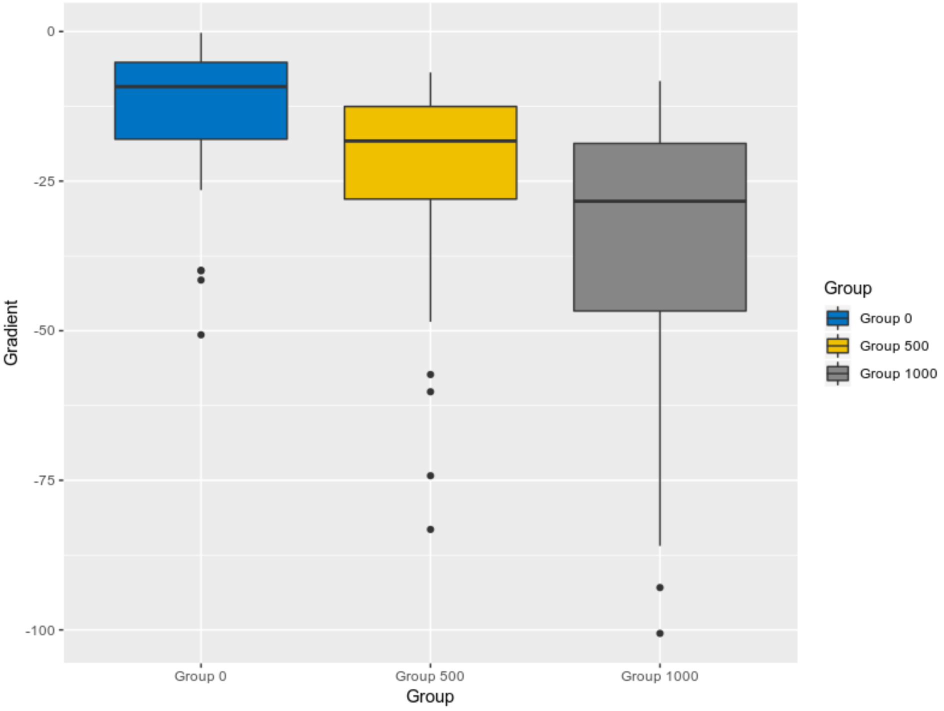


Figure 3.33: Box plot of Gradient values based on Group

Table 3.14: Games-Howell test for comparisons of Gradient values between groups of initial SF level

Comparison	p-value
Group 500-Group 0	2.2×10^{-5}
Group 1000-Group 0	1.5×10^{-12}
Group 1000-Group 500	4.1×10^{-5}

3.2.5.2 Association with Genotype

No clear correlation was found between **Gradient** and **Genotype** on initial correlation analysis. This was also confirmed by: i) inspecting the medians of **Gradient** between each **Genotype**, which seemed very similar (Table A.6), ii) observing the histogram (Figure A.6) and box plot (Figure A.7), which did not seemed to show noteworthy differences on the distribution of each **Genotype** and iii) performing the Kruskal-Wallis test, in which the null hypothesis was not rejected (p-value = 0.26). In fact, it seems that **Genotype** does not have any apparent influence on the rate of depletion of SF per week of a type-1 HH patient.

3.2.5.3 Association with Sex

After some slight positive correlation found between **Gradient** and **Sex**, extended analysis was needed to confirm this hypothesis. A minor median difference seems to exist between *male* and *female* patients regarding their **Gradient** values (Table A.7). Moreover, by observing the histogram in Figure 3.34 and the box plot in Figure 3.35, it seems that more *male* patients are associated with more negative **Gradient** values. The null hypothesis of the Levene's test was rejected, suggesting data variance heterogeneity (p-value = 0.027). The Kruskal-Wallis test' null hypothesis was also rejected (p-value $= 5.3 \times 10^{-4}$), indicating possible different rate of depletions of SF during the DP for *male* and *female* HH patients. This analysis suggests that *male* patients may be more likely to have higher weekly rates of depletion of SF during the DP.

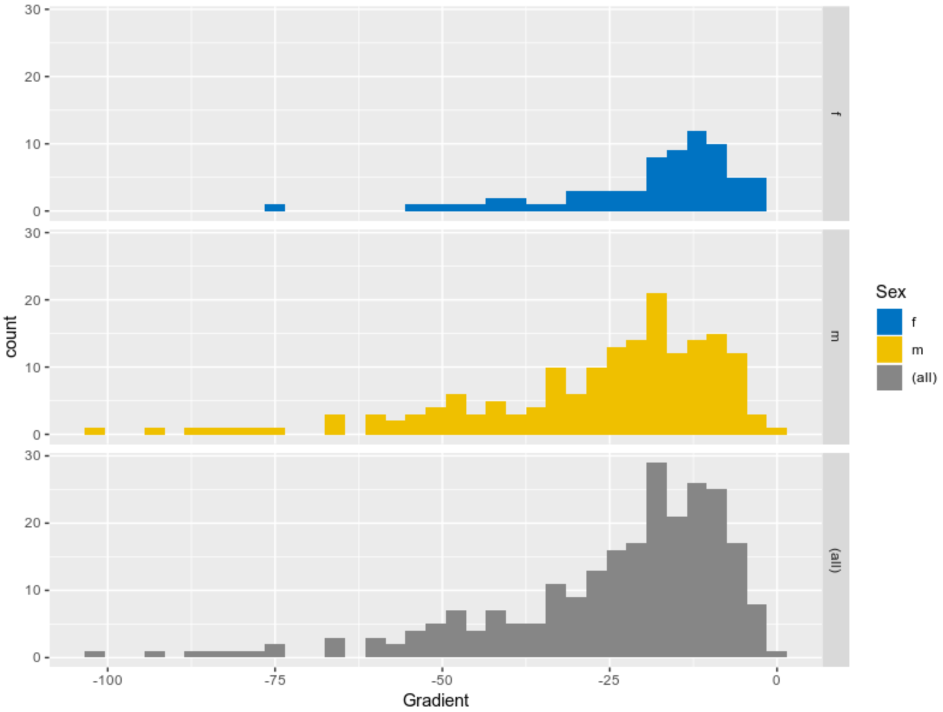


Figure 3.34: Histogram of Gradient values based on Sex

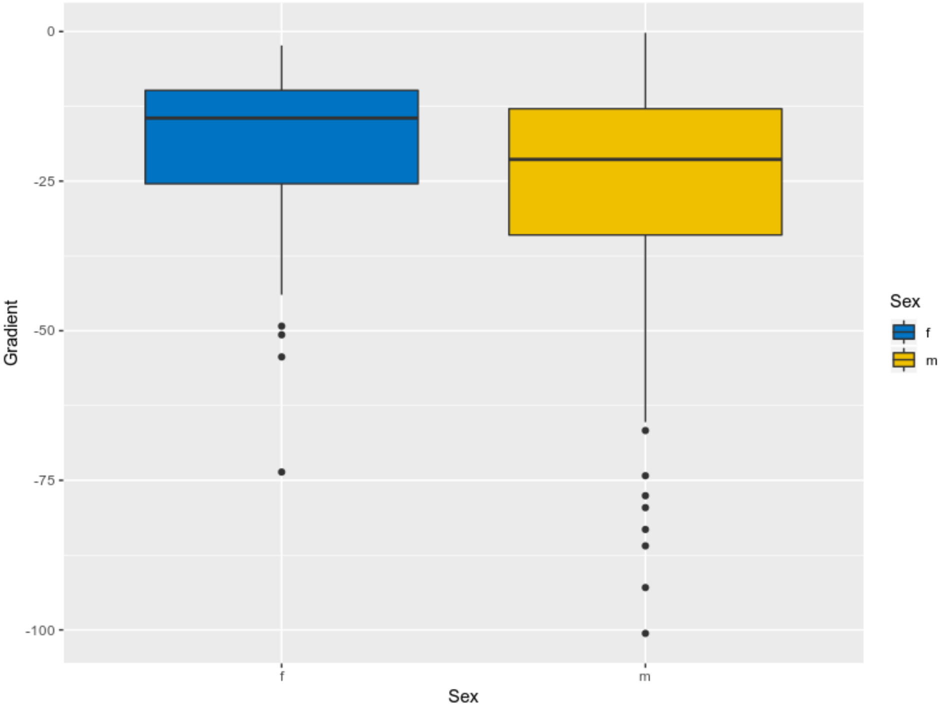


Figure 3.35: Box plot of Gradient values based on Sex

3.2.5.4 Association with Monthly Phlebotomies

Finally, the relationship between the **Gradient** variable and **MonthlyPhlebotomies** was assessed. The negative correlation found between these two variables earlier on this study suggests that patients that endure only one phlebotomy per month are more related to slower rates of depletion of SF during this treatment stage. To further investigate this hypothesis, an histogram and a box plot were drawn to examine the **Gradient** values based on the number of phlebotomies performed per month. Besides the median **Gradient** values appearing to be different depending on **MonthlyPhlebotomies** (Table A.9), the plots mentioned suggest varying distributions based on the frequency of therapeutic sessions (see Figures 3.36 and 3.37). Again, as data seems to violate the assumption of normality and variance homogeneity (Levene's test p-value $= 3.2 \times 10^{-5}$), the Kruskal-Wallis test was taken into account for group comparison. This test' null hypothesis was rejected (p-value $= 2.0 \times 10^{-15}$), implying differences of **Gradient** values depending on **MonthlyPhlebotomies**. Actually, this analysis suggests that: i) type-1 HH patients that perform one phlebotomy on a monthly basis are more associated with slower rates of depletion of SF during the DP and ii) the patients that endure more than one phlebotomy per month are more related to higher rates of depletion of SF on the same stage of therapy.

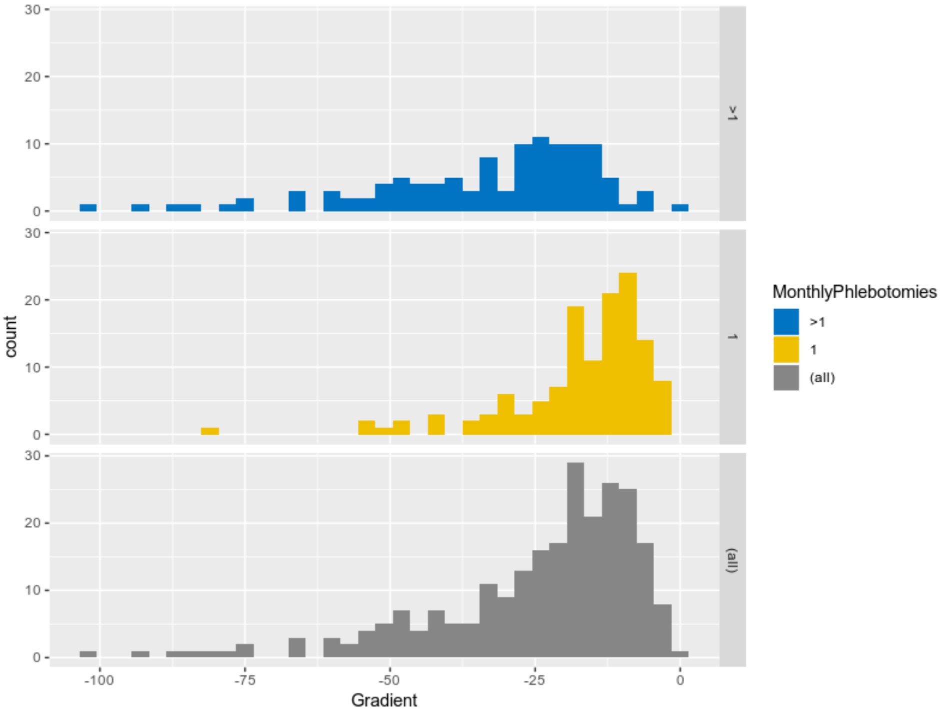


Figure 3.36: Histogram of Gradient values based on MonthlyPhlebotomies

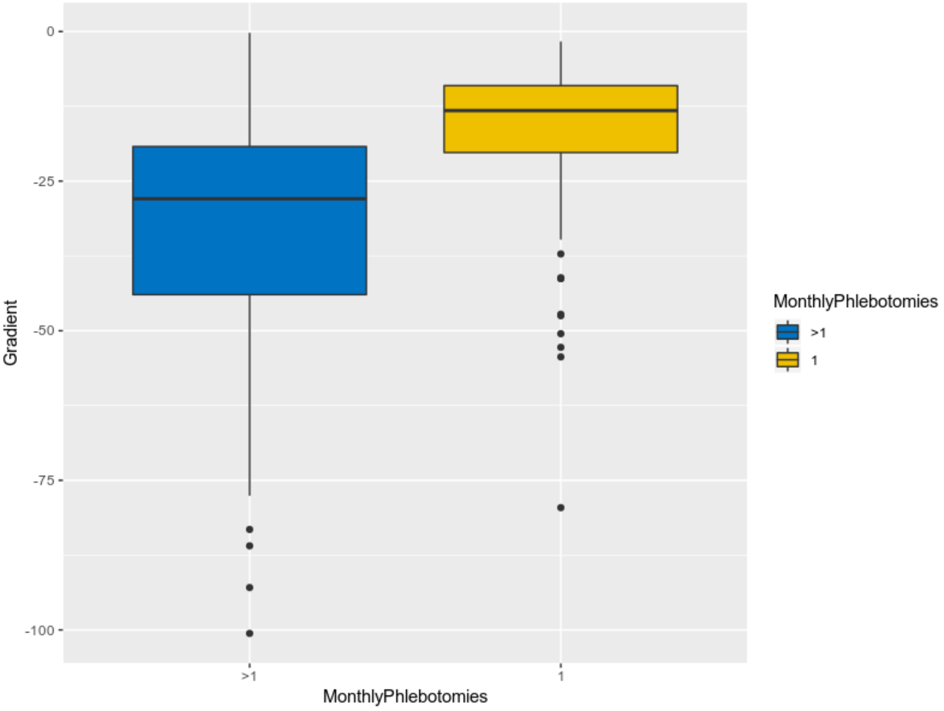


Figure 3.37: Box plot of Gradient values based on MonthlyPhlebotomies

3.2.6 AgeDiagnosis variable

Regarding the **AgeDiagnosis** variable, a more succinct analysis was performed as this variable does not seemed to be highly correlated to any other variable on the initial correlation assay. Correlation with the outcome variable, **DurationWeeks**, also seemed weak. Whilst a closer to zero skewness value was found this time (-0.21) (Table 3.15), the Shapiro-Wilk test' null hypothesis was rejected (p-value = 0.018) suggesting some degree of deviation from normality. The histogram in Figure 3.38 suggests a closer approximation to a normal distribution when compared to previously studied numerical variables. Although, the Kruskal-Wallis was still chosen: i) because the Shapiro-Wilk null hypothesis was rejected and ii) to coherently use the same test for comparisons between groups.

Table 3.15: Basic statistics for the AgeDiagnosis variable

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
AgeDiagnosis	1.00	246.00	49.65	13.53	50.00	49.96	14.83	17.00	80.00	63.00	-0.21	-0.64	0.86

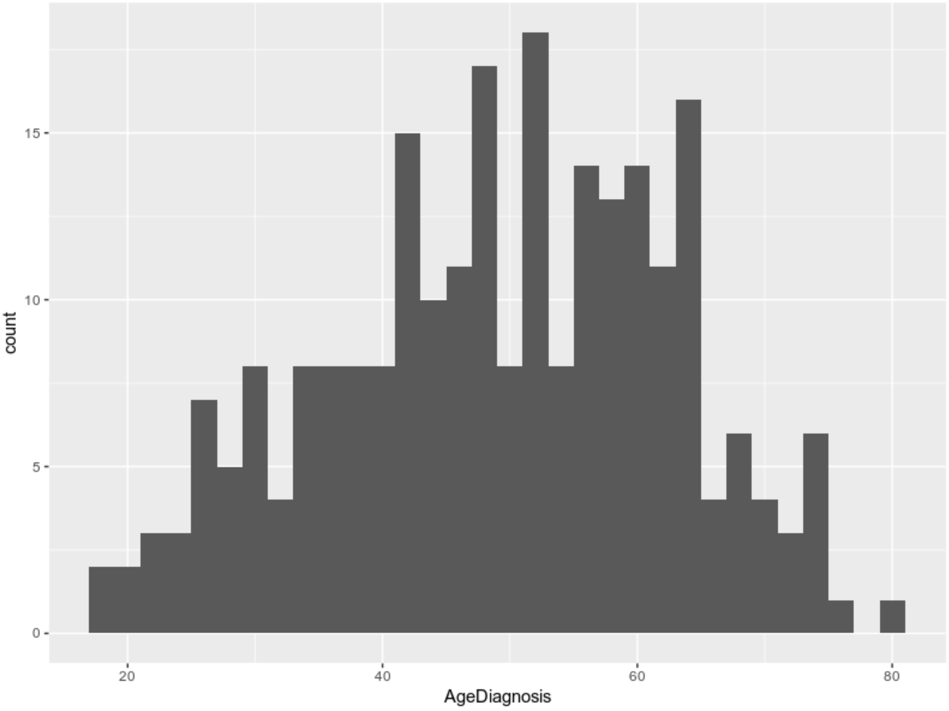


Figure 3.38: Histogram of AgeDiagnosis values

Kruskal-Wallis test' p-values are presented in Table 3.16, where the **AgeDiagnosis** variable was grouped based on the categorical variables. Results from this tests suggest differences in **AgeDiagnosis** values between groups for **Group** of SF at diagnosis and **Sex**. A Games-Howell post-hoc test is presented in the Appendix (Table A.10) and suggests differences between *Group 0* with the other two groups, but not between *Group 500* and *Group 1000*. In sum, these tests suggest: i) no association between **AgeDiagnosis** and **Genotype**, ii) no association between

AgeDiagnosis and **MonthlyPhlebotomies**, iii) an association between **AgeDiagnosis** and **Group**, in which patients with SF at diagnosis < 500 μg/L seem more associated with younger ages at diagnosis than patients with SF levels ≥ 500 μg/L (See Figure A.8 and Figure A.9) and iv) an association between **AgeDiagnosis** and **Sex**, in which *female* patients tend to be more associated with greater ages at diagnosis (See Figure A.10 and Figure A.11).

Table 3.16: Kruskal-Wallis tests for comparisons between groups, where AgeDiagnosis is grouped based on the categorical variables

Relationship	p-value
AgeDiagnosis and Group	1.9×10^{-2}
AgeDiagnosis and Genotype	9.8×10^{-1}
AgeDiagnosis and Sex	7.2×10^{-7}
AgeDiagnosis and MonthlyPhlebotomies	7.1×10^{-1}

3.2.7 Variables not included in the main analysis

Although some variables were not included on the data set for the main analysis of this thesis, some minor data analysis techniques were attempted. Indeed, unveiling their associations may be fruitful for further enhanced data collection and future work. Essentially, this subsection focuses on assessing correlations between previously disregarded variables due to their missing values or due to their not studied time-series behavior during the DP, for the case of biochemical parameters. The variables incorporated on the data set for this secondary analysis were: **LiverDisease**, **Alcohol**, **BMI**, **Transferrin**, **HB**, **ALT** and **AST**. The addition of these variables resulted in a large reduction on the number of patients available on the data set, due to missing values: 246 to 52 patients.

A correlation plot was drawn and can be found in the Appendix (Table A.12). Regarding the **LiverDisease** and **Alcohol** variables, their major correlations are with **ALT** and **AST** variables. These positive correlations suggest that patients with worse liver conditions are more associated with increased values of these two biochemical markers. Also, patients that consume alcohol more regularly seem to be more related to higher values of these parameters. These correlations seem to be in agreement with previous studies where these two biomarkers are identified as possible indicators for high alcohol consumption and liver disease [23, 43, 72, 98]. Besides the **LiverDisease** and **Alcohol** variables, the **BMI** also seems to be positively correlated to **ALT** and **AST**. Also, minor negative correlations between **BMI** and the two variables that try to describe the decay of SF over time, **kParameter** and **Gradient**, were found.

Further, the **Transferrin** variable, with TS values at diagnosis, seems to be positively correlated with **Ferritin** and **Group** of initial SF values, the **kParamater** and the **Gradient** and the target variable **DurationWeeks**. Contrarily, negative correlations are found with **Genotype** and **MonthlyPhlebotomies**. Summarizing, these findings suggest that patients with higher TS levels at diagnosis: i) are associated with higher initial SF values, ii) are associated with slower decays of SF during the DP, iii) are associated with larger durations of the DP, iv) are more associated with the *Homozygous* genotype and v) tend to perform only one phlebotomy per month.

The **HB** concentration at diagnosis seems to have a high negative correlation with **Sex**, suggesting that *male* patients are more associated with higher values of this parameter at diagnosis. In addition, less strong negative correlations could be found between **HB** and **kParamater**, **Gradient** and **DurationWeeks**. **ALT** and **AST** are highly correlated with each other and seem to have a similar pattern of correlations with the other variables. Both seem to be positively correlated with **Ferritin**, especially **AST**. Also, these markers seem to be negatively correlated to the two variables associated with the decay of SF, being the correlations with **Gradient** stronger than the ones found with **kParameter**. However, no noticeable correlations were found between **ALT** and **AST** with the outcome variable **DurationWeeks**. As a future approach, **ALT** and **AST** could be transformed in a another variable that has values regarding their ratio (**ALT/AST**) as they seem to be highly correlated, suggesting collinearity. Also, this ratio is a widely used biomarker to asses liver fibrosis, cirrhosis and heavy alcohol consumption [40, 72].

Of all of these variables, **Transferrin** seems to be the one more strongly correlated with the variable of interest, **DurationWeeks**, evidencing the relevance of this variable. These findings, along with the fact that TS is commonly used as a marker to initiate treatment [1, 60] and that it seems positively correlated with SF levels at diagnosis, suggest that further studies should be endured to assess whether this variable is pertinent to the aims of this work. Despite the fact that this correlation study is an initial exploratory analysis and the number of patients available is very restricted (n = 52), it seems that it unveiled some possibly insightful associations between some of these variables. Indeed, better data collection to reduce the number of lost patients due to missing values and performing a careful study of the behavior of the time-stamped biochemical parameters during the DP may allow to take these variables into account for an exhaustive statistical analysis and eventually consider them as potential explanatory variables to predict the duration of the DP of type-1 HH patients.

3.3 Data processing after EDA: dealing with unknown variables at diagnosis

This section precedes the actual modeling of the DP of HH patients and focuses on preparing the data for the model building, comparison and evaluation. Two key topics are discussed: i) the purpose of the **MonthlyPhlebotomies** for following models building and ii) the grouping of the **kParameter** and **Gradient** values according to other variables.

As established before, the main goal of this work is to build a model to predict the duration of the DP of newly diagnosed HFE-related HH patients. Hence, it is of utmost importance that all the explanatory variables' values are known at the time an individual is diagnosed with this genetic disorder. While **Ferritin**, **Genotype**, **AgeDiagnosis**, **Sex** are well known variables at the time of a patient's diagnosis, the **kParameter**, **Gradient** and **MonthlyPhlebotomies** values can only be obtained a posteriori as they depend on the value of the outcome variable **DurationWeeks**.

3.3.1 The MonthlyPhlebotomies variable

The categorical variable **MonthlyPhlebotomies** has two factors: i) patients that do one phlebotomy per month (*1*) and ii) patients that do more than one phlebotomy per month (*>1*). As this variable's values were retrieved using information from the end of this treatment stage like the duration and the number of phlebotomies performed during this period, no data would be available for a newly diagnosed patient. The phlebotomies' frequency is usually based on medical advice given by the patients' physician or nurse, as there are not any evidence based studies with protocols establishing starting points, frequency and endpoints of the treatment [64]. Also, it seems that the frequency of phlebotomies during the DP may vary from patient to patient, depending on their tolerance to the therapeutic [103]. For instance, the frequency may be altered if the patient is in risk of anemic state [16, 18]. Thus, the proposal of this work is to treat this variable as a recommendation, whereas, in the case of it being included as a predictor on a reasonably accurate predictive model, the patient's physician or nurse would attribute one of the possible factors. Specifically, if the medical practitioner were to advise an arbitrary patient weekly procedures, the correspondent **MonthlyPhlebotomies** factor would be *>1*, while a recommendation of only one phlebotomy per month would be associated with the factor *1*. This would enable the physician to understand the impact of recommending one frequency plan over the other on the predicted duration of the DP.

