Universidade do Minho
Escola de Ciências

Célia Cristina de Aguiar Carvalho

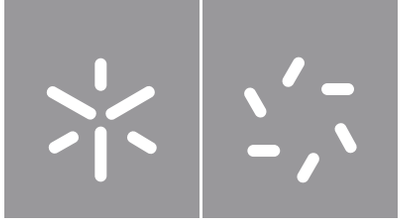
Modelação da Variação do Tempo de Resposta da Carga Viral e da Contagem de Células CD4 à Terapêutica Antirretroviral em Doentes com HIV

fevereiro de 2021

Modelação da Variação do Tempo de Resposta da Carga Viral e da Contagem de Células CD4 à Terapêutica Antirretroviral em Doentes com HIV

Célia Cristina de Aguiar Carvalho

UMinho | 2021



Universidade do Minho
Escola de Ciências

Célia Cristina de Aguiar Carvalho

Modelação da Variação do Tempo de Resposta da Carga Viral e da Contagem de Células CD4 à Terapêutica Antirretroviral em Doentes com HIV

Dissertação de Mestrado
em Estatística

Trabalho efetuado sob a orientação das Professoras
Carla Moreira
Irene Brito

fevereiro de 2021

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Agradecimentos

Desejo exprimir os meus agradecimentos a todos aqueles que, de algum modo, permitiram que esta tese se realizasse.

Agradeço às professoras Doutora Carla Moreira e Doutora Irene Brito pela orientação, disponibilidade, pelo acompanhamento e esclarecimentos ao longo da realização deste projeto.

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Universidade do Minho: 01/02/2021

Assinatura:

Resumo

Os indivíduos infetados com o vírus da imunodeficiência humana (HIV, do inglês *human immunodeficiency virus infection*) estão mais suscetíveis a outras infeções e alguns tipos de cancro. Atualmente, não existe cura para a infeção por HIV e os pacientes devem seguir tratamento ao longo de toda a vida. Dois marcadores clínicos usados para avaliar o estádio da infeção HIV e a eficácia do tratamento são a contagem das células CD4 e a carga viral. Este estudo avalia a progressão da carga viral do HIV e da contagem das células CD4 ao longo do tempo em resposta a diferentes regimes de terapia antirretroviral, numa *coort naïve* de pacientes com HIV e investiga os fatores associados à incapacidade de obter a supressão da carga viral (valores de carga viral iguais a zero) e os fatores associados aos valores de células CD4. Também é determinado o risco de o paciente desenvolver neoplasia.

Neste estudo, foram usados modelos de efeitos aleatórios para avaliar a evolução da carga viral e das células CD4, e depois foram construídos modelos de risco proporcionais de Cox para determinar o risco de neoplasia. Na construção destes últimos modelos foram usados os valores ajustados (dos modelos de efeitos aleatórios) para as células CD4 e para a carga viral.

Os fatores associados com a evolução das células CD4 ao longo do tempo são Sexo, Hemoglobina, Leucócitos, Neutrófilos, Albumina, Tipo de tratamento, Idade na primeira consulta e o Tempo de follow-up. Os fatores associados a valores de carga viral igual a zero são Sexo, Modo de transmissão, Neutrófilos, Albumina, Adesão ao tratamento na primeira consulta, Tipo de tratamento e Tempo de follow-up.

De acordo com estes modelos, o tratamento mais eficaz é INSTIs e o menos eficaz é PIs.

O risco de neoplasia está associado com as células CD4 e Infeções oportunistas (para o modelo com as células CD4) e associado com a carga viral, Infeções oportunistas, Plaquetas e Linfócitos, existindo um efeito de interação entre a carga viral e Infeções oportunistas (para o modelo com a carga viral).

Palavras-chave: modelo de efeitos aleatórios, análise de sobrevivência, HIV

Abstract

Individuals infected with human immunodeficiency virus (HIV) are more vulnerable to many infections and some types of cancer. Currently, there is no cure for HIV infection and patients must follow lifelong treatment. Two clinical markers used to evaluate HIV infection stage and treatment efficacy are CD4 cell count and viral load. This study evaluates HIV viral load progression over time in an antiretroviral naive cohort of pacients that received different treatments and determines the factors associated with the inability to reach viral load supression (viral load values equal to zero) and the factors associated with CD4 counts. Furthermore, the pacient risk of neoplasia is also determined.

This study used random effects models to evaluate CD4 and viral load progression, and then built Cox proportional hazards models to determine neoplasia risk. These last models were built using the adjusted values (obtained with the random effects models) for the CD4 cells and viral load.

The factors associated with CD4 cells progression over time are Sex, Hemoglobin, Leucocytes, Neutrophils, Albumin, Type of treatment, Age at first consultation and Follow-up time.

The factors associated with viral load equal to zero are Sex, Mode of transmission, Neutrophils, Albumin, Beginning treatment at the first consultation, Type of treatment and Follow-up time.

According to these models, the most effective treatment is INSTIs and the least effective is PIs.

The neoplasia risk is associated with CD4 cell count and Opportunistic infections (for the model with CD4 cell count) and is associated with viral load, Opportunistic infections, Platelets and Lymphocytes, existing an interaction effect between viral load and Opportunistic infections (for the model with viral load).