3.3.2 The kParameter and Gradient variables

Although both the **kParameter** and the **Gradient** variables seem to add value to the analysis and may be useful as explanatory variables for predictive model building, their values are not known at the time of type-1 HH diagnosis. Thereafter, it is proposed to take advantage of previously found insights regarding these variables correlations to characterize a newly diagnosed patient **k** and **Gradient** values. The basic idea was to attribute a **k** and **Gradient** value to a recently diagnosed patient based on other characteristics and values known at diagnosis. Statistically significant differences between groups were taken into account to perform this task. Indeed, it was found before that the **kParameter** values seem to be influenced by the **Genotype** of the patient and the number of phlebotomies performed per month. As such, a newly diagnosed HH patient **k** values would be defined based on these two values. This is achieved by grouping the available patients based on their values for these two variables and attributing a final **k** value to each group, that corresponds to the median **k** values of the patients with those specific characteristics. The median is the metric used due to the previously verified skewness of the data, as the mean is usually more deviated towards the tail of the distribution [32]. On the same fashion, **Gradient** would be characterized based on variables found to have statistically significant **Gradient** values between their groups: **Group** of SF at diagnosis, **Sex** and **MonthlyPhlebotomies**.

In the Appendix, Figure A.13 depicts a decision tree for the process of attributing a **kParameter** value for a newly diagnosed patient and Figure A.14 shows a decision tree with the same purpose for the **Gradient** values. These decision trees may aid on specifying **kParameter** and **Gradient** values for a recently diagnosed patient. Also, Table 3.17 shows the assigned **kParameter** values for a new HH patient and Table 3.18 shows the attributed **Gradient** values. Exponential and linear SF decays were plotted with these assigned values and can be seen in Figure 3.39 and 3.40 respectively. Although the patient specific **kParameter** and **Gradient** values were used later on the models training, two new variables, called **kParameter2** and **Gradient2**, were created with the assigned values to emulate newly diagnosed type-1 patients data. Hence, these new variables would correspond to the data available for a recently diagnosed patient and serve to test the predictive models built on the next section of this thesis.

Table 3.17: kParameter values assigned to a newly diagnosed patient based on Genotype and MonthlyPhlebotomies. The attributed value corresponds to the median of each group

Patient characteristics	k parameter value assigned
Homozygous, 1	-0.0055
Homozygous, >1	-0.0087
Heterozygous, 1	-0.0071
Heterozygous, >1	-0.017

Table 3.18: Gradient values assigned to a newly diagnosed patient based on Group of initial SF, Sex and MonthlyPhlebotomies. The attributed value corresponds to the median of each group

Patient characteristics	Gradient value assigned
Group 0, male, 1	-7.13
Group 0, male, >1	-18.37
Group 500, male, 1	-13.10
Group 500, male, >1	-25.25
Group 1000, male, 1	-20.22
Group 1000, male, >1	-40.34
Group 0, female, 1	-5.15
Group 0, female, >1	-17.99
Group 500, female, 1	-13.66
Group 500, female, >1	-22.71
Group 1000, female, 1	-15.27
Group 1000, female, >1	-28.19

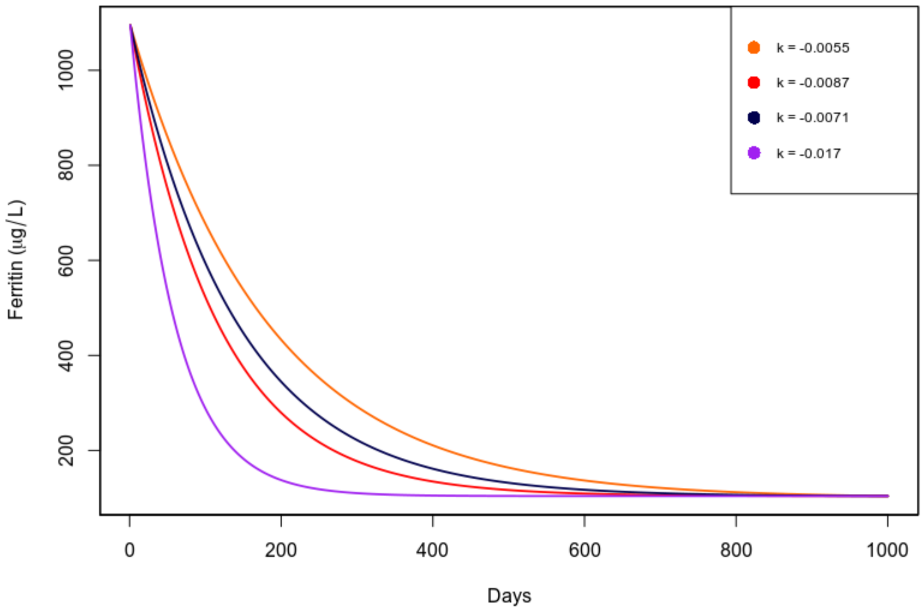


Figure 3.39: Plot of decays of SF with the assigned kParameter2 values

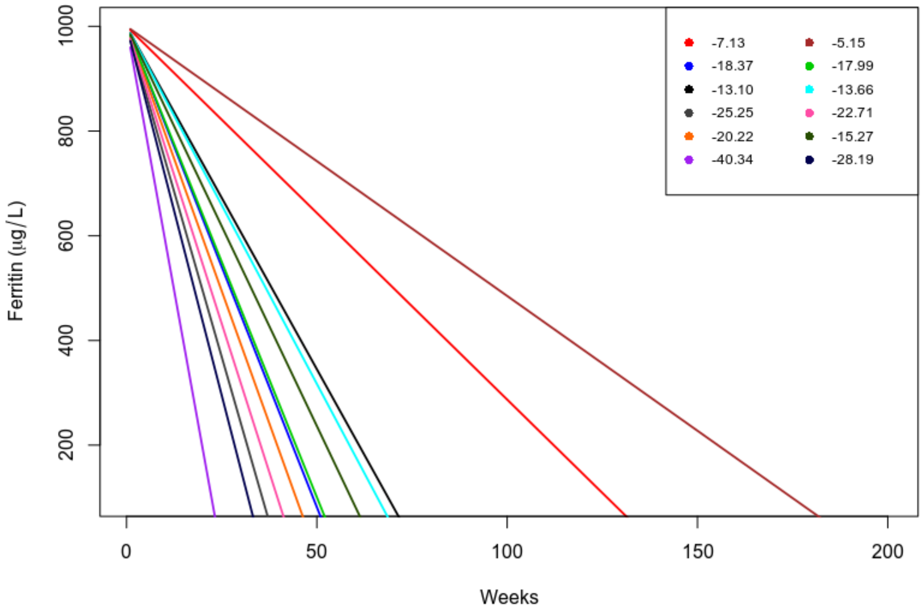


Figure 3.40: Plot of decays of SF with the assigned Gradient2 values

While the **Gradient** values were grouped according to three variables, resulting in twelve different **Gradient** medians, the **kParameter** values were only sub-grouped with two variables and thus, only four medians were obtained. This may result in oversimplified values' assignments, especially for the **kParameter** variable. In fact, the minimum **kParameter** value on the data set is -0.0361 and the maximum is 0.0003 (3.12) while on the **kParameter2** the minimum is -0.017 and the maximum is -0.0055. As for the **Gradient** variable, its minimum is -100.58 and the maximum is -0.21, while the minimum for **Gradient2** is -40.32 and the maximum is -5.15. This may mean that these new variables values do not cover the full range of values of the original ones, **kParameter** and **Gradient**. So, further correlations with these variables have to be found to test group differences and consequently group these variables more efficiently.

As an early analysis, the **Transferrin** variable was taken into account to assess differences of **kParameter** and **Gradient** values based on the TS level at diagnosis. As discussed before, even if this variable is not included on this study's main analysis, it was considered relevant due to its' found correlations with **DurationWeeks**, **Ferritin**, **kParameter** and **Gradient** and its' apparent clinical importance as a biomarker. Also, this variable does not have a large number of missing values (n = 17), only reducing the data set from 246 to 229 patients. This numerical variable was then encoded to categorical with two factors: i) patients with TS level at diagnosis *<45* or ii) patients with TS level at diagnosis *>=45*. These two factors' values were saved on a new variable called **TransfGroup**. This threshold was used as it is a common TS level used to establish the need to initiate iron-depletion treatment [1, 60]. Differences between groups of initial TS level on the **kParameter** and **Gradient** values were assessed with the Kruskal-Wallis test. While the null hypothesis of this test was not rejected for the **Gradient** values (p-value = 0.44), the rejection of the null hypothesis was possible when comparing **kParameter** values based on the **TransfGroup** (p-value = 4.1×10^{-3}). These findings suggest that the TS level at diagnosis influences the exponential decay of SF over time, where patients with TS *>=45%* tend to be more associated with slower speeds of decay (see Figure A.15). Using the **TransfGroup** variable to group the **kParameter** values, along with **Genotype** and **MonthlyPhlebotomies**, resulted in eight different medians of **kParameter** values (Table A.11). While this analysis was not taken into account for the next chapters of this thesis, as the data set without **Transferrin** and **TransfGroup** with 246 patients was the one used for model building, it seems that further studies are needed to confirm the hypothesis of using this variable to group the **kParameter** values.

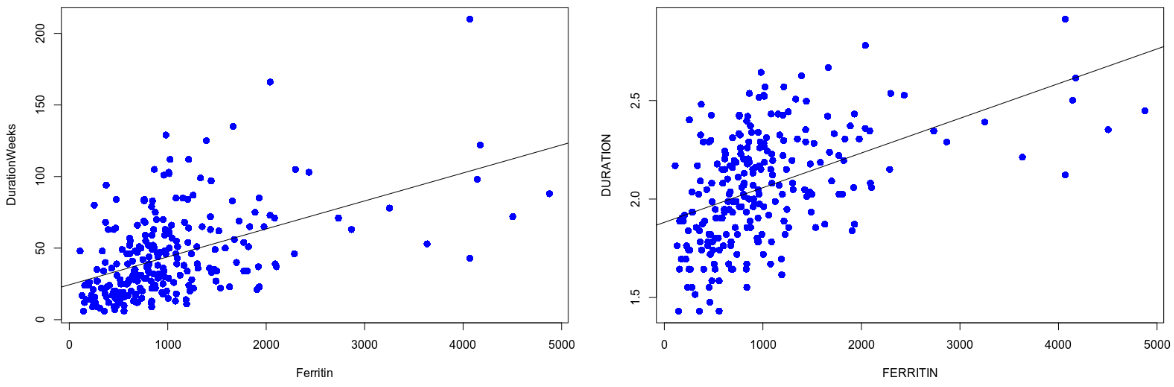
Moreover, a profound study of other variables is needed and may unveil more potential grouping candidates. For example, **ALT** and **AST** were found to be negatively correlated with both **kParameter** and **Gradient**, which hints at two possible variables for grouping these variables, or at least one if the ratio between the two is considered instead of their individual values. Additionally, as differences of **kParameter** and **Gradient** values depending on the number of phlebotomies performed per month seemed to be high, dividing the **MonthlyPhlebotomies** in more categories may aid enhancing the grouping of these variables and perhaps obtain a better cover of the range of these values. However, more data related to patients that endure weekly phlebotomies is needed to have balanced sample sizes.

3.4 Predictive models

This section comprises some procedures before building, evaluating and comparing models to predict the duration of the DP. First, the target variable was transformed to approximate the data to a normal distribution and to a more homogeneous variance for ensuing Linear Regressions. After, an approach was established for the use of **kParameter** and **Gradient** as explanatory variables for all the regressions built. Finally, an analysis to discover interactions between variables was conveyed. The following sections cover the actual model building, using Linear Regressions (section 3.5) and Generalized Linear Models (GLMs) (section 3.6).

3.4.1 Transformation of the outcome variable

Previous analysis suggested that the outcome variable, **DurationWeeks**, had an asymmetric distribution with a right tail (positive skewness) and, as such, deviates from normality. Data non-normality was also confirmed with the Shapiro-Wilk test. Moreover, data heteroscedasticity seemed to be present by using the Levene's test with other categorical variables. Data non-normality and heteroscedasticity are common obstacles when performing Linear Regressions as data normality and homoscedasticity are two assumptions of this method [81]. Data transformation techniques are widely used to improve the normality of a distribution and to equalize the variance to meet the assumptions of Linear Regression models [74]. Specifically, power transformations can be used for this purpose, in which a number is raised to a given exponent [74]. Box-Cox transformations are broadly used data transformations that try to estimate the ideal exponent, also called *Lambda* on this method, for the specific data at hand [74, 96]. As such, the Box-Cox transformation was used to tackle this problem, where a *Lambda* of 0.19 was obtained. Note that the exponent found to transform the data was calculated for a simple Linear Regression with only the **Ferritin** variable as a predictor. This variable was chosen as the only explanatory variable due to being the variable more correlated with **DurationWeeks** with known patient-specific values at diagnosis. Figure A.16 shows a plot with the interval of best *Lambda* values with 95% of confidence. The *Lambda* value was rounded to 0.2 for simplicity reasons. Figure 3.41 shows the **DurationWeeks** plotted against the **Ferritin** values with regression lines before and after the Box-Cox transformation of the target variable.



(a) (b)

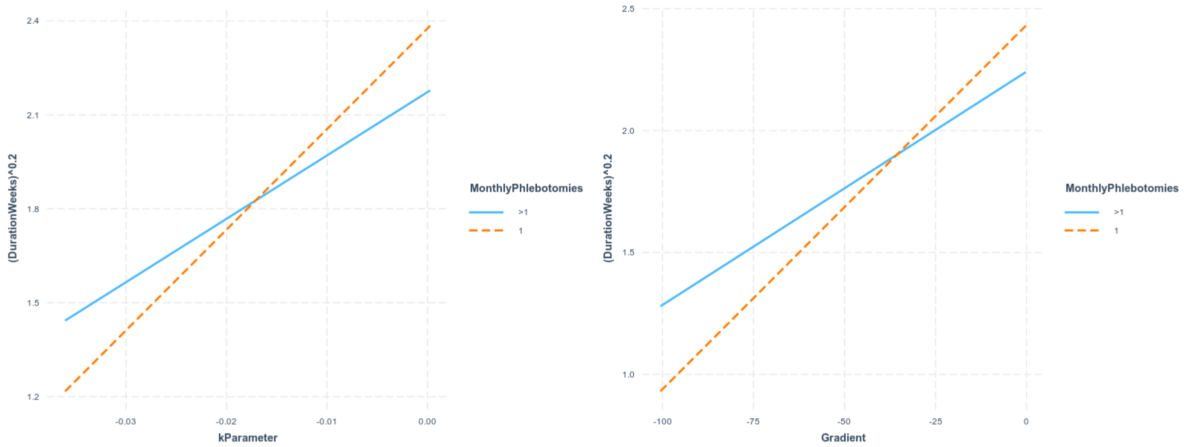
Figure 3.41: DurationWeeks values and associated Ferritin values with a regression line before (a) and after Box-Cox transformation (b)

3.4.2 kParameter vs Gradient

Even though both the **kParameter** and **Gradient** variables seem to add statistical value to model the duration of the DP, based on previous correlation studies, they describe the same physical phenomenon. The approach of this work is to try to infer which approximation of the decay of SF during this treatment stage is more statistically significant to model and predict the duration of the DP. Thus, alternate models were built in which these variables were not included simultaneously. Gaining insights regarding which variable describes more accurately the behavior of the time-stamped SF levels may aid on: i) further characterizing the behavior of the iron depletion on this treatment stage and ii) establish recommendations for future work or better data collection. The following sections follow this approach in which the **kParameter** and **Gradient** are studied separately.

3.4.2.1 Interactions

Prior to model building and comparison, an exploratory analysis of interactions between variables was conducted. Various combinations of possible interactions were carried out. Yet, two of them seemed to require a special attention for the following regressions: i) interaction between **kParameter** and **MonthlyPhlebotomies** and ii) interaction between **Gradient** and **MonthlyPhlebotomies**. Interaction plots were drawn as a preliminary tool to assess possible interactions and seemed to suggest the presence of interactions between the variables mentioned, as the fitted lines are not parallel (Figure 3.42).



(a) (b)

Figure 3.42: Interaction plots in which (a) depicts an interaction between kParameter and MonthlyPhlebotomies and (b) depicts an interaction between Gradient and MonthlyPhlebotomies

3.5 Linear Regressions

Several Linear Regressions were attempted with the **DurationWeeks** as the target variable. Patient specific **kParameter** and **Gradient** were used for the model building. The approach to assess variables relevance and therefore selection of statistically significant explanatory variables and model comparison was based on a stepwise addition of the variables. Also, six main criteria were established for the same purpose: i) *R-squared* value, ii) *Analysis of Variance* (ANOVA) tables and variables' statistical significance, iii) *Variance Inflation Factor* (VIF), iv) *Akaike Information Criterion* (AIC), v) clinical reasoning and vi) model simplicity and interpretability. The R-squared is a widely used metric to assess the goodness of fit of a regression, as it is an estimate of the proportion of variance explained in the outcome variable [68]. ANOVA tables were used to compare models and determine if a more complex model is significantly better than a simpler one [69] and VIFs were inspected to ascertain variables multicollinearity [66]. Further, the AIC balances the trade-off between model complexity and goodness of fit, allowing to compare models for the same sample, in which a lower AIC suggests a better trade-off [17]. Additionally, diagnostic plots like the *Residuals vs Fitted* and *Scale-Location* were reviewed to visually examine possible issues regarding data heteroscedasticity.

As stated, the first variable added was **Ferritin**. This simple model had an R-squared = 0.23 and the **Ferritin** was found to be statistically significant to the model. The second variable additions are summarized in Table A.12. All the variables were statistically significant except **Sex**. Yet, the two variables that are related to the SF decay seemed to increase more the R-squared compared to **AgeDiagnosis**, **Genotype** and **MonthlyPhlebotomies**. Indeed, **AgeDiagnosis** and **Genotype** only incremented the R-squared by decimals and **MonthlyPhlebotomies** doubled this metric's value, while the others at least tripled (**Gradient**) or almost tripled (**kParameter**) the R-squared value. From here, two paths were followed where the third variable addition was done to a model with **Ferritin** and **kParameter** and another with **Ferritin** and **Gradient**.

Starting with the model with **Ferritin** and **kParameter**, the addition of **Genotype** and **Sex** were not significant and the inclusion of **AgeDiagnosis**, **MonthlyPhlebotomies** and an interaction between **kParameter** and **MonthlyPhletomies** seemed statistically significant (Table A.13). Raises on the R-squared value were mild for all the additions mentioned, but adding **AgeDiagnosis** resulted in the slightest increment. Although the addition of the interaction increased slightly more the R-squared, the VIFs of the stated interaction seemed to be higher than the linear addition of **MonthlyPhlebotomies** (see Table A.14 and A.15). For this reason and the fact that the R-squared is similar on both additions, the model considered was with **Ferritin**, **kParameter** and **MonthlyPhlebotomies** without interaction. Next, the addition of **AgeDiagnosis** seemed to be statistically significant to the model mentioned, while only increasing slightly the R-squared value (Table A.16). Also, the AIC values decreased when adding these variables (Table A.17). Equation 3.1 shows the final regression including the **kParameter**.

$$\begin{aligned}(DurationWeeks_i)^{0.2} = 1.914 + 0.0002 \cdot Ferritin_i + 23.34 \cdot kParameter_i \\ + 0.109 \cdot MonthlyPhlebotomies_i + 0.003 \cdot AgeDiagnosis_i + \epsilon_i \quad (3.1)\end{aligned}$$

The results of the ANOVA for the addition of the variables discussed before are shown in Table 3.19. These results also seemed to corroborate the idea that the successive addition of the variables mentioned were significant to the model, implying that each variable addition statistically compensates its underlying error addition.

Table 3.19: ANOVA table where: model 1) Ferritin, model 2) Ferritin + kParameter, model 3) Ferritin + kParameter + MonthlyPhlebotomies and model 4) Ferritin + kParameter + MonthlyPhlebotomies + AgeDiagnosis

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	244	14.76				
2	243	7.76	1	7.00	245.08	0.0000
3	242	7.21	1	0.55	19.42	0.0000
4	241	6.88	1	0.33	11.51	0.0008

Additionally, Figure 3.43 depicts diagnostic plots for this last regression model. By observing the Residuals vs Fitted plot, it seemed that increased fitted values are associated with negative residuals, while reduced fitted values tend to be associated with positive residuals. Indeed, the Residuals vs Fitted and the Scale-Location plots seemed to present a pattern, to some extent, which is a sign of heteroscedasticity. Nonetheless, these patterns do not seemed to be very clear to the point of suggesting high variance heterogeneity. The Normal Q-Q plot suggested a reasonable approximation to a normal distribution of the residuals. The Residuals vs Leverage may aid identifying influential points, albeit the analysis of these points was assessed later on this thesis.

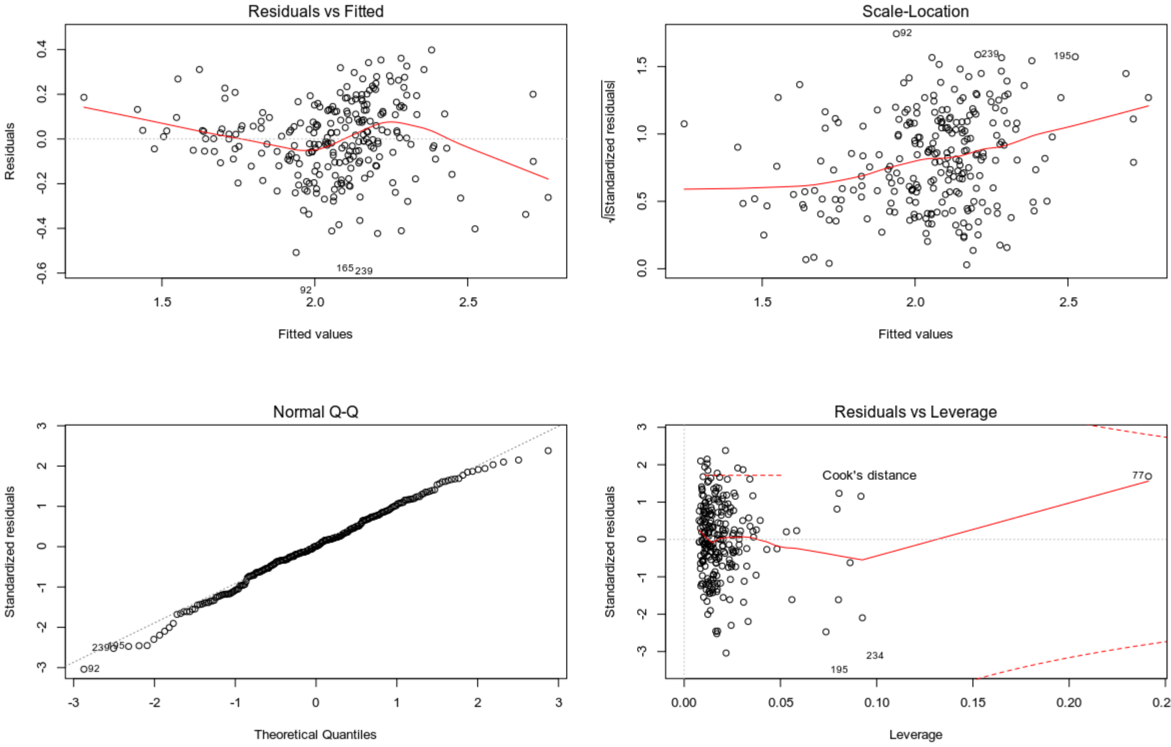


Figure 3.43: Diagnostic plots of final linear regression model considered including the kParameter variable

Similarly, a model with the **Gradient** variable was considered. Table A.18 shows the addition of a third variable to this model and suggests that all variables except **Genotype** are statistically significant. Although, **MonthlyPhlebotomies** linear addition and its' interaction with **Gradient** seemed to be the more significant, akin to what was found with the model with **kParameter**. This time, the interaction was selected to incorporate the model as the R-squared value increased more than the other additions and the VIFs did not seem excessively elevated (A.19). The next variable addition summary can be seen in Table A.20. Again, **AgeDiagnosis** seemed to be the most statistically significant variable and thus, was the one selected to include on the final model. Furthermore, even if the subsequent addition of **Genotype** and **Sex** seemed significant to the model according to the ANOVA results (see Table 3.20), these two variables were not included in order to keep the model simpler and easily interpretable and due to the fact that the R-squared did not increase much when adding these variables. Nevertheless, these findings seemed to suggest that these variables are, to some extent, statistically significant. Also, the AIC values seemed to decrease in each model (Table A.21). The equation regression of the final model selected is presented in Equation 3.2.

$$(DurationWeeks_i)^{0.2} = 1.834 + 0.0003 \cdot Ferritin_i + 0.009 \cdot Gradient_i + 0.191 \cdot MonthlyPhlebotomies_i + 0.005 \cdot Gradient_i \cdot MonthlyPhlebotomies + 0.002 \cdot AgeDiagnosis_i + \epsilon_i \quad (3.2)$$

Table 3.20: ANOVA table where: model 1) Ferritin, model 2) Ferritin + Gradient, model 3) Ferritin + Gradient * MonthlyPhlebotomies, model 4) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis, model 5) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis + Genotype and model 6) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis + Genotype + Sex

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	244	14.76				
2	243	5.75	1	9.02	466.65	0.0000
3	241	5.10	2	0.65	16.70	0.0000
4	240	4.95	1	0.15	7.53	0.0065
5	239	4.86	1	0.09	4.87	0.0282
6	238	4.60	1	0.26	13.54	0.0003

The Residuals vs Fitted and Scale-Location plots in Figure 3.44 suggest a similar but more accentuated pattern to the one found with the model with the **kParameter** variable. Once more, on the Residuals vs Fitted plot, it seemed that greater fitted values were associated with negative residuals. The normal Q-Q plot suggested that the residuals were not as much approximated to a normal distribution compared to the previous model with the **kParameter** variable.

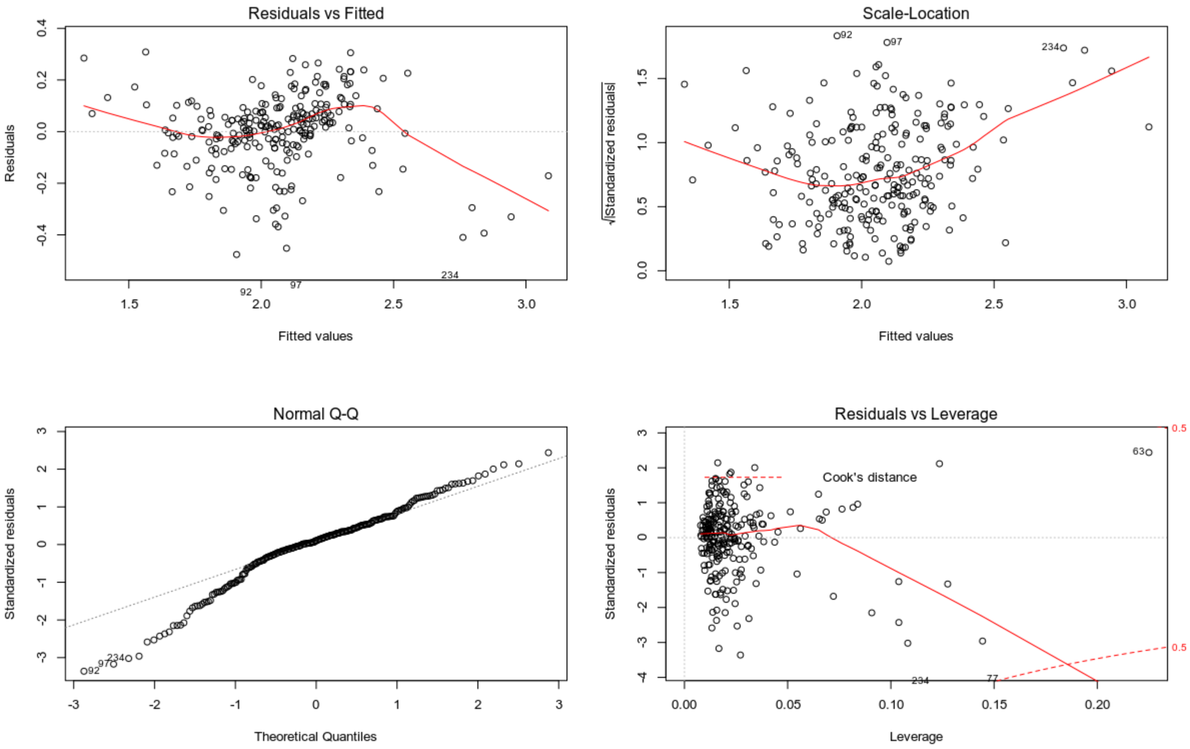
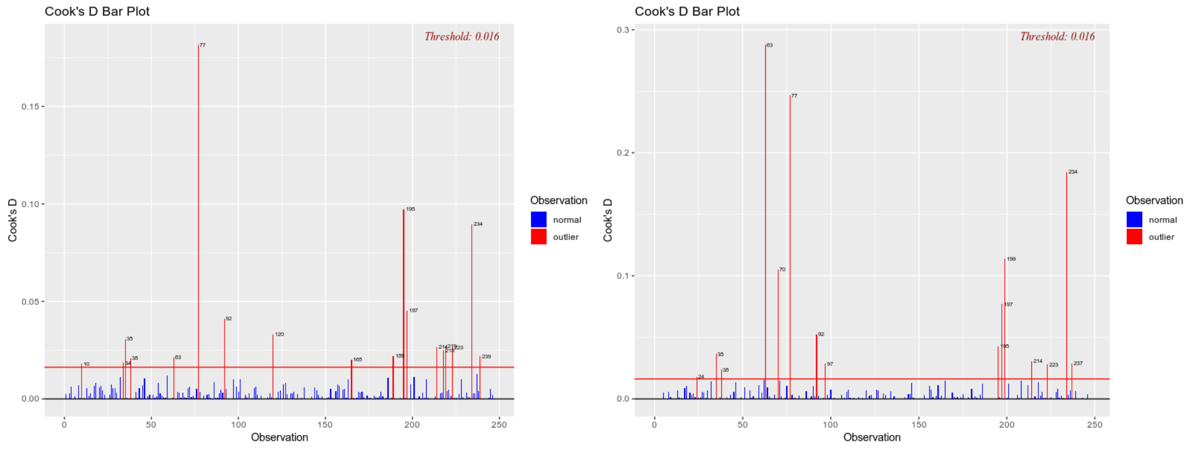


Figure 3.44: Diagnostic plots of final linear regression model considered including the Gradient variable

The approach described here consisted on building two linear regression models, each with one variable that tries to explain the behavior of the SF decay, the **kParameter** and **Gradient** variables. At this point, some insights could be formulated while inspecting the two models built: i) the predictors selected were the same on both regressions, while on the model with **Gradient** the interaction between this variable and **MonthlyPhlebotomies** was considered and seemed more statistically significant, ii) the R-squared value was higher for the model with **Gradient** (0.742) than the model with **kParameter** (0.642) by one unit, iii) after the addition of **AgeDiagnosis** to the model with **Gradient**, both **Sex** and **Genotype** additions seemed significant, albeit not raising much the R-squared, but were not considered to keep the model simpler and interpretable and iv) the model with **Gradient** seemed to have more issues regarding data homoscedasticity and residuals' normality.

3.5.1 Outlier detection

After building the two regression models discussed before, an analysis of influential points was performed. For this, the Cook's distance was taken into account to identify possible outliers. This method is an influence measure based on the difference between the regression estimates and their value if the i^{th} observation is deleted [55]. These differences are measured for all the data points and the influential points are identified when their Cook's distance is sufficiently high to influence the fitted values of the regression [55]. In this work, the function *ols_plot_cooksd_bar()* from the R package *olsrr* was used for this purpose [45]. Figure 3.45 shows the output of this function, in which it is possible to visualize the influential points found for both regression models built previously. Eighteen (18) influential data points were discovered on the model with the **kParameter** variable and fifteen (15) on the model with the **Gradient** variable.



(a)

(b)

Figure 3.45: Cook's distance plots to detect influential data points for (a) linear regression with the kParameter variable and (b) linear regression with the Gradient variable. Data points above red threshold line are considered influential points

The patients regarded as influential points are presented in Table A.22 and A.23. A preliminary examination of these tables seemed to suggest that: i) patients with extreme **Ferritin** values were deemed as influential, ii) patients with extreme **DurationWeeks** were deemed as influential, iii) patients with extreme **kParameter** values were considered influential, especially on the regression with this variable and iv) patients with extreme **Gradient** values were regarded as influential, especially on the regression with this variable. These findings suggest that regressions without these influential data points may fail on the prediction of the **DurationWeeks** for patients with extreme values of **Ferritin**, **DurationWeeks**, **kParameter** and **Gradient**.

However, the increased R-squared of the two models without these influential points suggest that these regressions may be more capable of explaining the variability of the outcome variable than models with all the data points. Indeed, the linear regression with the **kParameter** without influential points had an R-squared of 0.715 and with these points the R-squared was 0.642, while the linear regression with the **Gradient** variable without the influential points had an R-squared of 0.834 and with these points the R-squared was 0.742 (Table 3.21). These linear regression equations with the associated estimates' changes are presented in Equation 3.3 and 3.4. Although both Residuals vs Fitted and Scale-Location plots seemed to suggest a similar pattern to the one identified on the regression models with data from all the patients, these plots regarding the model with **Gradient** without the influential points seemed to show a slightly better approximation to data variance homogeneity (Figures A.17 and A.18).

$$\begin{aligned}(DurationWeeks_i)^{0.2} = 1.956 + 0.0002 \cdot Ferritin_i + 28.212 \cdot kParameter_i \\ + 0.085 \cdot MonthlyPhlebotomies_i + 0.003 \cdot AgeDiagnosis_i + \epsilon_i \end{aligned} \tag{3.3}$$

$$\begin{aligned}(DurationWeeks_i)^{0.2} = 1.792 + 0.0004 \cdot Ferritin_i + 0.011 \cdot Gradient_i \\ + 0.213 \cdot MonthlyPhlebotomies_i + 0.007 \cdot Gradient_i \cdot MonthlyPhlebotomies \\ + 0.002 \cdot AgeDiagnosis_i + \epsilon_i \end{aligned} \tag{3.4}$$

Table 3.21: Summary of linear regression models, one with kParameter and another with Gradient, on data without respective influential data points

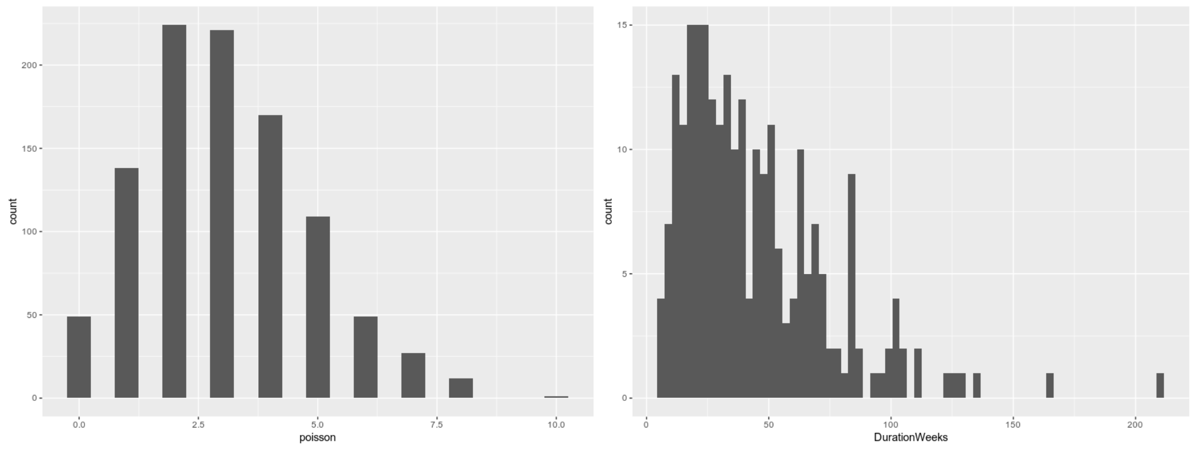
	Dependent variable:	
	(DurationWeeks)ˆ0.2	
	(1)	(2)
Ferritin	0.0002^{***}	0.0004^{***}
	(0.00002)	(0.00002)
kParameter	28.212^{***}	
	(1.968)	
Gradient		0.011^{***}
		(0.001)
MonthlyPhlebotomies1	0.085^{***}	0.213^{***}
	(0.022)	(0.029)
AgeDiagnosis	0.003^{***}	0.002^{***}
	(0.001)	(0.001)
Gradient:MonthlyPhlebotomies1		0.007^{***}
		(0.001)
Constant	1.956^{***}	1.792^{***}
	(0.050)	(0.035)
Observations	228	231
R^2	0.715	0.834
Adjusted R^2	0.709	0.831
Residual Std. Error	0.143 (df = 223)	0.110 (df = 225)
F Statistic	139.538^{***} (df = 4; 223)	226.578^{***} (df = 5; 225)
Note:		*p<0.1; **p<0.05; ***p<0.01

3.6 Generalized Linear Models

On this section, GLMs, specifically Poisson Regression and Negative Binomial Regression (NB) are attempted to model the duration of the DP. GLMs extend the regression models to other data distributions besides the Gaussian, like the Poisson distribution, allowing to fit skewed distributions and non-constant variance [62]. A similar approach to the one endured for the Linear Regressions with data transformation was followed to build Poisson and NB regression models. However, the analysis performed to select the final Poisson and NB regression models was not as thoroughly discussed.

3.6.1 Poisson Regressions

The Poisson Regression is a common technique to model count data of objects or events [39, 54]. These models can be used when the outcome variable distribution is entirely positive, as it is the case for the **DurationWeeks**. A random simulation of a Poisson distribution and the actual distribution of the target variable, **DurationWeeks**, are depicted in Figure 3.46. Indeed, the distribution of the outcome variable of this study seemed to be relatively approximate to a Poisson distribution, in which a right tail is present.



(a)

(b)

Figure 3.46: Random simulation of a Poisson distribution (a) and the distribution of DurationWeeks (b)

The same criteria used with Linear Regressions was used to select the exploratory variables on a stepwise addition fashion. Besides assessing the statistical significance of each variable addition, the AIC was used for model selection, taking into account decreases on this value for model comparison. As such, the variable considered in each addition was the one with higher statistical significance and that lead to an higher decrease on the AIC. Following this procedure resulted on the Equation 3.5 for a model with the **kParameter** and Equation 3.6 for a model with the **Gradient** variable. Both models are summarized in Table 3.22. Unlike the final linear regressions considered previously, the Poisson regression with the **kParameter** variable has the **Genotype** included as a predictor and the Poisson regression with the **Gradient** variable has **Genotype** and **Sex** as explanatory variables. Some degree of data heteroscedasticity was found on the Residuals vs Fitted and Scale-Location of these two models, in which higher fitted values seem to be more associated with more data variance (Figures A.19 and A.20).

$$\begin{aligned}\log E(DurationWeeks_i) &= 3.331 + 0.0003 \cdot Ferritin_i + 64.895 \cdot kParameter_i \\ &+ 0.326 \cdot MonthlyPhlebotomies_i + 0.006 \cdot AgeDiagnosis_i + 0.151 \cdot GenotypeHomozygous_i\end{aligned} \tag{3.5}$$

$$\begin{aligned}\log E(DurationWeeks_i) &= 2.991 + 0.001 \cdot Ferritin_i + 0.026 \cdot Gradient_i \\ &+ 0.441 \cdot MonthlyPhlebotomies_i + 0.012 \cdot Gradient_i \cdot MonthlyPhlebotomies \\ &+ 0.006 \cdot AgeDiagnosis_i + 0.151 \cdot GenotypeHomozygous_i + 0.155 \cdot Sexm_i\end{aligned} \tag{3.6}$$

Moreover, these two regression models were assessed with the data without the influential data points found before. The estimates of these Poisson regressions were updated on Equation 3.7 (model with **kParamter**) and Equation 3.8 (model with **Gradient**) and both models are summarized on Table 3.23. Inspection of the Residuals vs Fitted and Scale-Location plots for these models suggested that they seem to approximate better the data to homoscedasticity than the Poisson regressions with all the data points, especially for the case of the regression with the **Gradient** variable (Figures A.21 and A.22).

$$\begin{aligned}\log E(DurationWeeks_i) &= 3.599 + 0.0003 \cdot Ferritin_i + 86.473 \cdot kParameter_i \\ &+ 0.229 \cdot MonthlyPhlebotomies_i + 0.006 \cdot AgeDiagnosis_i + 0.091 \cdot GenotypeHomozygous_i\end{aligned} \tag{3.7}$$

$$\begin{aligned}\log E(DurationWeeks_i) &= 3.008 + 0.001 \cdot Ferritin_i + 0.032 \cdot Gradient_i \\ &+ 0.456 \cdot MonthlyPhlebotomies_i + 0.014 \cdot Gradient_i \cdot MonthlyPhlebotomies \\ &+ 0.005 \cdot AgeDiagnosis_i + 0.133 \cdot GenotypeHomozygous_i + 0.099 \cdot Sexm_i\end{aligned} \tag{3.8}$$

Table 3.22: Summary of Poisson regression models, one with kParameter and another with Gradient, on data with all the patients

	Dependent variable:	
	DurationWeeks	
	(1)	(2)
Ferritin	0.0003***	0.001***
	(0.00001)	(0.00001)
kParameter	64.895***	
	(2.515)	
Gradient		0.026***
		(0.001)
MonthlyPhlebotomies1	0.326***	0.441***
	(0.022)	(0.040)
AgeDiagnosis	0.006***	0.006***
	(0.001)	(0.001)
Sexm		0.155***
		(0.023)
Gradient:MonthlyPhlebotomies1		0.012***
		(0.002)
GenotypeHomozygous	0.151***	0.224***
	(0.030)	(0.029)
Constant	3.331***	2.991***
	(0.058)	(0.065)
Observations	246	246
Log Likelihood	−1,467.926	−1,183.839
Akaike Inf. Crit.	2,947.852	2,383.678
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3.23: Summary of Poisson regression models, one with kParameter and another with Gradient, on data without influential data points

	Dependent variable:	
	DurationWeeks	
	(1)	(2)
Ferritin	0.0003***	0.001***
	(0.00002)	(0.00002)
kParameter	86.473***	
	(3.090)	
Gradient		0.032***
		(0.001)
MonthlyPhlebotomies1	0.229***	0.456***
	(0.024)	(0.046)
AgeDiagnosis	0.006***	0.005***
	(0.001)	(0.001)
Sexm		0.099***
		(0.024)
Gradient:MonthlyPhlebotomies1		0.014***
		(0.002)
GenotypeHomozygous	0.091***	0.133***
	(0.030)	(0.030)
Constant	3.599***	3.008***
	(0.064)	(0.070)
Observations	228	231
Log Likelihood	−1,127.593	−911.608
Akaike Inf. Crit.	2,267.187	1,839.216
Note:	*p<0.1; **p<0.05; ***p<0.01	

3.6.2 Negative Binomial Regressions

The NB regression, akin to the Poisson regressions, can be used to model count data [48]. Although, the traditional NB is derived from a Poisson-gamma distribution [48]. Unlike the Poisson regression, that assumes a Poisson distribution with equidispersion (equal mean and variance), the NB may be more appropriate for data with overdispersion, in which the variance is greater than the mean [39, 48]. Indeed, in the NB regression, a random term to reflect unexplained between-subject differences and thus, account for unexplained variance, is introduced on the regression models [39]. Regarding the target variable of this study, **DurationWeeks**, it seemed that its variance (≈ 903) is much greater than its mean (≈ 44). Hence, NB regressions were considered as a possible modeling strategy.

Again, a similar stepwise variables' addition procedure was endured. The final model considered with the **kParameter** variable is displayed in Equation 3.9 and the final model with the **Gradient** variable is shown in Equation 3.10. These two models are summarized in Table 3.24. The same predictors as the ones included on the Poisson regression were considered for the NB model with the **Gradient** variable, while on the NB regression with the **kParameter** the **Genotype** variable was not found to be statistically significant contrarily to what was discovered with the Poisson regression with the **kParameter**. Similar patterns to the ones observed for the Poisson regressions were found on Residuals vs Fitted and Scale-Location plots of NB regressions (Figures A.23 and A.24).

$$\begin{aligned}\log E(DurationWeeks_i) = 3.363 + 0.0004 \cdot Ferritin_i + 63.594 \cdot kParameter_i \\ + 0.278 \cdot MonthlyPhlebotomies_i + 0.006 \cdot AgeDiagnosis_i \end{aligned} \tag{3.9}$$

$$\begin{aligned}\log E(DurationWeeks_i) = 2.834 + 0.001 \cdot Ferritin_i + 0.025 \cdot Gradient_i \\ + 0.444 \cdot MonthlyPhlebotomies_i + 0.012 \cdot Gradient_i \cdot MonthlyPhlebotomies \\ + 0.006 \cdot AgeDiagnosis_i + 0.178 \cdot GenotypeHomozygous_i + 0.169 \cdot Sexm_i \end{aligned} \tag{3.10}$$

Further, these two models were assessed with the data without influential data points. The NB regressions with the new estimates are presented in Equation 3.11 (model with the **kParameter** variable) and in Equation 3.12 (model with the **Gradient** variable). Also, the models are summarized in Table 3.25 and the diagnostic plots are shown in Figures A.25 and A.26. Equivalently to what was assessed through the examination of the Residuals vs Fitted and Scale-Location plots of the Poisson regressions without influential data points, it seems that the NB regression with the **Gradient** variable was the one that approximated better the data to homoscedasticity.

$$\begin{aligned}\log E(DurationWeeks_i) = 3.495 + 0.0004 \cdot Ferritin_i + 78.604 \cdot kParameter_i \\ + 0.206 \cdot MonthlyPhlebotomies_i + 0.007 \cdot AgeDiagnosis_i\end{aligned} \tag{3.11}$$

$$\begin{aligned}\log E(DurationWeeks_i) = 2.906 + 0.001 \cdot Ferritin_i + 0.030 \cdot Gradient_i \\ + 0.474 \cdot MonthlyPhlebotomies_i + 0.016 \cdot Gradient_i \cdot MonthlyPhlebotomies \\ + 0.005 \cdot AgeDiagnosis_i + 0.099 \cdot GenotypeHomozygous_i + 0.097 \cdot Sexm_i\end{aligned} \tag{3.12}$$

Table 3.24: Summary of NB regression models, one with kParameter and another with Gradient, on data with all the patients

	Dependent variable:	
	DurationWeeks	
	(1)	(2)
Ferritin	0.0004^{***} (0.00003)	0.001^{***} (0.00003)
kParameter	63.594^{***} (4.957)	
Gradient		0.025^{***} (0.002)
MonthlyPhlebotomies1	0.278^{***} (0.056)	0.444^{***} (0.078)
GenotypeHomozygous		0.178^{***} (0.055)
AgeDiagnosis	0.006^{***} (0.002)	0.006^{***} (0.002)
Sexm		0.169^{***} (0.049)
Gradient:MonthlyPhlebotomies1		0.012^{***} (0.003)
Constant	3.363^{***} (0.120)	2.834^{***} (0.127)
Observations	246	246
Log Likelihood	−1,001.686	−953.107
θ	8.137^{***} (0.876)	13.772^{***} (1.683)
Akaike Inf. Crit.	2,013.371	1,922.215
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3.25: Summary of NB regression models, one with kParameter and another with Gradient, on data without influential data points

	Dependent variable:	
	DurationWeeks	
	(1)	(2)
Ferritin	0.0004^{***}	0.001^{***}
	(0.00004)	(0.00004)
kParameter	78.604^{***}	
	(5.195)	
Gradient		0.030^{***}
		(0.002)
MonthlyPhlebotomies1	0.206^{***}	0.474^{***}
	(0.051)	(0.070)
GenotypeHomozygous		0.099^{**}
		(0.046)
AgeDiagnosis	0.007^{***}	0.005^{***}
	(0.002)	(0.001)
Sexm		0.097^{**}
		(0.040)
Gradient:MonthlyPhlebotomies1		0.016^{***}
		(0.003)
Constant	3.495^{***}	2.906^{***}
	(0.119)	(0.107)
Observations	228	231
Log Likelihood	−890.822	−837.622
θ	12.374^{***} (1.517)	28.590^{***} (4.442)
Akaike Inf. Crit.	1,791.643	1,691.244
Note:		*p<0.1; **p<0.05; ***p<0.01

3.7 Model evaluation and comparison

After establishing the final regression models, it was crucial to evaluate their performance. The K-fold Cross Validation (CV) technique was used to evaluate the predictive power of the models considered previously. Indeed, K-fold CV has been widely used for model selection in regression tasks and works by dividing the data set in random and evenly K parts [52]. Then, the K - 1 parts (training set) are used to train the model, while the remaining part (test set) is used to evaluate the model, by predicting its values and comparing these predictions to the true values under a given metric [8, 52]. This method is performed on a iterative fashion, in which each of the K parts are used as the test set. Finally, the values obtained for the error metric chosen for each iteration is averaged [8]. The metric used in this analysis to assess models performance is the Mean Absolute Percentage Error (MAPE). This quality measure is commonly used when the quantity to predict is known to be above zero and has the advantages of being scale-independent and relatively easy to interpret [28, 56]. Actually, MAPE is the average of the absolute percentage errors, meaning that lower percentages are associated with more accurate predictions [56].

So, the models evaluation consisted on performing a 10-fold CV with the MAPE as the quality measure. This procedure was carried on the linear regression, Poisson regression and NB regression models, some with the **kParameter** and some with the **Gradient** variables as predictors, trained with the data set with all the data and with the data set without influential data points, and tested on data of patients with the real **kParameter** and **Gradient** values (Table 3.26). This allows to gain insights regarding the predictive accuracy of these models on an ideal setting in which the true **kParameter** and **Gradient** values are available, allowing to compare with models tested on simulated newly diagnosed patients. Indeed, the same procedure was endured to all of the models but the test set values for the **kParameter** and **Gradient** variables were replaced with the corresponding values of **kParameter2** and **Gradient2**, allowing to judge the models predictive power on newly diagnosed patients (Table 3.27).

Generally, the three regression techniques seemed to achieve similar MAPE values, albeit the linear regressions with data transformation seemed to have at least slightly lower MAPE values overall. In fact, Poisson and NB regressions seemed to have more similar results. The models built with the data set without influential data points seemed to have an increased performance, especially for models in which the **Gradient** variable was used, where some had reductions of $\approx 10\%$ in the MAPE. Also, it seemed that reductions in the MAPE on models with these data sets were more pronounced when the test set had the real **kParameter** and **Gradient** values. Furthermore, MAPE values seemed to duplicate, or almost duplicate, when the **kParameter2** and **Gradient2** values were considered on the test set for the models with the **Gradient** as an explanatory variable. While the models with this predictor seemed to have a better predictive power when the test set had the true values of this variable, models with the **kParameter** had lower MAPE values when the assigned **kParameter2** values were considered on the test set. In reality, these results seem to suggest that, at this point, for a newly diagnosed patient, models with the **kParameter** may be preferable. Indeed, the models that may be currently used on newly diagnosed patients that had lower MAPE values were the linear regressions with the **kParameter** variable as a predictor, on both data sets assessed. Yet, the lowest MAPE values obtained with test sets with the real **kParameter** and **Gradient** values were with regressions with the **Gradient** as an explanatory variable, indicating that a better characterization of these values for newly diagnosed patients may lead to MAPE values in the range of ≈ 20% to ≈ 30%, depending on the inclusion of influential points on the data set. Thus, even if the **Gradient2** values assigned were more varied (twelve) than the ones attributed on **kParameter2** (four), these findings suggest that the characterization of the **Gradient** values is not as accurate.

Table 3.26: 10-fold cross-validation results using the MAPE metric for all the final models considered, in which the test set has the real patient specific values for the variables that describe the decay of SF over time (kParameter and Gradient)

Metric and data	Regression type					
	Linear		Poisson		NB	
	Gradient	k	Gradient	k	Gradient	k
MAPE						
all data	30.2%	36.4%	34.2%	39.5%	31.4%	38.6%
no influential points	21.2%	30.5%	24.0%	31.9%	22.3%	31.1%

Table 3.27: 10-fold cross-validation results using the MAPE metric for all the final models considered, in which the test set has the assigned k and gradient values (kParameter2 and Gradient2)

Metric and data	Regression type					
	Linear		Poisson		NB	
	Gradient2	k2	Gradient2	k2	Gradient2	k2
MAPE						
all data	59.6%	49.2%	63.7%	54.5%	64.2%	53.1%
no influential points	54.1%	46.8%	55.8%	50.9%	55.3%	49.3%

As stated before, the model that seemed to perform better for newly diagnosed patients was the linear regression with data transformation, with the **kParameter** as a predictor, built with the data set without the influential data points (n = 228). Thus, this model was used to simulate an hypothetical situation in which the model would try to predict the duration of DP of some newly diagnosed patients, with the assigned **kParameter2** values. The regression was trained with 90% of the patients (n = 205), leaving 10% for the test set (n = 23). The MAPE obtained was of 34.9%, a reduction of ≈ 12% compared to what was found with the averaging process of the K-fold CV.

The prediction values and the true values of the duration in weeks of the DP can be found in Figure 3.47. While in this test set a large number of the predictions seemed to be relatively close to the real duration, some of them had predictions with an elevated error. Indeed, some predictions seemed to have a reasonable error of less than ten weeks, but other predictions had errors of more than fifty weeks, which is excessively high. Generally, the plot seems to suggest that patients with a real **DurationWeeks** value close to the mean (≈ 43) or the median (≈ 37) of the data were associated with the lowest prediction errors, while patients with more extreme values of durations (≈ 100 weeks) tend to have higher prediction errors.

Furthermore, the prediction intervals (95% confidence) were also determined and are displayed in Table 3.28, along with the values of the variables on the data set, and seem to be in line with the previous hypothesis. In fact, patients with large errors seem to be patients whose real durations are closer to the upper bound of the prediction interval, revealing that the regression model is predicting relatively lower values for these individuals. Although, there seems to be an exception in which one of the patients (number 311) had a real duration of 122 weeks and the regression model predicted 146 weeks. Even if the error is higher than twenty weeks, it does not seem to compromise the model as much as an error of the same magnitude for an individual like patient number 37, whose true duration is 42 weeks and the model predicted only 16 weeks. Overall, the prediction intervals seem to be very large, covering a large period of time, which may reflect that they may not be helpful enough as a prediction tool to assess a time window for the end of the DP for the majority of the HH patients.

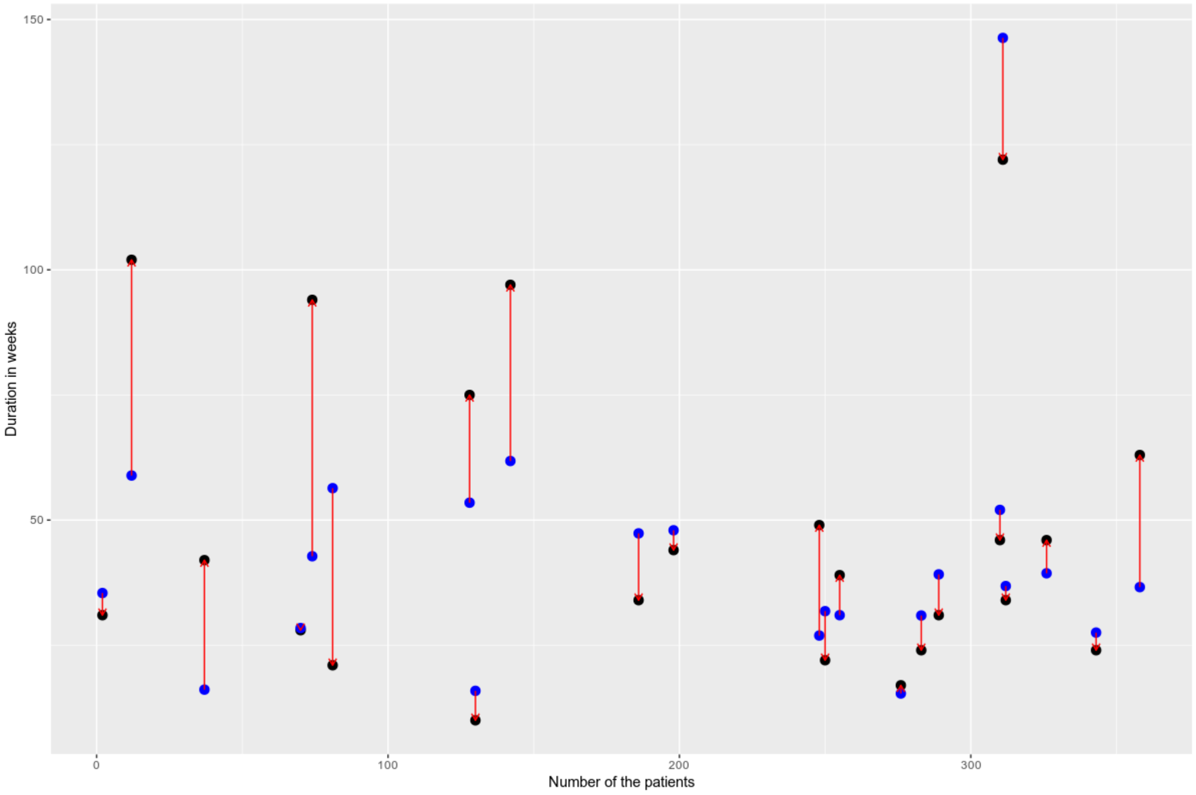


Figure 3.47: Plot showing the predicted values (blue circles) and the true values (black circles) of the duration in weeks, in which arrows link the corresponding predictions to their true values. The model used was the linear regression with data transformation, with the kParameter as a predictor, trained on 90% of the data set without influential data points and with the assigned kParameter2 and Gradient2 values

Table 3.28: Patients from test set, consisting of 10% of the data set without influential data points and with assigned kParameter2 and Gradient2 values, with predictions (fit) and prediction intervals (lower and upper bounds) using the linear regression with data transformation with the kParameter as an explanatory variable

Number	DurationWeeks	fit	lwr	upr	Ferritin	Group	kParameter2	MonthlyPhlebotomies	AgeDiagnosis	Sex	Genotype	Gradient2
2	31	35	17	68	1125	Group 1000	-0.0087	>1	46	m	Homozygous	-40.34
12	102	59	30	107	1014	Group 1000	-0.0055	1	70	m	Homozygous	-20.23
37	42	16	7	35	748	Group 500	-0.0169	>1	46	m	Heterozygous	-25.25
70	28	28	13	56	608	Group 500	-0.0087	>1	50	m	Homozygous	-25.25
74	94	43	21	81	375	Group 0	-0.0055	1	64	f	Homozygous	-5.15
81	21	56	28	103	1905	Group 1000	-0.0087	>1	63	m	Homozygous	-40.34
128	75	53	27	99	1887	Group 1000	-0.0087	>1	56	m	Homozygous	-40.34
130	10	16	6	34	482	Group 0	-0.0169	>1	62	m	Heterozygous	-18.37
142	97	62	32	112	1441	Group 1000	-0.0055	1	49	m	Homozygous	-20.23
186	34	47	23	88	1774	Group 1000	-0.0087	>1	45	f	Homozygous	-28.19
198	44	48	24	89	899	Group 500	-0.0055	1	46	f	Homozygous	-13.66
248	49	27	12	54	842	Group 500	-0.0087	>1	27	m	Homozygous	-25.25
250	22	32	15	62	719	Group 500	-0.0087	>1	58	f	Homozygous	-22.71
255	39	31	14	61	481	Group 0	-0.0071	1	26	m	Heterozygous	-7.13
276	17	15	6	33	674	Group 500	-0.0169	>1	45	m	Heterozygous	-25.25
283	24	31	14	61	1214	Group 1000	-0.0087	>1	21	m	Homozygous	-40.34
289	31	39	19	75	647	Group 500	-0.0071	1	48	m	Heterozygous	-13.10
310	46	52	26	96	980	Group 500	-0.0055	1	53	m	Homozygous	-13.10
311	122	146	79	252	4175	Group 1000	-0.0087	>1	75	m	Homozygous	-40.34
312	34	37	17	71	348	Group 0	-0.0055	1	44	m	Homozygous	-7.13
326	46	39	19	75	614	Group 500	-0.0071	1	51	m	Heterozygous	-13.10
343	24	28	12	55	408	Group 0	-0.0087	>1	59	m	Homozygous	-18.37
358	63	37	17	70	395	Group 0	-0.0055	1	40	m	Homozygous	-7.13

Chapter 4

Discussion

In this work, we conducted an extensive correlation analysis of some variables related to HFE-associated HH patients, trying to understand their influence on the duration of the DP. Besides, regression models were built to estimate the duration of the DP for newly diagnosed patients. This study seemed to unveil relevant information regarding the DP of HFE-related HH patients. Initially, the data was processed to obtain the SF levels during the DP of each patient to study their behavior. Feature extraction allowed to create two variables to model the decay of SF during this treatment stage (**kParameter** and **Gradient**) and one variable reflecting an approximation of the frequency of phlebotomies per month (**MonthlyPhlebotomies**). It was found that a large number of patients did not have enough data to establish an end point of the DP or to characterize exponentially the SF decay. Also, some patients seemed to have large phlebotomy interruptions. Dr. Annick Vanclooster, from the Leuven University Hospital, was consulted and suggested that this lack of usable data may be related to: i) some patients performing some phlebotomies in the hospital and some with a general practitioner, ii) some patients that did not start venesection treatment, either due to the decision of the treating physician or own choice, iii) some patients that may have skipped phlebotomies and did not reschedule and iv) some patients that were not in regular follow-up. After the data processing procedures assessed in this work, main recommendations for future data collection include: i) gather at least two time-stamped SF values to allow to determine the **Gradient** values, ii) gather at least three time-stamped SF values to allow to determine the **kParameter** values and iii) gather data with established dates of beginning of the DP and end of the DP, as in this work a SF concentration threshold of 100 μg/L had to be assumed to establish an end point for the DP.

Regarding the other variables, the majority has a large number of missing values and others were not deemed as relevant for this study. Even then, some of the initially investigated variables were not included in the main analysis. The fact that they have missing values, even if they seemed clinically relevant, was decisive to disregard them for this work. Also, unbalanced sample sizes, especially for the variables related to liver disease and alcohol consumption, were another negative evidence verified. While for the case of liver disease values it may be related to actual biological findings, as the patients with cirrhosis may be naturally outnumbered, the alcohol consumption values were too categorized, with very unbalanced categories, which may be associated with poor data collection. In reality, a better data collection approach may solve these issues and allow to explore further the relevance of these variables. In fact, studies have shown that cirrhotic patients may have an elevated rate of iron mobilization during depletion treatment [2].

Moreover, the time-stamped biochemical parameters, besides the SF, considered for an initial analysis had missing values on their first data point (at diagnosis). A study of their levels as time-series during the DP was not assessed in this thesis but, eventually, may aid unveiling insights regarding their behavior during this therapy. Having the values for these parameters over time, for the corresponding dates associated with the SF values, may be of utmost importance to perform a similar analysis to the one endured for the SF levels. Nonetheless, an initial correlation assessment with these variables, on a very limited data set ($n = 55$), suggested that they may add value to this study, assuming that an enhanced data collection procedure is performed to avoid missing values either on static and time-stamped variables. Regarding these discarded variables, the major correlation found with the target variable was a positive correlation with the initial TS level (**Transferrin** variable).

Furthermore, a comprehensive correlation study was performed on the variables considered for the main analysis. Differences between groups of the initial SF value (**Group** variable) and the duration of the DP (**DurationWeeks**) were found, in which longer treatment periods are associated with increased threshold of SF at diagnosis (*Group 0* < *Group 500* < *Group 1000*). Also, the *Homozygous* genotype was found to be more associated with longer durations of the DP. These findings seem to be in line with what is described in the literature. Indeed, studies have shown that higher SF levels at diagnosis are associated with more treatment procedures during the DP [14, 91] and with iron-overload liver disease [4] and that the C282Y/C282Y genotype is more associated with increased SF and TS levels at diagnosis than compound heterozygotic genotypes [14, 88, 92]. Actually, the present study revealed that *Homozygous* patients are more associated with higher initial SF concentrations, as well as *male* individuals and older patients, relationships also described in the literature [14, 61]. Also, the number of phlebotomies performed may be highly related to the duration of this treatment stage, as it is expected that more blood withdrawals lead to more depletion of SF [4, 84]. This analysis revealed that patients that perform more than one phlebotomy per month are significantly more associated with shorter durations of the DP when compared to patients that perform one therapy per month.

Although, the frequency of the venesections may not be the only relevant factor when considering patient specific treatment plans. The volume of blood removed, which may reflect different rates of iron depletion, may be an important value to account for. Indeed, the volume of blood extracted may vary depending on the tolerance of the patient, its BMI, gender, age or risk of anemia [1, 4, 14]. Despite not having data available regarding the volume of blood removed for each patient, Dr. Annick Vanclooster, when consulted, stated that in the Leuven University Hospital the usual amount is 400 ml. Notwithstanding, having the patient specific values of blood removed in each phlebotomy may allow to, for instance, determine the average volume for each patient and, eventually, assess differences of iron depletion depending on the patients.

The **kParameter** variable and the **Gradient** variable were two patient specific characteristics extracted, aiming to obtain approximate values of the speed of decay of SF over time, modeling it exponentially and linearly, respectively. While the **Gradient** values can be interpreted as the average rate of depletion of SF per week, the **kParameter** values are related to the speed of decay of the exponential curve that defines each patients' time-stamped SF levels. Therefore, more negative values of these two parameters are associated with higher speeds of decay of SF during the DP and thus greater rates of SF depletion. On newly diagnosed patients, that do not have these *a posteriori* values, median values of these parameters were determined based on grouping the patients according to previously found statistically significant characteristics (**kParameter2** and **Gradient2**). As a matter of fact, these values were assigned on patients belonging to the test set, either when evaluating the models with 10-fold CV or a train/test split (90%/10%), to simulate predictions of newly diagnosed patients.

Beforehand, a variable selection procedure was assessed, on each regression type. The variables **Ferritin**, **MonthlyPhlebotomies** and **AgeDiagnosis** were statistically significant independently of the type of regression and of the type of approximation of the decay of SF used (**kParameter** and **Gradient**). In addition, **Genotype** and **Sex** were considered on some of the models. Overall, the three regression approaches tested - Linear, Poisson and NB regressions - obtained similar MAPE values between each other depending on the use of **kParameter** or **Gradient** and the use of the data set with all the patients or without the influential data points. Likewise, it seemed that the choice of the variable to approximate the behavior of the decay of SF (**kParameter** or **Gradient**), the data set used to train the models and the choice of values on the test set (either **kParameter** or **kParameter2** and **Gradient** or **Gradient2**) influenced more the MAPE values than the type of regression used. Regardless, the Linear regressions with data transformation, with $Lambda = 0.20$, seemed to have a slightly better performance (inferior MAPE) on all the different conditions examined. Also, in this study, same type of regressions but with different predictors were not compared through their predictive accuracy, as the approach was to establish a final set of predictors for each regression type and then test each one. Although it could reveal some insights regarding the impact of each variable on the prediction power, the number of possible combinations was considered extremely high.

Moreover, models built with the data set without the influential points seemed to have higher reductions on the MAPE value when tested on patients with their real **kParameter** and **Gradient** values (reductions between $\approx 6\%$ and $\approx 10\%$). However, testing on patients with the assigned **kParameter2** and **Gradient2** still resulted, generally, in decreases on the MAPE between $\approx 2\%$ and $\approx 5\%$, with two exceptions of $\approx 8\%$ and $\approx 11\%$, when comparing the use of the complete data set against the data set without the influential data points.

As expected, the predictive accuracy is superior when assessed on patients with their real **kParameter** and **Gradient** values, in which the MAPE is reduced between $\approx 15\%$ and $\approx 30\%$ when compared to test sets with the assigned values **kParameter2** and **Gradient2**, meaning that in some cases the MAPE duplicated. Also, models with the **kParameter** seemed to have more accurate predictions than the models with the **Gradient** when testing on newly diagnosed patients with the **kParameter2** and **Gradient2**, in which MAPE decreases between ≈ 5% and ≈ 10% were found. Altogether, these findings suggest that: i) the **kParameter2** may be more appropriate than the **Gradient2** at this point, ii) even if the **kParameter2** has only four different values and the **Gradient2** has twelve, it seems that more information was lost when characterising the **Gradient** of the patients, iii) characterising further these values for each patient may aid enhancing the predictive power of the models, perhaps even halve the MAPE. Notably, finding other significant variables to guarantee a better sub-grouping of these values may be one of the key tasks to assess in future work. In this thesis secondary analysis, it was found that **Transferrin** at diagnosis may be used to sub-group the **kParameter** values, as patients with TS *>45%* were more prone to have slower speeds of decay of SF over time. Furthermore, the ratio **ALT/AST** may also be relevant for the same task as both were found to be negatively correlated with the **kParameter** and the **Gradient** variables. Other static variables like the **BMI**, **AgeDiagnosis**, **LiverDisease**, **Alcohol** or **HB** should also be considered. Indeed, as discussed before, the volume of blood extracted may vary depending on the patients' **BMI**, **AgeDiagnosis** or possible risk of anemia, which can be assessed through the **HB** values, and cirrhotic patients were found to be associated with higher rates of SF depletion in previous studies [1, 2, 4, 14]. Essentially, to further explore these hypothesis, two key recommendations can be made: i) enhance data collection procedures to diminish the number of missing values and increase data quality and ii) perform an extensive study of the time-series levels of the biochemical parameters. Also, Decision Trees could be considered to predict **Gradient** and **kParameter** values, instead of the approach assessed in this work, or even to predict the duration of the DP, instead of the regressions tested.

Alternatively, to tackle this problem related to the characterisation of the **kParameter** and **Gradient** values, a predictive model may be taken into account to predict updated durations of the DP after a given patient has performed more than one phlebotomy. Besides the prediction at diagnosis, which is the main goal of this work, a patient could benefit from these models if more predictions were assessed mid-treatment, after some phlebotomies were performed, with updated **kParameter** and/or **Gradient** values instead of the use of the assigned **kParameter2** and/or **Gradient2**. Markedly, to obtain patient specific **kParameter** values, patients would need to have performed at least three phlebotomies and two therapeutic sessions to obtain the **Gradient** values. Eventually, this alternative approach could allow to compare the assigned **kParameter2** and **Gradient2** to real patient values of **kParameter** and **Gradient**, after some data collection, regarding their impact on the predictive models accuracy. Understanding if these regression models predict better durations of the DP of patients with some time-stamped SF data available may be of great relevance. Indeed, if proven true, it may mean that patients could have an enhanced prediction after enduring a given number of phlebotomies and eventually receive regular updates on the estimated end of the treatment. This could also aid the treating physician decide if the current treatment plan is appropriate or if the frequency of venesections should be altered.

In addition, the **MonthlyPhlebotomies** variable may function as a knob that allows to understand the influence of the frequency of the phlebotomies on the total time of the DP, aiding the physician on the decision-making. In the future, having more data regarding patients that performed more phlebotomies per month, especially weekly, may improve our understanding of the impact of the frequency of therapy sessions on the duration of the DP.

Ultimately, in this study, the more appropriate model found to predict the duration of the DP for a newly diagnosed patient was a Linear regression with Box-Cox transformation, using the **kParameter** as a predictor and using the data set without influential data points (MAPE = 46.8%, with 10-fold CV). Prediction intervals were also determined for this regression using 90% of the data to train the model and 10% to test it (MAPE = 34.9%). While this regression model seemed to have a reasonable accuracy prediction for the majority of the patients, in which fourteen patients had predictions with an error < 15 weeks, it seemed to present great errors for some of the patients, in which some had errors of ≈ 40 weeks. Also, the real duration values seemed to be within the prediction intervals on almost all the cases tested, except on three patients. Yet, these prediction intervals seemed excessively large, as some comprise prediction windows of approximately seventy weeks or more. Nevertheless, these prediction intervals could benefit a newly diagnosed patient as they indicate that with 95% of confidence: i) the DP will not end for at least the number of weeks determined on the lower bound and ii) the DP may last for the number of weeks determined on the upper bound. Equally, these prediction windows may aid the physician on the treatment planning. Undoubtedly, being able to determine shorter prediction intervals would increase these models usefulness for both the HFE-related HH patients and their treating physicians.

The variability in treatment scheduling recommended by the physician and patients' degree of commitment to it are two factors that may have influenced negatively the analysis performed in this study. Indeed, having data regarding patients to whom were recommended similar frequencies of phlebotomies and to which they committed without delays or rescheduling, could allow an improved understanding of the relevance of the variables and, eventually, lead to a model with enhanced predictive power. For instance, a randomized control trial to gather data from patients that perform weekly, every two weeks or no phlebotomies, would allow to collect equally spaced time-stamped data points and consequently remove the influence of the frequency of the therapeutic sessions and interruptions on the treatment on the total duration of the DP. Analyzing data from patients that have endured similar treatment plans may allow to understand better the influence of other static variables on the duration of the DP.

Chapter 5

Conclusion

In sum, this study allowed to unveil some correlations with the duration of the DP of type-1 HH patients. Indeed, higher SF concentration at diagnosis and the *homozygous* genotype seemed to be more associated with longer treatments. Besides the *homozygous* genotype, *male* patients and older individuals, at diagnosis, seemed to be related to higher SF levels at diagnosis. Also, we were able to characterise the decay of SF over time exponentially and linearly and discover that both approximations were correlated with the duration of the DP. The frequency of the therapeutic sessions was also found to influence the total duration of the treatment, in which patients that performed only one phlebotomy per month were more prone to increased treatment stages.

While building regression models, the SF values at diagnosis, the speed of decay of SF, either exponential or linear, the number of phlebotomies per month performed and the age at diagnosis seemed to be statistically significant characteristics. Some models also included the genotype and the sex of the patients as explanatory variables. Similar predictive accuracy power was found on the three types of regressions assessed. When disregarding influential data points, the predictive accuracy seemed to increased on all the models built. Moreover, the prediction errors seemed to increase greatly when testing on data from newly diagnosed patients, with assigned values to describe the speed of decay of SF, instead of the real ones. Nonetheless, the prediction intervals determined may aid on the physicians' decision making, by at least giving an estimate of the minimum and maximum duration of the DP. Thus, this study seemed to be an advantageous first step to predict the length of the DP for HFE-related HH patients.

Further analysis of other patients characteristics, with enhanced data collection techniques, may improve our understanding of the underlying factors affecting the duration of the HH iron depletion treatment. Hence, more accurate regression models could be built. In the future, a novel tool could be developed, taking advantage of these predictive models, to estimate the duration of the DP for type-1 HH patients, aiding the physicians on establishing patient specific treatment plans.

Appendix A

Appendix

Days	Ferritin
1	370
684	319
1026	281
1056	291
1116	254
1119	234
1164	257
1206	194
1234	172
1258	135
1272	107

Table A.1: Time-stamped SF values during the DP for a patient from Group 0 (initial SF value < 500 μg/L)

Variables in data set 1 with time-stamped biochemical parameters
Number
Date
Ferritin
Transferrin
Glycohemoglobine (HbA1c) in % (bloed) / %
Glycohemoglobine (HbA1c) in mmol/mol (bloed) / mmol/mol
Hepatitis A As (bloed)
Hepatitis A As (bloed) - % INH / % INH
Hepatitis A IgG (bloed)
Hepatitis A IgG (bloed) - kwantitatief / S/CO
Hepatitis B surface As (bloed) / mIU/mL
Hepatitis B surface As (bloed) - kwalitatief
Testosteron (bloed) / ng/dL
Vrij testosteron (bloed) / ng/dL
SHBG - Roche (bloed) / nmol/L
TSH (bloed) / mIU/L
AST
ID

Table A.2: Variables on the data set 1 that has information regarding time-stamped biochemical parameters

Variables in data set 2 with time-stamped biochemical parameters
Number
Date
HB
ALT

Table A.3: Variables on the data set 2 that has information regarding time-stamped biochemical parameters

Variables in data set 3
Number
Sex
birth year
Age Diagnosis
Deceased
Reason of Death
Year
Age...8
Smoking

number of pack years
Alcohol intake
Alcohol standardized
Weight (diagnosis)
Height(diagnosis)
BMI (diagnosis)
year diagnosis...16
blood donor
Family screening
Genotype
Diagnosis made via
year diagnosis...21
Biopsy
Fibrosis
Stage
made via
year diagnosis...26
Age...27
Cirrosis
Year diagnosis
Age...30
oesophagus varices
Ascites
Geelzucht (Bili >1mg/dl)
Encephalopathy
Protrombine tijd < 40%
HCC
year diagnosis...37
age...38
Diabetes
year diagnosis...40
age...41
transplantation
year...43
age...44
prothesis
year...46
age...47
what
fatigue
joint complaints
sexual dysfunction

Ferritine (diagnosis)
Tijd tot ferritine < 100 g/L (phlebotomies)
start phlebotomies
maintenance phase
Stop phlebotomy (year)
Aantal jaar AL
Hepatitis A IgG
Hepatitis B AS > 10
Pneumococcen (/5 jaar)
Proton Pump Inhibitors
Jaar start
Jaar stop
Andere
...65
Fibroscan_0-6m
kPa <7:OK_0-6m
Echo lever_0-6m
Echo leverres_0-6m
MRI/CT lever_0-6m
MRI/CT leverres_0-6m
Echo cardiores_0-6m...72
Echo cardiores_0-6m...73
BMC_0-6m
BMCres_0-6m
Hb_0-6m (g/dL)
Ferritine_0-6m (g/L)
Tf sat_0-6m (%)
AST_0-6m (U/L)
ALT_0-6m (U/L)
Glycemie_0-6m (mg/dL)
HbA1C_0-6m (%)
Testosteron_0-6m (ng/dL)
SHBG_0-6m (ng/dL)
SHBG_0-6m (nmol/L)
TSH_0-6 (mIU/L)
Vrij testosteron_0-6m
-FP_0-6m
Fibroscan_6-12m
kPa_6-12m
Echo lever_6-12m
Echo leverres_6-12m
MRI lever_6-12m

MRI leverres_6-12m
Echo cardio_6-12m
Echo cardiores_6-12m
BMC_6-12m
BMCres_6-12m
Hb_12m (g/dL)
Ferritine_12m (g/L)
Tf sat_6-12m (%)
AST_6-12m (U/L)
ALT_6-12m (U/L)
Glycemie_6-12m (mg/dL)
HbA1C_6-12m (%)
Testosteron_6-12m (ng/dL)
SHBG_6-12m (ng/dL)
SHBG_6-12m (nmol/L)
TSH_6-12m (mIU/L)
Vrij test_6-12m
-Fp_6-12m
Fibroscan_trsf
kPa_trsf
Jaar_trsf...114
Tijd sinds_Trsf diagnose
Echo Lever_trsf
Echo leverres_trsf
Jaar_trsf...118
Tijd sinds diagnose_tsfr...119
MRI lever_tsfr
Jaar_tsfr...121
Tijd sinds diagnose_tsfr...122
Echo cardio_tsfr
Echo cardiores_trsf
Jaar_tsfr...125
Tijd sinds diagnose_tsfr...126
BMC_trsf
BMCres_tsfr
Jaar_tsfr...129
Tijd sinds diagnose_tsfr...130
Hb (g/dL)
Ferritine_tsfr (g/L)
Tf sat_tsfr (%)
AST_tsfr (U/L)
ALT_tsfr (U/L)

Glycemie_tsfr (mg/dL)
HbA1C_tsfr (%)
Testosteron (ng/dL)_tsfr
SHBG_tsfr (ng/dL)
SHBG_tsfr (nmol/L)
TSH_tsfr (mIU/L)
Vrij test_tsfr
-Fp_tsfr
Fibroscan_dienst
kPa_dienst
Jaar_dienst...146
Tijd sinds diagnose_dienst...147
Echo lever_dienst
Echo leverres_dienst
Jaar_dienst...150
Tijd sinds diagnose_dienst...151
MRI lever_dienst
MRI leverres_dienst
Jaar_dienst...154
Tijd sinds diagnose_dienst...155
Echo cardio_dienst
Echo cardio_dienstres
Jaar_dienst...158
Tijd sinds diagnose_dienst...159
BMC_dienstopv
BMC_dienst
Jaar_dienst...162
Tijd sinds diagnose_dienst...163
Hb_dienst (g/dL)
Ferritine_dienst (g/L)
Tf sat_dienst (%)
AST_dienst (U/L)
ALT_dienst (U/L)
Glycemie_dienst (mg/dL)
HbA1C_dienst (%)
Testosteron_dienst (ng/dL)
SHBG_dienst (ng/dL)
SHBG_dienst (nmol/L)
TSH_dienst (mIU/L)
Vrij test_dienst
-Fp_dienst
jaar

Table A.4: Variables on the data set 3 that has information regarding static parameters

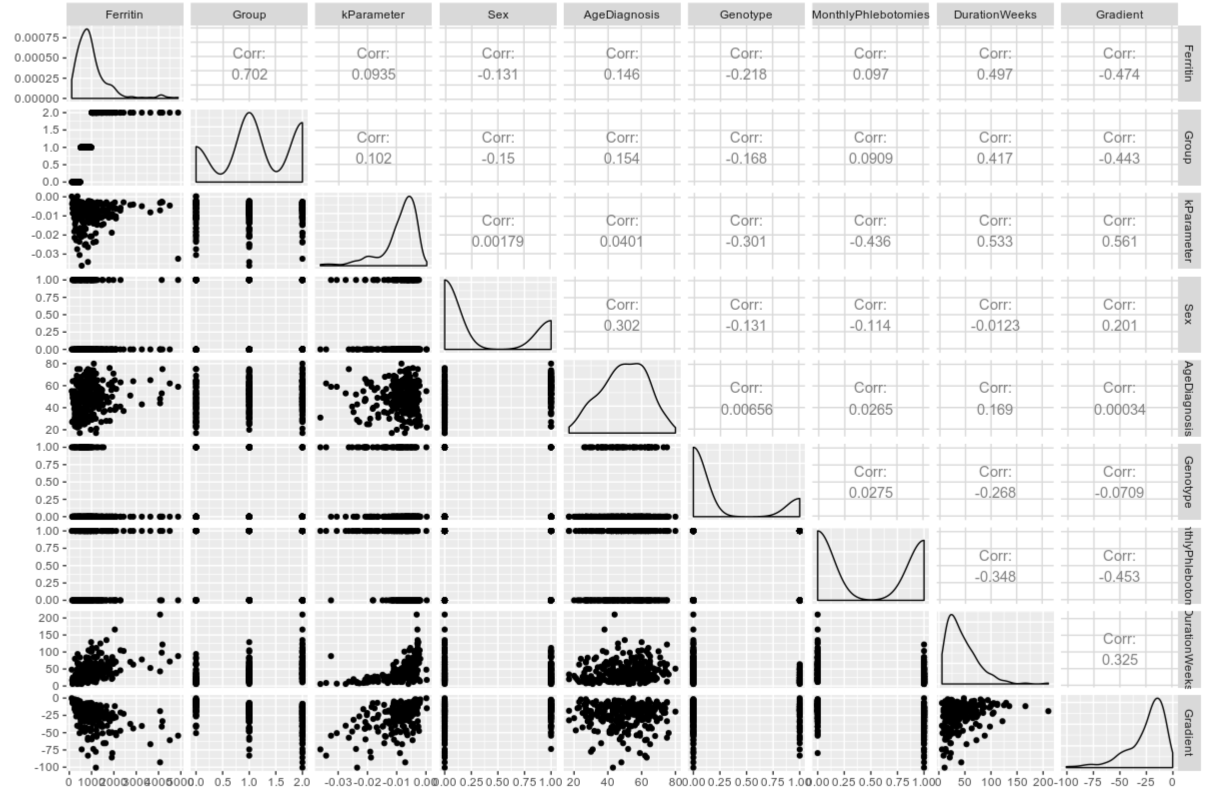


Figure A.1: Correlation plot showing scatter plots between variables and associated r correlation values

Group	DurationWeeksMedian	kMedian	FerritinMedian	GradientMedian	AgeMedian
Group 0	20	-0.0079	357	-9.20	44
Group 500	34	-0.0074	765	-18.30	52
Group 1000	54	-0.0067	1434	-28.36	50

Table A.5: Median values of the numerical variables based on Group of initial SF level

Genotype	DurationWeeksMedian	kMedian	FerritinMedian	GradientMedian	AgeMedian
Heterozygous	25	-0.0099	669	-19.19	49
Homozygous	39	-0.0067	907	-18.89	51

Table A.6: Median values of the numerical variables based on Genotype

Sex	DurationWeeksMedian	kMedian	FerritinMedian	GradientMedian	AgeMedian
f	37	-0.0067	714	-14.46	58
m	36	-0.0072	873	-21.37	48

Table A.7: Median values of the numerical variables based on Sex

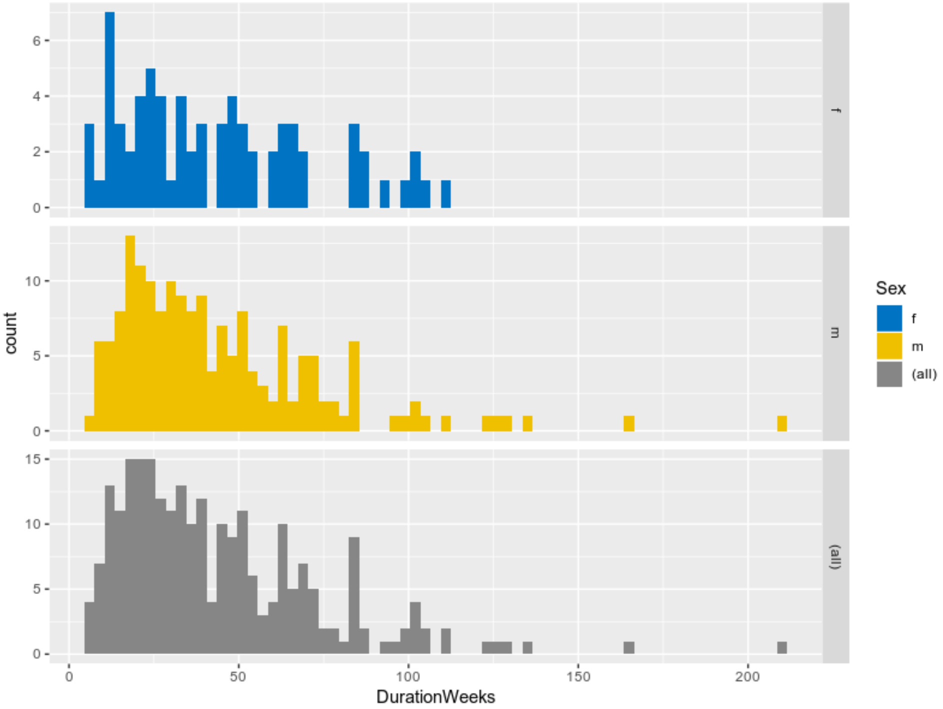


Figure A.2: Histogram of DurationWeeks values based on Sex

Sex	n
f	72
m	174

Table A.8: Sample sizes of each Sex

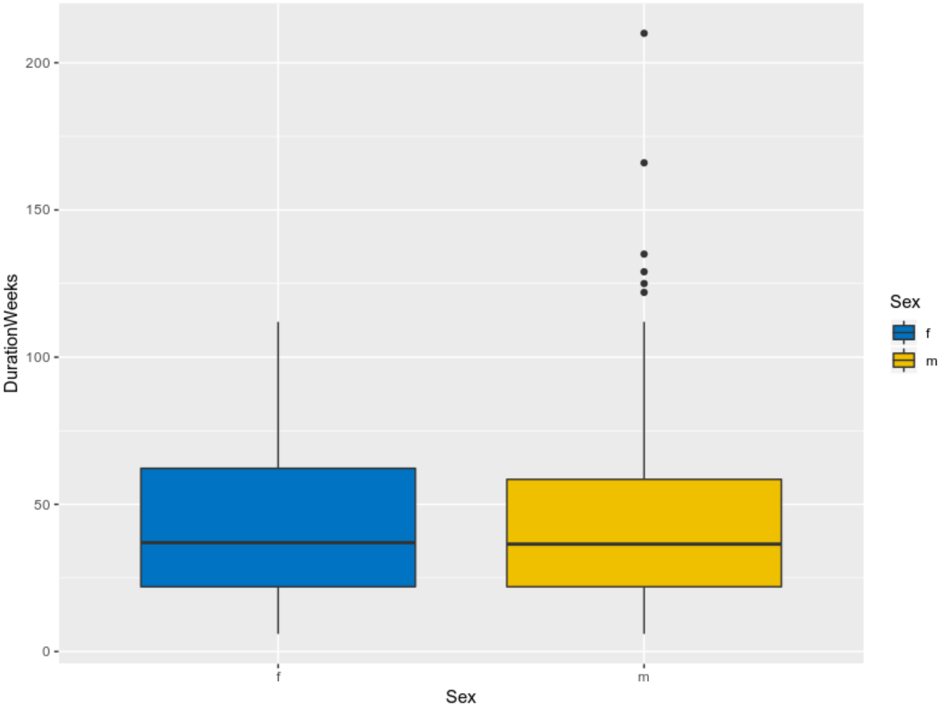


Figure A.3: Box plot of DurationWeeks values based on Sex

MonthlyPhlebotomies	DurationWeeksMedian	kMedian	FerritinMedian	GradientMedian	AgeMedian
>1	28	-0.0099	900	-27.95	50
1	48	-0.0058	808	-13.21	50

Table A.9: Median values of the numerical variables based on MonthlyPhlebotomies

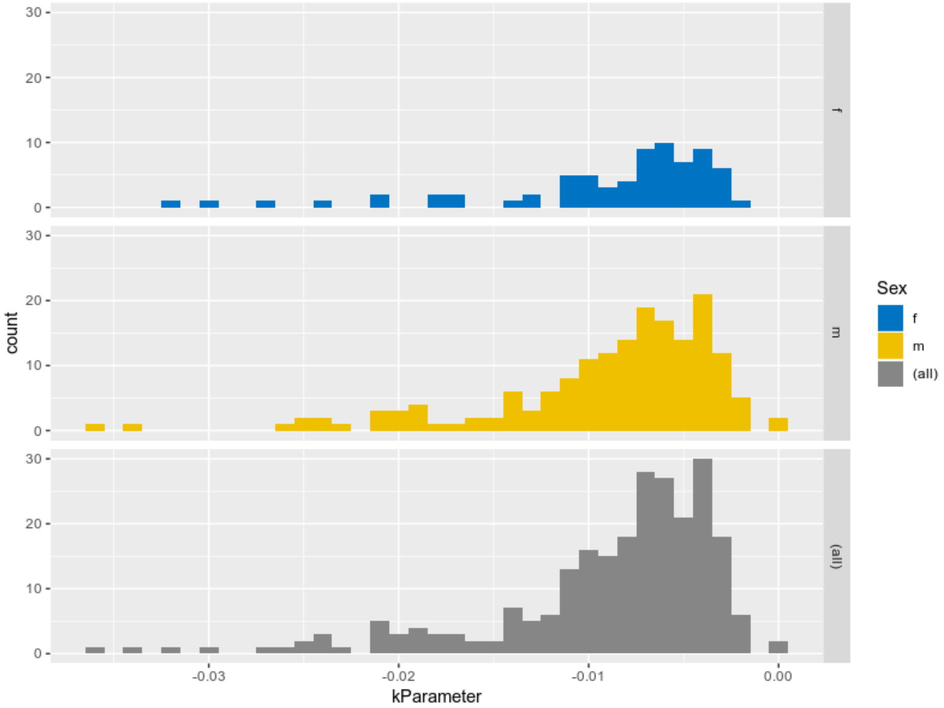


Figure A.4: Histogram of kParameter values based on Sex

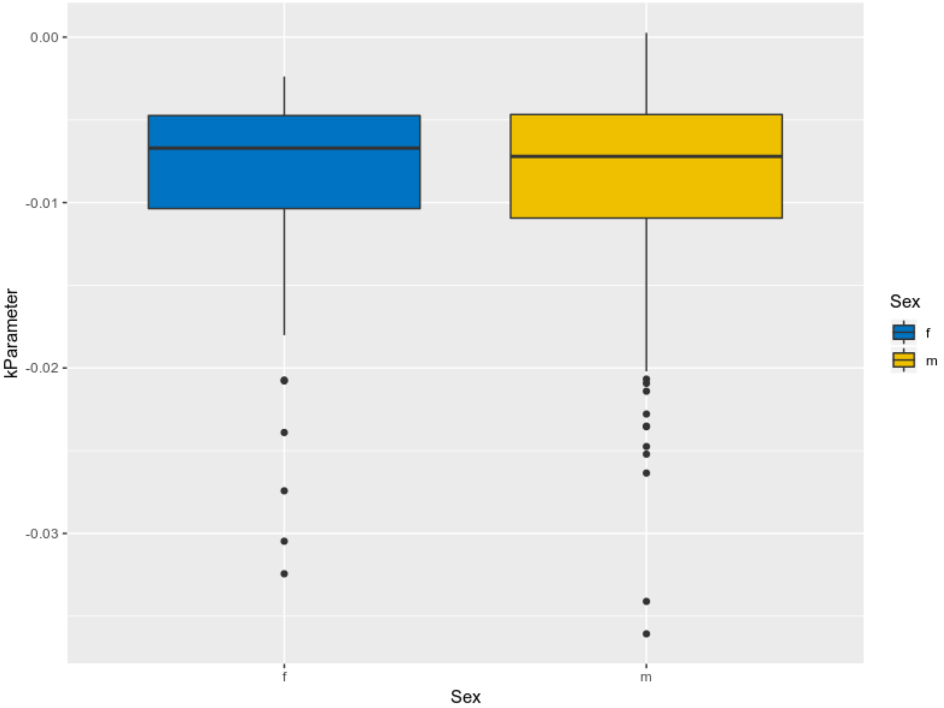


Figure A.5: Box plot of kParameter values based on Sex

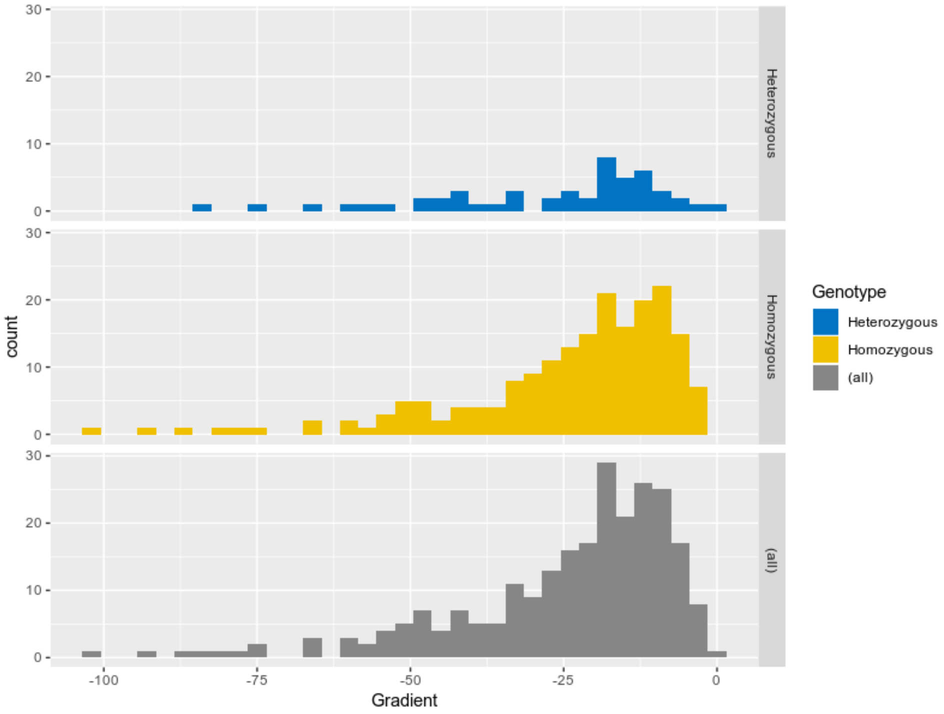


Figure A.6: Histogram of Gradient values based on Genotype

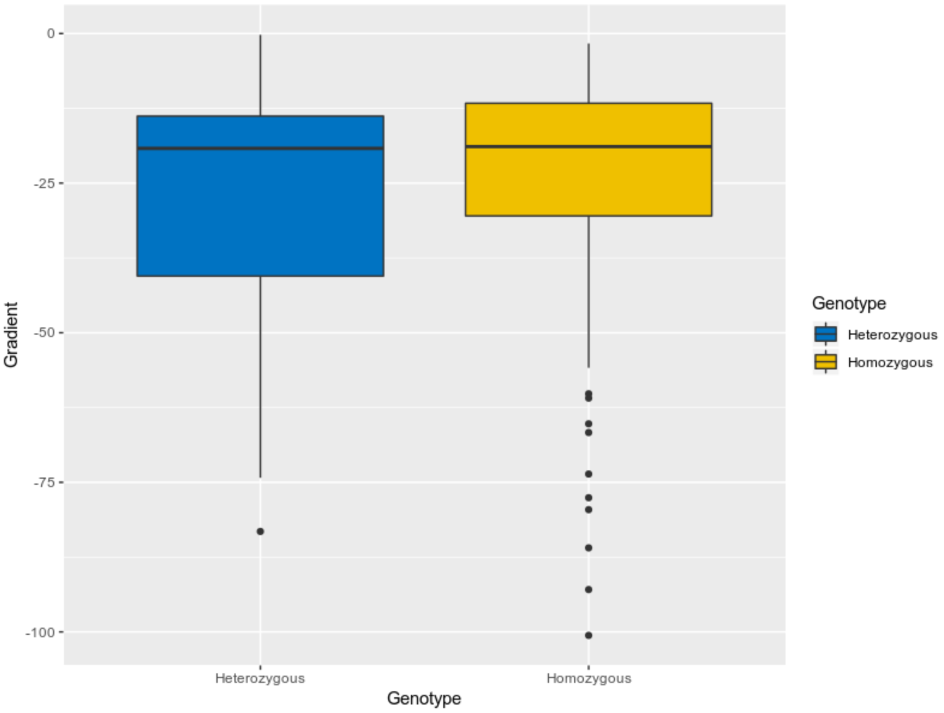


Figure A.7: Box plot of Gradient values based on Genotype

Comparison	p-value
Group 500-Group 0	2.7×10^{-2}
Group 1000-Group 0	3.3×10^{-2}
Group 1000-Group 500	1.0

Table A.10: Games-Howell test for comparisons of AgeDiagnosis values between groups of initial SF level

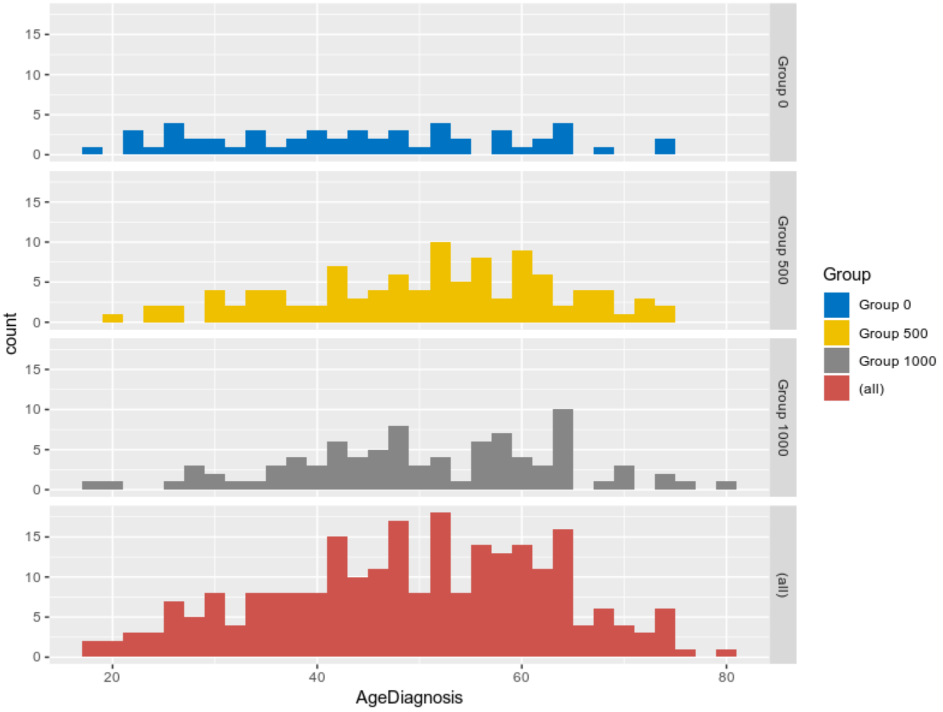


Figure A.8: Histogram of AgeDiagnosis values based on Group

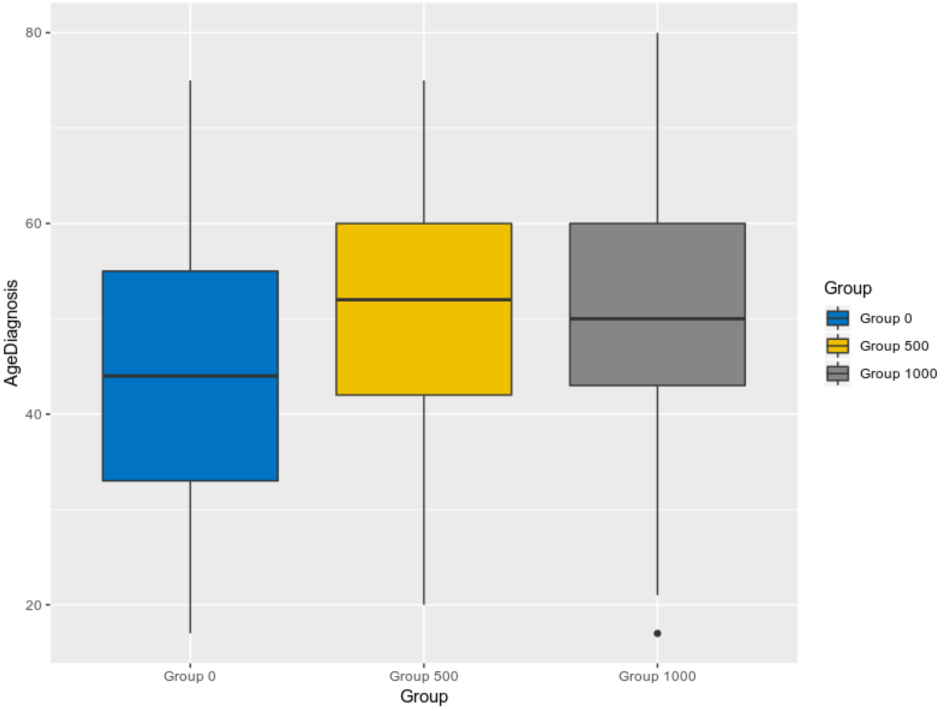


Figure A.9: Box plot of AgeDiagnosis values based on Group

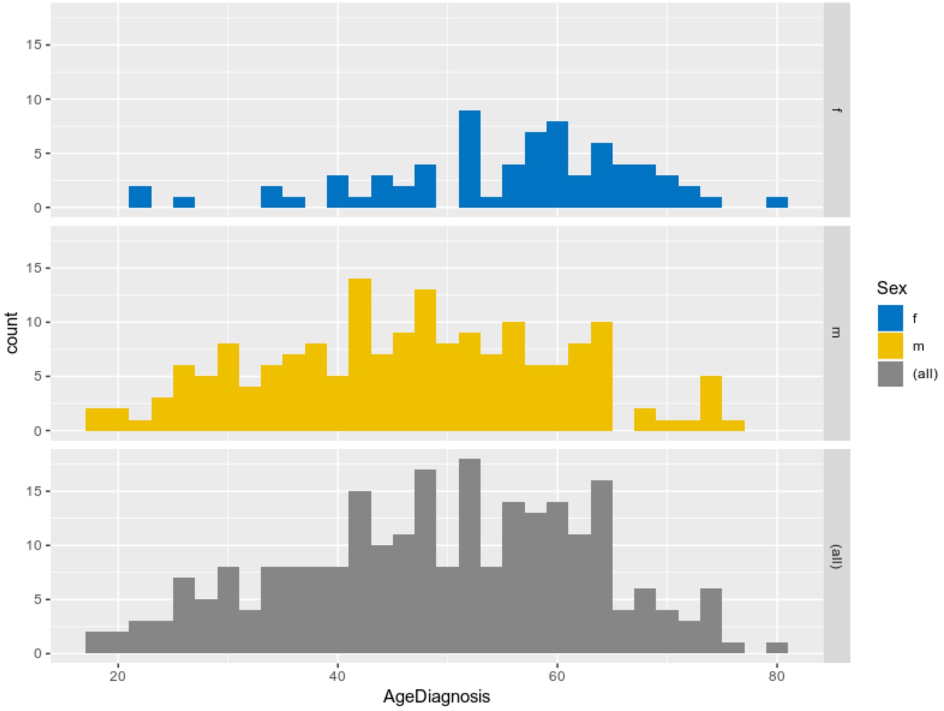


Figure A.10: Histogram of AgeDiagnosis values based on Sex

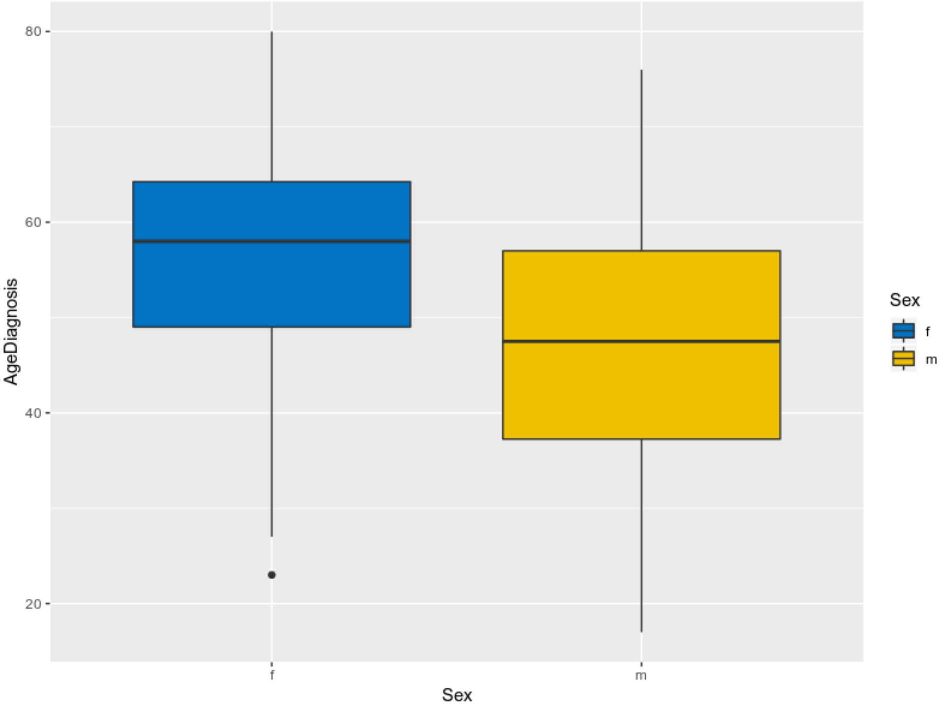


Figure A.11: Box plot of AgeDiagnosis values based on Sex

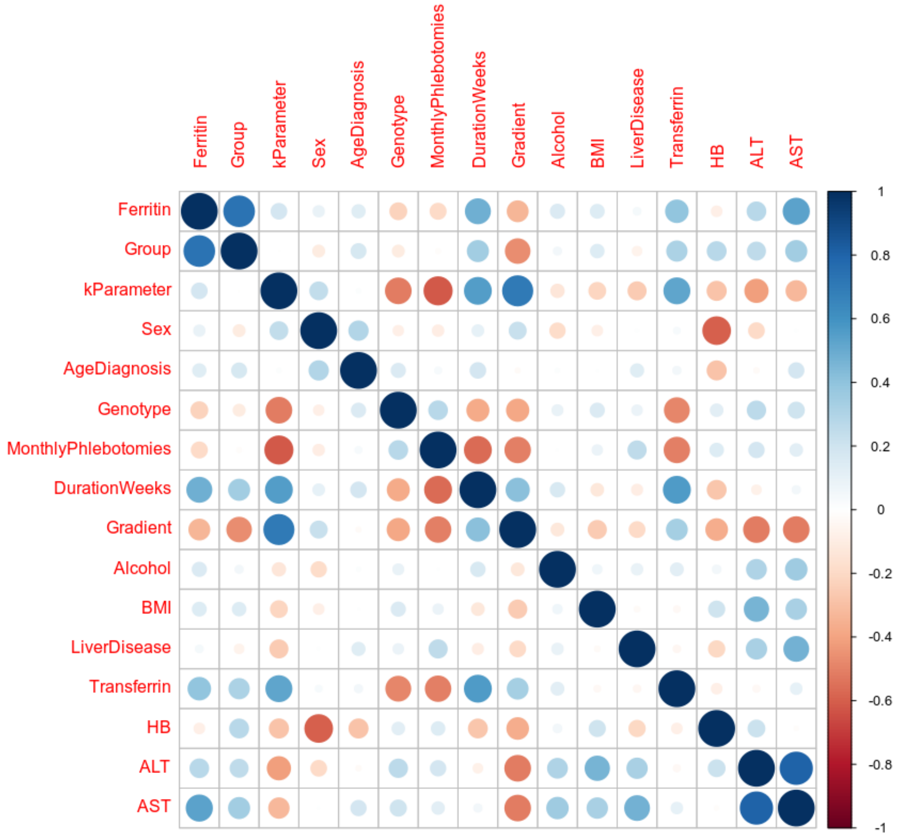


Figure A.12: Correlation plot with coloured scale where red implies negative correlation and blue colors imply positive correlations. Data set with variables disregarded from the main analysis (n = 52)

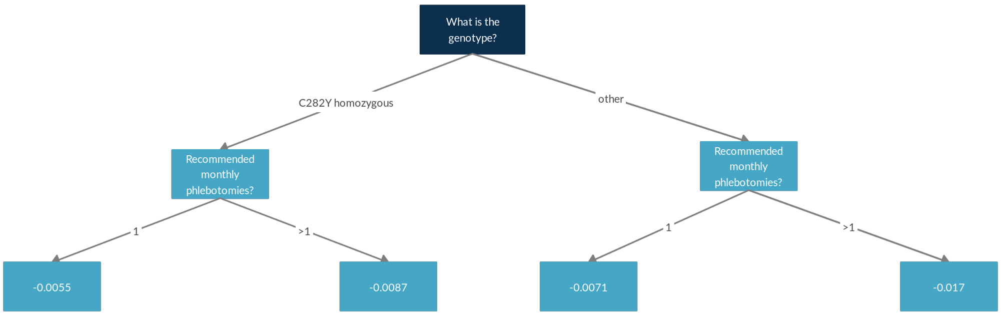


Figure A.13: Decision Tree depicting the grouping of the kParameter values based on Genotype and MonthlyPhlebotomies. Gradient values correspond to the median of each group

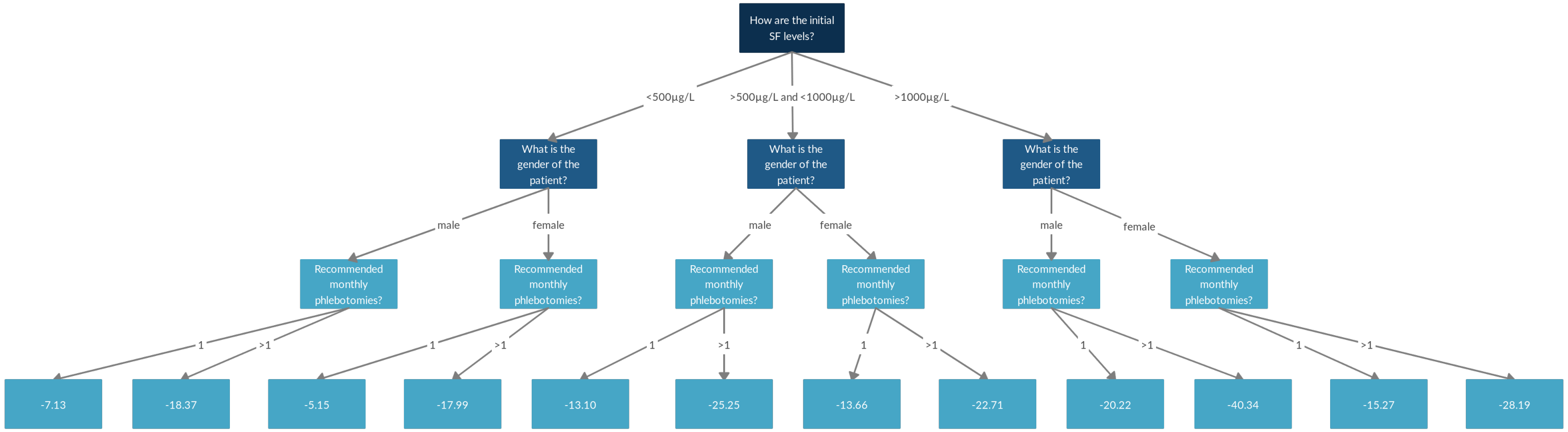

How are the initial SF levels?
<500µg/L
>500µg/L and <1000µg/L
>1000µg/L
What is the gender of the patient?
What is the gender of the patient?
What is the gender of the patient?
male
female
male
female
male
female
Recommended monthly phlebotomies?
Recommended monthly phlebotomies?
Recommended monthly phlebotomies?
Recommended monthly phlebotomies?
Recommended monthly phlebotomies?
Recommended monthly phlebotomies?
1
>1
1
>1
1
>1
1
>1
1
>1
1
>1
-7.13
-18.37
-5.15
-17.99
-13.10
-25.25
-13.66
-22.71
-20.22
-40.34
-15.27
-28.19

Figure A.14: Decision Tree depicting the grouping of the Gradient values based on Group of initial SF level, Sex and MonthlyPhlebotomies. Gradient values correspond to the median of each group

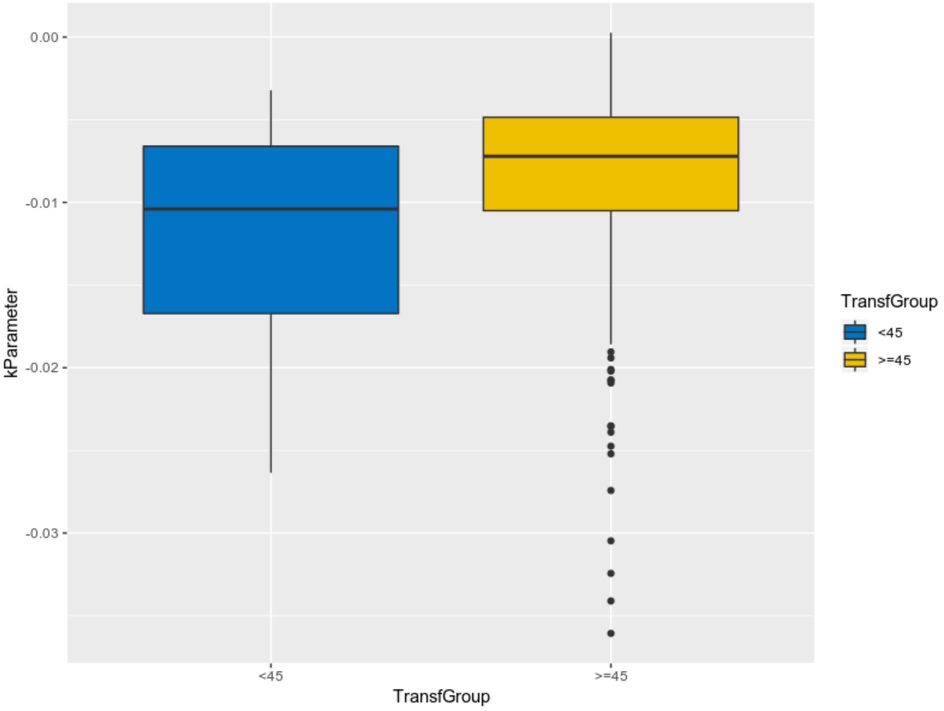


Figure A.15: Box plot of kParameter values based on TransfGroup

Patient characteristics	k parameter value assigned
Homozygous, >=45, 1	-0.0057
Homozygous, >=45, >1	-0.0088
Homozygous, <45, 1	-0.0058
Homozygous, <45, >1	-0.014
Heterozygous, >=45, 1	-0.0062
Heterozygous, >=45, >1	-0.018
Heterozygous, <45, 1	-0.0091
Heterozygous, <45, >1	-0.017

Table A.11: kParameter values assigned to a newly diagnosed patient based on Genotype, TransfGroup and MonthlyPhlebotomies. The attributed value corresponds to the median of each group

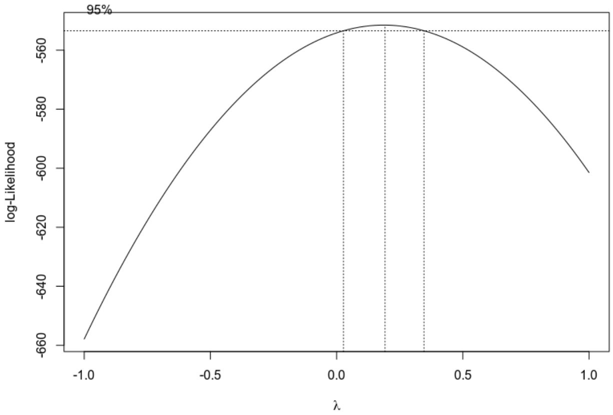


Figure A.16: Plot depicting the interval of best Lambdas found with 95% of confidence using the Box-Cox method

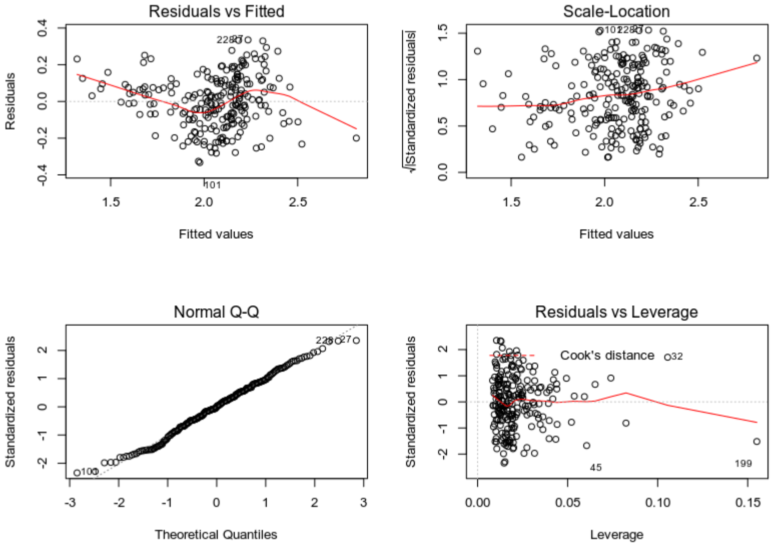


Figure A.17: Diagnostic plots of final linear regression model considered including the kParameter variable with the data without influential data points

Table A.12: Summary of second variable addition, on a linear regression model with Ferritin as explanatory variable

	Dependent variable:						
	(DurationWeeks)ˆ0.2						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ferritin	0.0002*** (0.00002)	0.0002*** (0.00001)	0.0003*** (0.00001)	0.0002*** (0.00002)	0.0002*** (0.00002)	0.0002*** (0.00002)	0.0002*** (0.00002)
kParameter		27.475*** (1.857)					
Gradient			0.012*** (0.001)				
GenotypeHomozygous				0.135*** (0.039)			
AgeDiagnosis					0.003** (0.001)		
Sexm						−0.027 (0.035)	
MonthlyPhlebotomies1							0.238*** (0.028)
Constant	1.882*** (0.026)	2.148*** (0.026)	2.042*** (0.018)	1.790*** (0.037)	1.742*** (0.060)	1.898*** (0.034)	1.739*** (0.028)
Observations	246	246	246	246	246	246	246
R^2	0.231	0.596	0.701	0.268	0.252	0.233	0.411
Adjusted R^2	0.228	0.592	0.698	0.262	0.246	0.227	0.406
Residual Std. Error	0.246 (df = 244)	0.179 (df = 243)	0.154 (df = 243)	0.241 (df = 243)	0.243 (df = 243)	0.246 (df = 243)	0.216 (df = 243)
F Statistic	73.493*** (df = 1; 244)	179.082*** (df = 2; 243)	284.708*** (df = 2; 243)	44.533*** (df = 2; 243)	40.871*** (df = 2; 243)	36.975*** (df = 2; 243)	84.613*** (df = 2; 243)
Note:							*p<0.1; **p<0.05; ***p<0.01

Table A.13: Summary of third variable addition, on a linear regression model with Ferritin + kParameter as explanatory variables

	Dependent variable:				
	(DurationWeeks)ˆ0.2				
	(1)	(2)	(3)	(4)	(5)
Ferritin	0.0002***	0.0002***	0.0001***	0.0002***	0.0002***
	(0.00002)	(0.00002)	(0.00001)	(0.00001)	(0.00002)
kParameter	27.227***	27.453***	27.321***	23.582***	20.240***
	(1.942)	(1.858)	(1.824)	(2.007)	(2.374)
GenotypeHomozygous	0.013				
	(0.030)				
Sexm		−0.021			
		(0.025)			
AgeDiagnosis			0.003***		
			(0.001)		
MonthlyPhlebotomies1				0.107***	0.203***
				(0.025)	(0.045)
kParameter:MonthlyPhlebotomies1					11.903**
					(4.641)
Constant	2.137***	2.162***	2.023***	2.046***	1.995***
	(0.037)	(0.030)	(0.047)	(0.035)	(0.040)
Observations	246	246	246	246	246
R^2	0.596	0.597	0.612	0.625	0.635
Adjusted R^2	0.591	0.592	0.607	0.620	0.629
Residual Std. Error	0.179 (df = 242)	0.179 (df = 242)	0.176 (df = 242)	0.173 (df = 242)	0.171 (df = 241)
F Statistic	119.058*** (df = 3; 242)	119.477*** (df = 3; 242)	127.070*** (df = 3; 242)	134.241*** (df = 3; 242)	104.646*** (df = 4; 241)

Note: *p<0.1; **p<0.05; ***p<0.01

Predictors	VIF
Ferritin	1.12
kParameter	1.81
MonthlyPhlebotomies	4.23
kParameter:MonthlyPhlebotomies	3.35

Table A.14: VIFs of model with Ferritin and an interaction between kParameter and MonthlyPhlebotomies

Predictors	VIF
Ferritin	1.03
kParameter	1.26
MonthlyPhlebotomies	1.27

Table A.15: VIFs of model with Ferritin, kParameter and MonthlyPhlebotomies

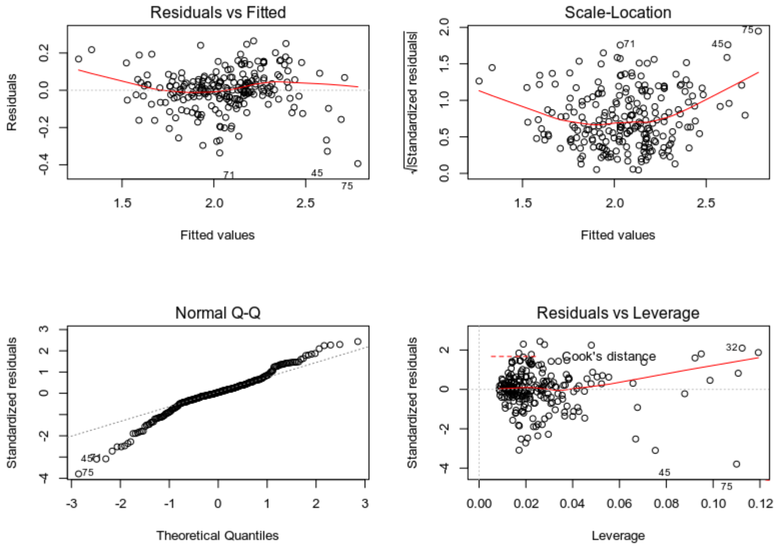


Figure A.18: Diagnostic plots of final linear regression model considered including the Gradient variable with the data without influential data points

Table A.16: Summary of fourth variable addition, on a linear regression model with Ferritin + kParameter + MonthlyPhlebotomies as explanatory variables

	Dependent variable:		
	(DurationWeeks)ˆ0.2		
	(1)	(2)	(3)
Ferritin	0.0002^{***}	0.0002^{***}	0.0002^{***}
	(0.00001)	(0.00001)	(0.00001)
kParameter	23.039^{***}	23.612^{***}	23.340^{***}
	(2.104)	(2.012)	(1.966)
MonthlyPhlebotomies1	0.109^{***}	0.106^{***}	0.109^{***}
	(0.025)	(0.025)	(0.024)
GenotypeHomozygous	0.025		
	(0.029)		
Sexm		−0.010	
		(0.025)	
AgeDiagnosis			0.003^{***}
			(0.001)
Constant	2.023^{***}	2.054^{***}	1.914^{***}
	(0.044)	(0.039)	(0.052)
Observations	246	246	246
R^2	0.626	0.625	0.642
Adjusted R^2	0.620	0.619	0.636
Residual Std. Error (df = 241)	0.173	0.173	0.169
F Statistic (df = 4; 241)	100.763^{***}	100.375^{***}	107.934^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01		

Model	df	AIC
1	3	12.069
2	4	-143.990
3	5	-160.214
4	6	-169.695

Table A.17: AIC values where: model 1) Ferritin, model 2) Ferritin + kParameter, model 3) Ferritin + kParameter + MonthlyPhlebotomies and model 4) Ferritin + kParameter + MonthlyPhlebotomies + AgeDiagnosis

Table A.18: Summary of third variable addition, on a linear regression model with Ferritin + Gradient as explanatory variables

	Dependent variable:				
	(DurationWeeks)ˆ0.2				
	(1)	(2)	(3)	(4)	(5)
Ferritin	0.0003***	0.0003***	0.0003***	0.0003***	0.0003***
	(0.00002)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Gradient	0.012***	0.012***	0.012***	0.011***	0.010***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GenotypeHomozygous	0.039				
	(0.025)				
Sexm		0.042*			
		(0.022)			
AgeDiagnosis			0.002**		
			(0.001)		
MonthlyPhlebotomies1				0.075***	0.191***
				(0.022)	(0.035)
Gradient:MonthlyPhlebotomies1					0.005***
					(0.001)
Constant	2.013***	2.018***	1.954***	1.982***	1.923***
	(0.026)	(0.022)	(0.039)	(0.025)	(0.028)
Observations	246	246	246	246	246
R^2	0.704	0.705	0.709	0.715	0.734
Adjusted R^2	0.700	0.702	0.705	0.711	0.730
Residual Std. Error	0.153 (df = 242)	0.153 (df = 242)	0.152 (df = 242)	0.150 (df = 242)	0.145 (df = 241)
F Statistic	191.651*** (df = 3; 242)	193.096*** (df = 3; 242)	196.130*** (df = 3; 242)	202.206*** (df = 3; 242)	166.678*** (df = 4; 241)
Note:					*p<0.1; **p<0.05; ***p<0.01

Predictors	VIF
Ferritin	1.37
Gradient	2.07
MonthlyPhlebotomies	3.47
Gradient:MonthlyPhlebotomies	2.86

Table A.19: VIFs of model with Ferritin and an interaction between Gradient and MonthlyPhlebotomies

Table A.20: Summary of fourth variable addition, on a linear regression model with Ferritin + Gradient * MonthlyPhlebotomies as explanatory variables

	Dependent variable:		
	(DurationWeeks)ˆ0.2		
	(1)	(2)	(3)
Ferritin	0.0003^{***}	0.0003^{***}	0.0003^{***}
	(0.00001)	(0.00001)	(0.00001)
Gradient	0.009^{***}	0.010^{***}	0.009^{***}
	(0.001)	(0.001)	(0.001)
MonthlyPhlebotomies1	0.195^{***}	0.192^{***}	0.191^{***}
	(0.035)	(0.034)	(0.034)
GenotypeHomozygous	0.047^{*}		
	(0.024)		
Sexm		0.044^{**}	
		(0.021)	
AgeDiagnosis			0.002^{***}
			(0.001)
Gradient:MonthlyPhlebotomies1	0.005^{***}	0.005^{***}	0.005^{***}
	(0.001)	(0.001)	(0.001)
Constant	1.885^{***}	1.897^{***}	1.834^{***}
	(0.034)	(0.030)	(0.043)
Observations	246	246	246
R^2	0.739	0.739	0.742
Adjusted R^2	0.733	0.734	0.737
Residual Std. Error (df = 240)	0.145	0.144	0.144
F Statistic (df = 5; 240)	135.693^{***}	136.150^{***}	138.097^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01		

Model	df	AIC
1	3	12.069
2	4	-218.072
3	6	-243.390
4	7	-248.509

Table A.21: AIC values where: model 1) Ferritin, model 2) Ferritin + Gradient, model 3) Ferritin + Gradient * MonthlyPhlebotomies and model 4) Ferritin + Gradient * MonthlyPhlebotomies + AgeDiagnosis

Number	Ferritin	Group	kParameter	Sex	AgeDiagnosis	Genotype	MonthlyPhlebotomies	DurationWeeks	Gradient
16	1809	Group 1000	-0.0045	m	76	Homozygous	>1	34	-50.48
56	742	Group 500	-0.0247	m	36	Homozygous	>1	20	-31.43
57	3634	Group 1000	-0.0082	m	65	Homozygous	>1	53	-66.68
62	254	Group 0	-0.0019	m	24	Homozygous	1	80	-1.92
104	1930	Group 1000	-0.0089	m	59	Homozygous	1	23	-79.57
123	4877	Group 1000	-0.0324	f	59	Homozygous	1	88	-54.37
145	146	Group 0	-0.0060	f	52	Homozygous	>1	6	-7.16
193	365	Group 0	0.0003	m	22	Homozygous	1	18	-14.72
256	250	Group 0	-0.0043	f	34	Heterozygous	1	12	-12.50
296	692	Group 500	-0.0201	m	25	Homozygous	>1	27	-21.58
306	4069	Group 1000	-0.0071	m	48	Homozygous	>1	43	-92.92
308	4143	Group 1000	-0.0026	f	53	Homozygous	1	98	-41.26
334	4070	Group 1000	-0.0032	m	44	Homozygous	1	210	-18.89
340	2041	Group 1000	-0.0029	m	38	Homozygous	1	166	-11.67
341	556	Group 500	-0.0361	m	31	Heterozygous	>1	6	-74.23
346	172	Group 0	-0.0021	m	28	Homozygous	1	14	-5.14
363	4504	Group 1000	-0.0046	m	62	Homozygous	>1	72	-60.92
369	948	Group 500	-0.0024	m	32	Homozygous	1	18	-47.49

Table A.22: Influential data points discovered on linear regression model with the kParameter variable

Number	Ferritin	Group	kParameter	Sex	AgeDiagnosis	Genotype	MonthlyPhlebotomies	DurationWeeks	Gradient
39	1186	Group 1000	-0.0190	m	30	Homozygous	>1	14	-77.57
57	3634	Group 1000	-0.0082	m	65	Homozygous	>1	53	-66.68
62	254	Group 0	-0.0019	m	24	Homozygous	1	80	-1.92
104	1930	Group 1000	-0.0089	m	59	Homozygous	1	23	-79.57
113	1192	Group 1000	-0.0235	m	43	Homozygous	>1	11	-100.58
123	4877	Group 1000	-0.0324	f	59	Homozygous	1	88	-54.37
145	146	Group 0	-0.0060	f	52	Homozygous	>1	6	-7.16
152	153	Group 0	-0.0086	f	49	Homozygous	1	12	-4.52
306	4069	Group 1000	-0.0071	m	48	Homozygous	>1	43	-92.92
308	4143	Group 1000	-0.0026	f	53	Homozygous	1	98	-41.26
311	4175	Group 1000	-0.0028	m	75	Homozygous	>1	122	-33.44
334	4070	Group 1000	-0.0032	m	44	Homozygous	1	210	-18.89
346	172	Group 0	-0.0021	m	28	Homozygous	1	14	-5.14
363	4504	Group 1000	-0.0046	m	62	Homozygous	>1	72	-60.92
366	129	Group 0	-0.0037	m	27	Homozygous	1	17	-1.66

Table A.23: Influential data points discovered on linear regression model with the Gradient variable

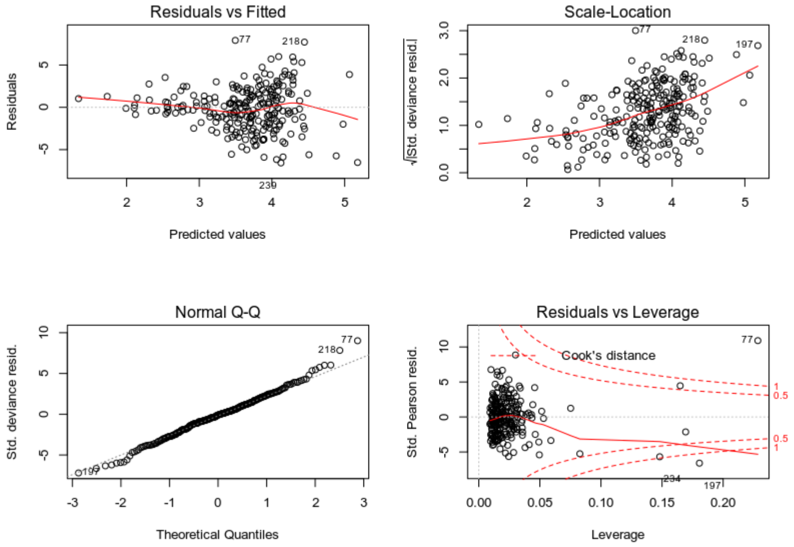


Figure A.19: Diagnostic plots of final poisson regression model considered including the kParameter variable with the data from all the patients

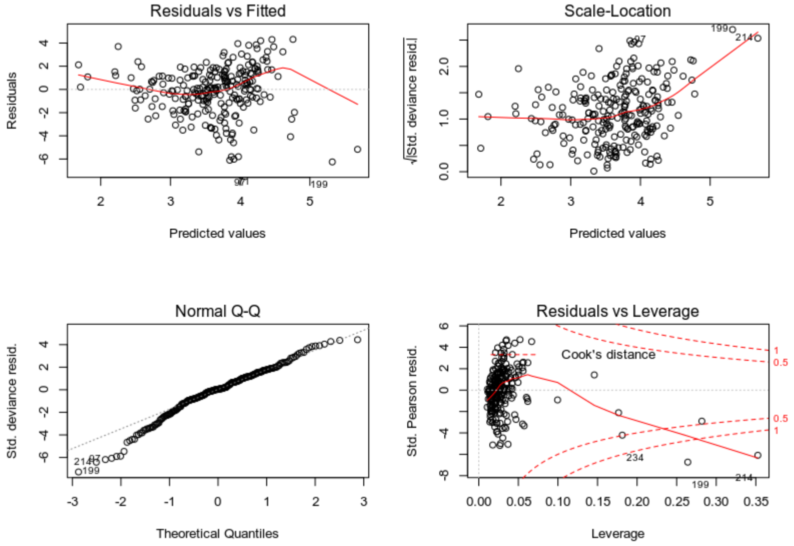


Figure A.20: Diagnostic plots of final poisson regression model considered including the Gradient variable with the data from all the patients

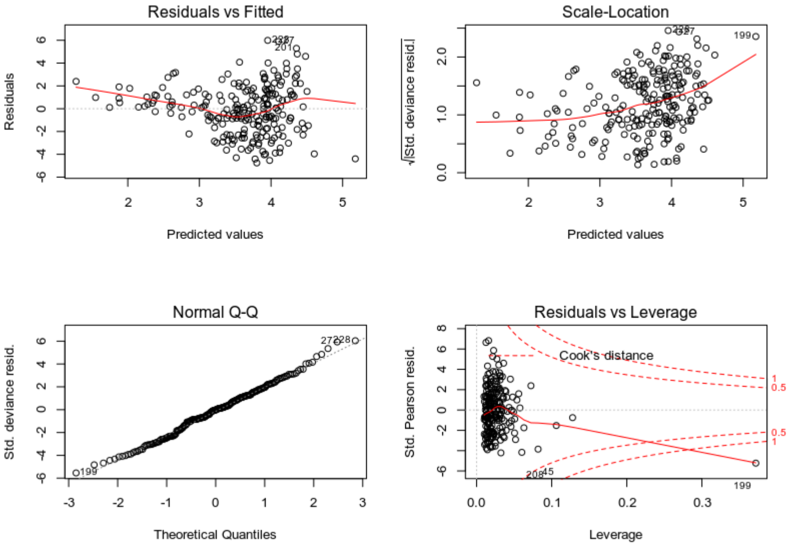


Figure A.21: Diagnostic plots of final poisson regression model considered including the kParameter variable with the data without influential data points

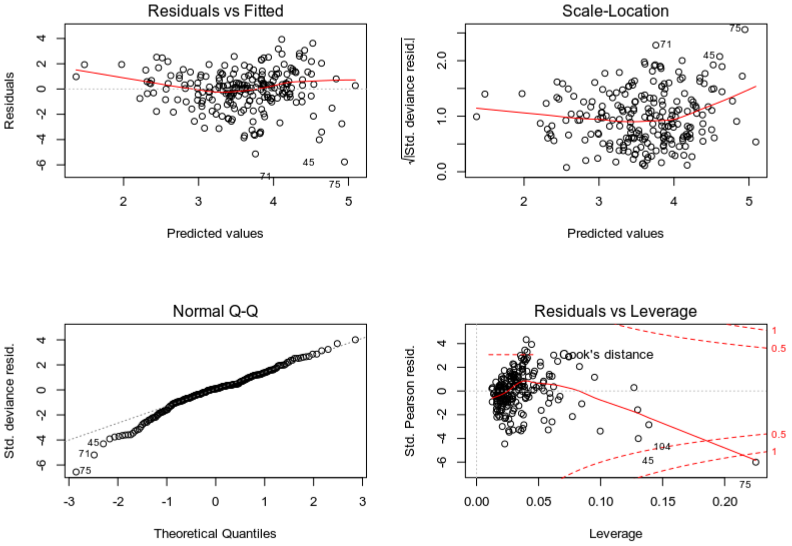


Figure A.22: Diagnostic plots of final poisson regression model considered including the Gradient variable with the data from without influential data points

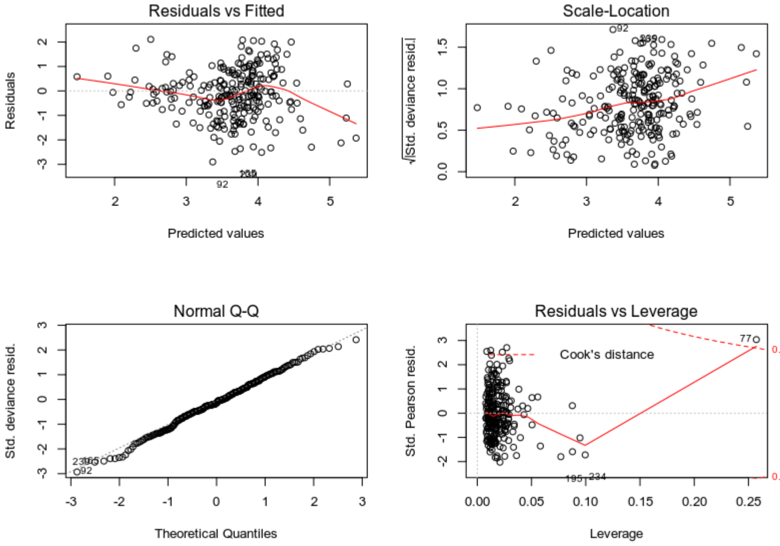


Figure A.23: Diagnostic plots of final NB regression model considered including the kParameter variable with the data from all the patients

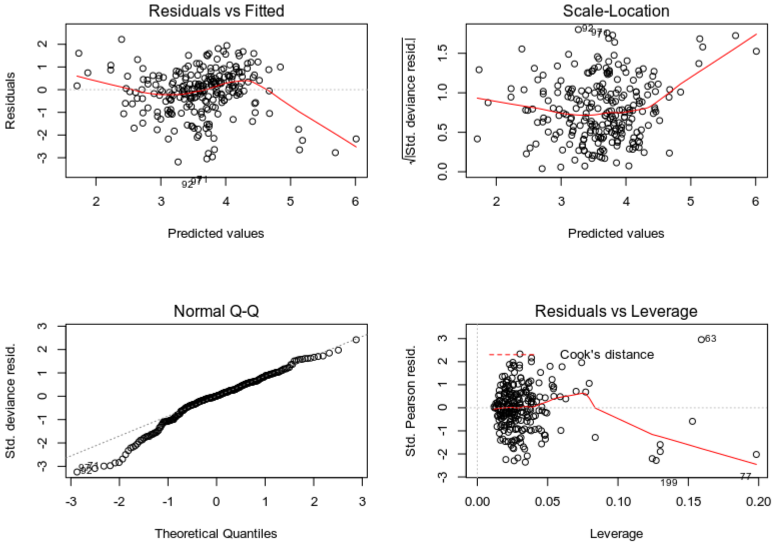


Figure A.24: Diagnostic plots of final NB regression model considered including the Gradient variable with the data from all the patients

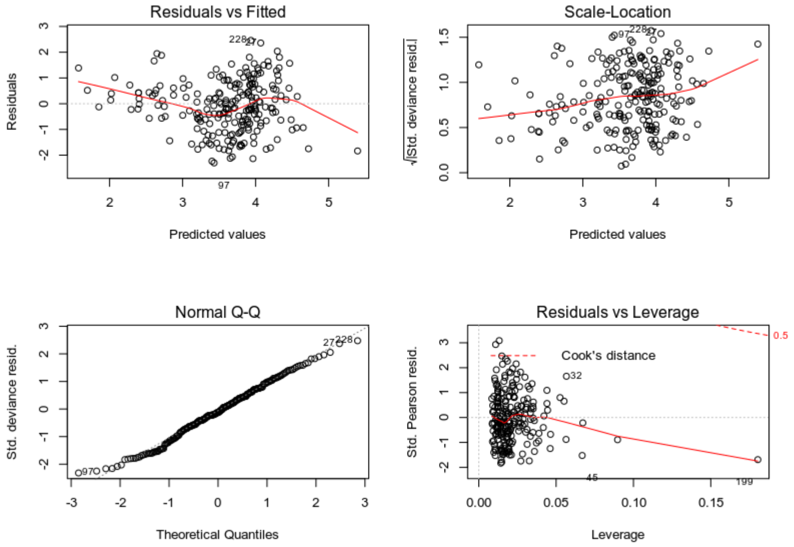


Figure A.25: Diagnostic plots of final NB regression model considered including the kParameter variable with the data without influential data points

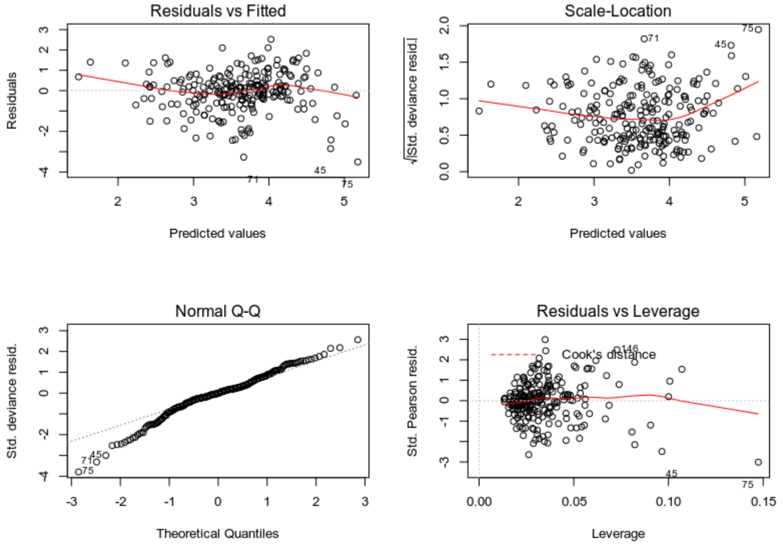


Figure A.26: Diagnostic plots of final NB regression model considered including the Gradient variable with the data from without influential data points

Bibliography

[1] Paul Adams, Albert Altes, Pierre Brissot, Barbara Butzeck, Ioav Cabantchik, Rodolfo Cançado, Sonia Distante, Patricia Evans, Robert Evans, Tomas Ganz, et al. Therapeutic recommendations in hfe hemochromatosis for p. cys282tyr (c282y/c282y) homozygous genotype. *Hepatology international*, 12(2):83–86, 2018.

[2] Paul C Adams. Factors affecting the rate of iron mobilization during venesection therapy for genetic hemochromatosis. *American journal of hematology*, 58(1):16–19, 1998.

[3] Paul C Adams and Shane Agnew. Alcoholism in hereditary hemochromatosis revisited: prevalence and clinical consequences among homozygous siblings. *Hepatology*, 23(4):724–727, 1996.

[4] Paul C Adams and James C Barton. How i treat hemochromatosis. *Blood*, 116(3):317–325, 2010.

[5] Paul C Adams, David M Reboussin, James C Barton, Christine E McLaren, John H Eckfeldt, Gordon D McLaren, Fitzroy W Dawkins, Ronald T Acton, Emily L Harris, Victor R Gordeuk, et al. Hemochromatosis and iron-overload screening in a racially diverse population. *New England Journal of Medicine*, 352(17):1769–1778, 2005.

[6] Katrina J Allen, Lyle C Gurrin, Clare C Constantine, Nicholas J Osborne, Martin B Delatycki, Amanda J Nicoll, Christine E McLaren, Melanie Bahlo, Amy E Nisselle, Chris D Vulpe, et al. Iron-overload–related disease in hfe hereditary hemochromatosis. *New England Journal of Medicine*, 358(3):221–230, 2008.

[7] Nancy C Andrews and Paul J Schmidt. Iron homeostasis. *Annu. Rev. Physiol.*, 69:69–85, 2007.

[8] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[9] Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2014. R package version 1.5.

[10] Bruce R Bacon, Paul C Adams, Kris V Kowdley, Lawrie W Powell, and Anthony S Tavill. Diagnosis and management of hemochromatosis: 2011 practice guideline by the american association for the study of liver diseases. *Hepatology*, 54(1):328–343, 2011.

[11] Stan K. Bardal, Jason E. Waechter, and Douglas S. Martin. Chapter 2 - pharmacokinetics. In Stan K. Bardal, Jason E. Waechter, and Douglas S. Martin, editors, *Applied Pharmacology*, pages 17 – 34. Content Repository Only!, Philadelphia, 2011. ISBN: 978-1-4377-0310-8. doi:https://doi.org/10.1016/B978-1-4377-0310-8.00002-6.

[12] Edouard Bardou-Jacquet, Jeff Morcet, Ghislain Manet, Fabrice Lainé, Michèle Perrin, Anne-Marie Jouanolle, Dominique Guyader, Romain Moirand, Jean-François Viel, and Yves Deugnier. Decreased cardiovascular and extrahepatic cancer-related mortality in treated patients with mild hfe hemochromatosis. *Journal of hepatology*, 62(3):682–689, 2015.

[13] James C Barton and Ronald T Acton. Diabetes in hfe hemochromatosis. *Journal of diabetes research*, 2017, 2017.

[14] James C Barton, Sharon M McDonnell, Paul C Adams, Pierre Brissot, Lawrie W Powell, Corwin Q Edwards, James D Cook, and Kris V Kowdley. Management of hemochromatosis. *Annals of internal medicine*, 129(11_Part_2):932–939, 1998.

[15] James C Barton, Christine E McLaren, Wen-pin Chen, Grant A Ramm, Gregory J Anderson, Lawrie W Powell, V Nathan Subramaniam, Paul C Adams, Pradyumna D Phatak, Lyle C Gurrin, et al. Cirrhosis in hemochromatosis: Independent risk factors in 368 hfe p. c282y homozygotes. *Annals of hepatology*, 17(5):871–879, 2018.

[16] Ernest Beutler and Jill Waalen. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? *Blood*, 107(5):1747–1750, 2006.

[17] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[18] Pierre Brissot. Optimizing the diagnosis and the treatment of iron overload diseases. *Expert review of gastroenterology & hepatology*, 10(3):359–370, 2016.

[19] Pierre Brissot, Thibault Cavey, Martine Ropert, Pascal Guggenbuhl, and Olivier Loréal. Clinical management of hemochromatosis: current perspectives. *International Journal of Clinical Transfusion Medicine*, 5:1, 2017.

[20] Pierre Brissot, Antonello Pietrangelo, Paul C Adams, Barbara de Graaff, Christine E McLaren, and Olivier Loréal. Haemochromatosis. *Nature Reviews Disease Primers*, 4: 18016, 2018.

[21] Guillem Casanovas, Katarzyna Mleczko-Sanecka, Sandro Altamura, Matthias W Hentze, and Martina U Muckenthaler. Bone morphogenetic protein (bmp)-responsive elements located in the proximal and distal hepcidin promoter are critical for its response to hjv/bmp/smad. *Journal of molecular medicine*, 87(5):471–480, 2009.

[22] Yvonne Chan and Roy P Walmsley. Learning and understanding the kruskal-wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Physical therapy*, 77(12):1755–1761, 1997.

[23] Katherine M Conigrave, Peter Davies, Paul Haber, and John B Whitfield. Traditional markers of excessive alcohol use. *Addiction*, 98:31–43, 2003.

[24] R Dennis Cook and Sanford Weisberg. Diagnostics for heteroscedasticity in regression. *Biometrika*, 70(1):1–10, 1983.

[25] Osman Dag, Anil Dolgun, N. Meric Konar, Sam Weerahandi, and Malwane Ananda. *onewaytests: One-Way Tests in Independent Groups Designs*, 2019. R package version 2.4.

[26] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. R package version 1.8-4.

[27] Marie Davidian and David M. Giltinan. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(4):387, Dec 2003.

[28] Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016.

[29] M Elmberg, R Hultcrantz, JF Simard, P Stål, K Pehrsson, and J Askling. Risk of ischaemic heart disease and cardiomyopathy in patients with haemochromatosis and in their first-degree relatives: a nationwide, population-based study. *Journal of internal medicine*, 272 (1):45–54, 2012.

[30] John N Feder, Zenta Tsuchihashi, Alivelu Irrinki, Vincent K Lee, Felipa A Mapa, Ebenezer Morikang, Cynthia E Prass, Steven M Starnes, Roger K Wolff, Seppo Parkkila, et al. The hemochromatosis founder mutation in hla-h disrupts β2-microglobulin interaction and cell surface expression. *Journal of Biological Chemistry*, 272(22):14025–14028, 1997.

[31] Betty J Feir-Walsh and Larry E Toothaker. An empirical comparison of the anova f-test, normal scores test and kruskal-wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34(4):789–799, 1974.

[32] Barbara Finlay and A Agresti. *Statistical methods for the social sciences*. Dellen, 1986.

[33] Sam Firke. *janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2019. R package version 1.2.0.

[34] Robert E Fleming and Prem Ponka. Iron overload in human disease. *New England Journal of Medicine*, 366(4):348–359, 2012.

[35] Linda M Fletcher, Jeannette L Dixon, David M Purdie, Lawrie W Powell, and Darrell HG Crawford. Excess alcohol greatly increases the prevalence of cirrhosis in hereditary hemochromatosis. *Gastroenterology*, 122(2):281–289, 2002.

[36] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019.

[37] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2019. R package version 6.0-84.

[38] Junwei Gao, Juxing Chen, Maxwell Kramer, Hidekazu Tsukamoto, An-Sheng Zhang, and Caroline A Enns. Interaction of the hereditary hemochromatosis protein hfe with transferrin receptor 2 is required for transferrin-induced hepcidin expression. *Cell metabolism*, 9(3): 217–227, 2009.

[39] William Gardner, Edward P Mulvey, and Esther C Shaw. Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological bulletin*, 118(3):392, 1995.

[40] Edoardo Giannini, Federica Botta, Alberto Fasoli, Paola Ceppa, Domenico Risso, Pasquale Bruno Lantieri, Guido Celle, and Roberto Testa. Progressive liver functional impairment is associated with an increase in ast/alt ratio. *Digestive diseases and sciences*, 44(6):1249–1253, 1999.

[41] Gene V Glass. Testing homogeneity of variances. *American Educational Research Journal*, 3(3):187–190, 1966.

[42] David Gohel. *flextable: Functions for Tabular Reporting*, 2019. R package version 0.5.5.

[43] Deepak V Gopal and Hugo R Rosen. Abnormal findings on liver function tests: interpreting results to narrow the diagnosis and establish a prognosis. *Postgraduate medicine*, 107(2): 100–114, 2000.

[44] Stephen A Harrison and Bruce R Bacon. Hereditary hemochromatosis: update for 2003. *Journal of hepatology*, 38:14–23, 2003.

[45] Aravind Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2018. R package version 0.5.2.

[46] Paul Hendricks. *scorer: Quickly Score Models in Data Science and Machine Learning*, 2016. R package version 0.2.0.

[47] Bret L Hicken, Diane C Tucker, and James C Barton. Patient compliance with phlebotomy therapy for iron overload associated with hemochromatosis. *The American Journal of Gastroenterology*, 98(9):2072 – 2077, 2003. ISSN: 0002-9270. doi:https://doi.org/10.1016/S0002-9270(02)06014-8.

[48] Joseph M Hilbe. *Negative binomial regression.* Cambridge University Press, 2011.

[49] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. R package version 5.2.2.

[50] James Jaccard, Michael A Becker, and Gregory Wood. Pairwise multiple comparison procedures: A review. *Psychological Bulletin*, 96(3):589, 1984.

[51] Michael L. Johnson and Susan G. Frasier. [16] nonlinear least-squares analysis. In *Enzyme Structure Part J*, volume 117 of *Methods in Enzymology*, pages 301 – 342. Academic Press, 1985. doi:https://doi.org/10.1016/S0076-6879(85)17018-7.

[52] Yoonsuh Jung and Jianhua Hu. Ak-fold averaging cross-validation procedure. *Journal of nonparametric statistics*, 27(2):167–179, 2015.

[53] Abidullah Khan, Wazir Muhammad Khan, Maimoona Ayub, Mohammad Humayun, and Mohammad Haroon. Ferritin is a marker of inflammation rather than iron deficiency in overweight and obese people. *Journal of obesity*, 2016, 2016.

[54] André I Khuri, Bhramar Mukherjee, Bikas K Sinha, and Malay Ghosh. Design issues for generalized linear models: A review. *Statistical Science*, pages 376–399, 2006.

[55] Choongrak Kim and Barry E Storer. Reference values for cook's distance. *Communications in Statistics-Simulation and Computation*, 25(3):691–708, 1996.

[56] Sungil Kim and Heeyoung Kim. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679, 2016.

[57] Caitriona King and David E Barton. Best practice guidelines for the molecular genetic diagnosis of type 1 (hfe-related) hereditary haemochromatosis. *BMC medical genetics*, 7 (1):81, 2006.

[58] Yutaka Kohgo, Katsuya Ikuta, Takaaki Ohtake, Yoshihiro Torimoto, and Junji Kato. Body iron metabolism and pathophysiology of iron overload. *International journal of hematology*, 88(1):7–15, 2008.

[59] Alexander Kowarik and Matthias Templ. Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16, 2016. doi:10.18637/jss.v074.i07.

[60] Kris V Kowdley, Kyle E Brown, Joseph Ahn, and Vinay Sundaram. Acg clinical guideline: Hereditary hemochromatosis. *American Journal of Gastroenterology*, 114(8):1202–1218, 2019.

[61] Adriana Lazarescu, Beverly M Snively, and Paul C Adams. Phenotype variation in c282y homozygotes for the hemochromatosis gene. *Clinical Gastroenterology and Hepatology*, 3 (10):1043–1046, 2005.

[62] James K Lindsey and Bradley Jones. Choosing among generalized linear models applied to medical data. *Statistics in medicine*, 17(1):59–68, 1998.

[63] Hangcheng Liu. Comparing welch's anova, a kruskal-wallis test and traditional anova in case of heterogeneity of variance. 2015.

[64] European Association For The Study Of The Liver et al. Easl clinical practice guidelines for hfe hemochromatosis. *Journal of hepatology*, 53(1):3–22, 2010.

[65] Jacob A. Long. *interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions*, 2019. R package version 1.1.0.

[66] Edward R Mansfield and Billy P Helms. Detecting multicollinearity. *The American Statistician*, 36(3a):158–160, 1982.

[67] Marius Marusteri and Vladimir Bacarea. Comparing groups for statistical differences: how to choose the right statistical test? *Biochemia medica: Biochemia medica*, 20(1):15–32, 2010.

[68] Jeremy Miles. R squared, adjusted r squared. *Wiley StatsRef: Statistics Reference Online*, 2014.

[69] Jason Millman and Gene V Glass. Rules of thumb for writing the anova table. *Journal of Educational Measurement*, 4(2):41–51, 1967.

[70] Adriana Maria Neghina and Andrei Anghel. Hemochromatosis genotypes and risk of iron overload—a meta-analysis. *Annals of epidemiology*, 21(1):1–14, 2011.

[71] Albina Nowak, Rebekka S Giger, and Pierre-Alexandre Krayenbuehl. Higher age at diagnosis of hemochromatosis is the strongest predictor of the occurrence of hepatocellular carcinoma in the swiss hemochromatosis cohort: A prospective longitudinal observational study. *Medicine*, 97(42), 2018.

[72] Helena Nyblom, Ulf Berggren, Jan Balldin, and Rolf Olsson. High ast/alt ratio may indicate advanced alcoholic liver disease rather than heavy drinking. *Alcohol and alcoholism*, 39(4): 336–339, 2004.

[73] Derek H. Ogle, Powell Wheeler, and Alexis Dinno. *FSA: Fisheries Stock Analysis*, 2019. R package version 0.8.25.

[74] Jason W Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation*, 15(12):1–9, 2010.

[75] George Papanikolaou and Kostas Pantopoulos. Iron metabolism and toxicity. *Toxicology and applied pharmacology*, 202(2):199–211, 2005.

[76] C Pelusi, DI Gasparini, N Bianchi, and R Pasquali. Endocrine dysfunction in hereditary hemochromatosis. *Journal of endocrinological investigation*, 39(8):837–847, 2016.

[77] Gjalt-Jorn Ygram Peters. *userfriendlyscience: Quantitative analysis made accessible*, 2018. R package version 0.7.2. doi:10.17605/osf.io/txequ.

[78] Antonello Pietrangelo. Hereditary hemochromatosis. *Annu. Rev. Nutr.*, 26:251–270, 2006.

[79] Antonello Pietrangelo. Hereditary hemochromatosis: Pathogenesis, diagnosis, and treatment. *Gastroenterology*, 139(2):393 – 408.e2, 2010. ISSN: 0016-5085. doi:https://doi.org/10.1053/j.gastro.2010.06.013.

[80] Alberto Piperno, Maurizio Sampietro, Antonello Pietrangelo, Cristina Arosio, Loredana Lupica, Giuliana Montosi, Anna Vergani, Mirella Fraquelli, Domenico Girelli, Paolo Pasquero, et al. Heterogeneity of hemochromatosis in italy. *Gastroenterology*, 114(5): 996–1002, 1998.

[81] Michael A Poole and Patrick N O'Farrell. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, pages 145–158, 1971.

[82] Graça Porto, Pierre Brissot, Dorine W Swinkels, Heinz Zoller, Outi Kamarainen, Simon Patton, Isabel Alonso, Michael Morris, and Steve Keeney. Emqn best practice guidelines for the molecular genetic diagnosis of hereditary hemochromatosis (hh). *European Journal of Human Genetics*, 24(4):479, 2016.

[83] Lawrie W Powell, D Keith George, Sharon M McDonnell, and Kris V Kowdley. Diagnosis of hemochromatosis. *Annals of Internal Medicine*, 129(11_Part_2):925–931, 1998.

[84] Lawrie W Powell, Rebecca C Seckington, and Yves Deugnier. Haemochromatosis. *The Lancet*, 388(10045):706–716, 2016.

[85] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[86] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[87] Kun Ren. *rlist: A Toolbox for Non-Tabular Data Manipulation*, 2016. R package version 0.4.6.1.

[88] J Rochette, G Le Gac, K Lassoued, C Ferec, and KJH Robson. Factors influencing disease phenotype and penetrance in hfe haemochromatosis. *Human genetics*, 128(3):233–248, 2010.

[89] E Rombout-Sestrienkova. *Erythrocytapheresis, a treatment modality in hereditary hemochromatosis*. Maastricht University, 2016.

[90] Eva Rombout-Sestrienkova, Marian GJ van Kraaij, and Ger H Koek. How we manage patients with hereditary haemochromatosis. *British journal of haematology*, 175(5):759–770, 2016.

[91] Eva Rombout-Sestrienkova, Bjorn Winkens, Marian van Kraaij, Cees Th BM van Deursen, Mirian CH Janssen, Alexander MJ Rennings, Dorothea Evers, Jean-Louis Kerkhoffs, Ad Masclee, and Ger H Koek. Predicting the number of treatments in

naïve hereditary hemochromatosis patients treated by phlebotomy or erythrocytapheresis. *Erythrocytapheresis, a treatment modality in hereditary hemochromatosis*, page 67, 2016.

[92] Enrico Rossi, Max K Bulsara, John K Olynyk, Digby J Cullen, Lesa Summerville, and Lawrie W Powell. Effect of hemochromatosis genotype and lifestyle factors on iron and red cell indices in a community population. *Clinical chemistry*, 47(2):202–208, 2001.

[93] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016.

[94] Mark H. Russell, Helena Nilsson, and Robert C. Buck. Elimination kinetics of perfluorohexanoic acid in humans and comparison with mouse, rat and monkey. *Chemosphere*, 93(10): 2419 – 2425, 2013. ISSN: 0045-6535. doi:https://doi.org/10.1016/j.chemosphere.2013.08.060.

[95] Jeffrey A. Ryan and Joshua M. Ulrich. *xts: eXtensible Time Series*, 2018. R package version 0.11-2.

[96] RM Sakia. The box-cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(2):169–178, 1992.

[97] Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. *GGally: Extension to 'ggplot2'*, 2018. R package version 1.4.0.

[98] PC Sharpe, R McBride, and GPR Archbold. Biochemical markers of alcohol abuse. *QJM: An International Journal of Medicine*, 89(2):137–144, 1996.

[99] Khulood K Shattnawi, Mahmoud Alomari, Nihaya Al-Sheyab, and Ayman Bani Salameh. The relationship between plasma ferritin levels and body mass index among adolescents. *Scientific reports*, 8(1):15307, 2018.

[100] Mital C Shingala and Arti Rajyaguru. Comparison of post hoc tests for unequal variance. *International Journal of New Technologies in Science and Engineering*, 2(5):22–33, 2015.

[101] Chiang W Siah, Debbie Trinder, and John K Olynyk. Iron overload. *Clinica Chimica Acta*, 358(1-2):24–36, 2005.

[102] DW Swinkels, AT Jorna, and RAP Raymakers. Synopsis of the dutch multidisciplinary guideline for the diagnosis and treatment of hereditary haemochromatosis. 2007.

[103] Anthony S Tavill. Diagnosis and management of hemochromatosis. *Hepatology*, 33(5): 1321–1328, 2001.

[104] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[105] Maja Vujic Spasic. Molecular basis of hfe-hemochromatosis. *Frontiers in pharmacology*, 5: 42, 2014.

[106] Charles D Warne, Sophie G Zaloumis, Nadine A Bertalli, Martin B Delatycki, Amanda J Nicoll, Christine E McLaren, John L Hopper, Graham G Giles, Greg J Anderson, John K Olynyk, et al. Hfe p. c282y homozygosity predisposes to rapid serum ferritin rise after menopause: A genotype-stratified cohort study of hemochromatosis in australian women. *Journal of gastroenterology and hepatology*, 32(4):797–802, 2017.

[107] Taiyun Wei and Viliam Simko. *R package "corrplot": Visualization of a Correlation Matrix*, 2017. (Version 0.84).

[108] Evelyn P Whitlock, Betsy A Garlitz, Emily L Harris, Tracy L Beil, and Paula R Smith. Screening for hereditary hemochromatosis: a systematic review for the us preventive services task force. *Annals of internal medicine*, 145(3):209–223, 2006.

[109] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[110] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.

[111] Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2019. R package version 1.3.1.

[112] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2019. R package version 0.8.0.1.

[113] Nan Xiao. *ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'*, 2018. R package version 2.9.

[114] Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.

[115] Andronicos Yiannikourides and Gladys O Latunde-Dada. A short review of iron metabolism and pathophysiology of iron disorders. *Medicines*, 6(3):85, 2019.

[116] Achim Zeileis and Gabor Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005. doi:10.18637/jss.v014.i06.

[117] An-Sheng Zhang, Junwei Gao, Dwight D Koeberl, and Caroline A Enns. The role of hepatocyte hemojuvelin in the regulation of bone morphogenic protein-6 and hepcidin expression in vivo. *Journal of Biological Chemistry*, 285(22):16416–16423, 2010.