Keywords: random effects model, survival analysis, HIV

Conteúdo

1 Introdução **1**

2 Análise de Dados Longitudinais **5**

2.1 Modelos 5

2.2 Conceitos Básicos 6

2.2.1 Variograma 6

2.2.2 Fontes de Variabilidade dos Dados Longitudinais 7

2.3 Modelo de Efeitos Aleatórios 8

2.3.1 Alguns Modelos Paramétricos Longitudinais 10

2.3.1.1 Correlação em Série Pura 10

2.3.1.2 *Random Intercept* com Correlação em Série e Erro de Medida 10

2.3.1.3 Efeitos Aleatórios e Erro de Medida 11

2.3.2 Estimação de Parâmetros 11

2.3.2.1 Estimador de Máxima Verosimilhança Restrita 12

2.4 Dados Omissos 13

2.4.1 *Intermittent Missing* e *Dropout* 15

2.5 Análise de Diagnóstico 16

2.6 Modelos para Excesso de Zeros 17

2.6.1 Modelos Hurdle 17

2.6.2 Modelo de Zeros Inflacionados 19

2.6.3 Análise de Diagnóstico 20

3 Análise de Sobrevivência **21**

3.1 Funções Básicas em Análise de Sobrevivência 21

3.2 Censura 22

3.3 Modelo de Riscos Proporcionais de Cox 24

3.4 Testes de Hipóteses para os Coeficientes de Regressão 25
3.5 Diagnósticos do Modelo 26

4 Descrição e Análise Exploratória da Base de Dados 27
4.1 Descrição da Base de Dados 27
4.2 Análise Exploratória 35
4.2.1 Análise Exploratória para os Indivíduos que Completaram o Tratamento 39

5 Modelos de Efeitos Aleatórios para os Biomarcadores 43
5.1 Modelo para as Células CD4 43
5.2 Modelos para a Carga Viral 49
5.2.1 Modelo de Poisson com Efeitos Aleatórios 49
5.2.2 Modelo Binomial Negativo com Efeitos Aleatórios 50
5.2.3 Modelos para Excesso de Zeros 50
5.2.4 Modelo Escolhido para a Carga Viral 55

6 Modelos de Riscos Proporcionais de Cox para a Neoplasia 59
6.1 Seleção das Variáveis Explicativas 59
6.2 Diagnósticos do Modelo 61
6.3 Modelo com as Células CD4 62
6.4 Modelo com a Carga Viral 63
6.5 Limitações dos Modelos Apresentados 64

7 Conclusão 65

Bibliografia 69

Lista de Acrónimos

AIC - do inglês, *Akaike information criterion*

EM - do inglês, *Expectation maximization*

HIV- do inglês, *Human immunodeficiency virus infection*

HSH - Homens que fazem sexo com homens

INSTIs - do inglês, *integrase strand transfer inhibitors*

MAR - do inglês, *Missing at random*

MCAR - do inglês, *Missing completely at random*

ML - do inglês, *Maximum likelihood*

MNAR- do inglês, *Missing not at random*

NNRTIs - do inglês, *Non-nucleoside reverse transcriptase inhibitors*

PIs - do inglês, *Protease inhibitors*

REML - do inglês, *Restricted maximum likelihood*

SIDA - Síndrome da imunodeficiência adquirida

TARV - Tratamento antirretrovírico

VIF - do inglês, *Variance inflation factor*

ZINB - do inglês, *Zero-inflated negative binomial*

ZIP - do inglês, *Zero-inflated Poisson*

Lista de Figuras

4.1 Número de pacientes por modo de transmissão: heterossexuais, HSH e toxicodependentes. 29
4.2 Número de pacientes que adeririam aos diferentes tratamentos. 30
4.3 Distribuição das idades dos pacientes na primeira consulta. 31
4.4 Distribuição das idades dos pacientes ao início do tratamento. 32
4.5 Progressão dos níveis das células CD4 por mm^3 ao longo do tempo para cada paciente. 35
4.6 Progressão dos níveis da carga viral (cópias/mL) ao longo do tempo para cada paciente. 36
4.7 Gráficos de caixa de bigodes para as variáveis contínuas que não variam ao longo do tempo. 37
4.8 Variograma empírico sem pontos para as células CD4. 41

5.1 Variogramas empírico e teórico para as células CD4. 47
5.2 Análise de resíduos para o modelo das células CD4. 48
5.3 Avaliação dos pressupostos do modelo para a Carga viral. 58

6.1 Curva de sobrevivência. 61

Lista de Tabelas

4.1 Variáveis registadas e usadas neste estudo. 28

4.2 Número de pacientes que seguiram sempre o mesmo tratamento, morreram, mudaram de tratamento e perderam-se de follow-up (FU) globalmente e por tratamento. 31

4.3 Estatísticas descritivas para as variáveis contínuas registadas e usadas neste estudo. 33

4.3 Estatísticas descritivas para as variáveis contínuas registadas e usadas neste estudo (continuação). 34

4.4 Resumo dos testes de Mann-Whitney para as variáveis contínuas que não variam com o tempo. 38

4.5 Características dos indivíduos que completaram o tratamento e dos que mudaram de tratamento. 39

5.1 AIC para os modelos ajustados. 44

5.2 Estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente do modelo para CD4. 45

5.3 Modelos para a Carga viral. 52

5.3 Modelos para a Carga viral(continuação). 53

5.3 Modelos para a Carga viral(continuação). 54

5.4 Estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente para a componente dos zeros do modelo para a Carga viral. . . . 56

5.5 Estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente para a componente dos valores positivos do modelo para a Carga viral. 56

5.6 Componentes da variância para o modelo hurdle binomial negativo de efeitos aleatórios. 57

6.1 Razão de risco para cada covariável do modelo final e respetivos intervalos de confiança (IC) e valores de prova. 60
6.2 Teste para avaliação do pressuposto da proporcionalidade dos riscos. 61
6.3 Razão de risco para cada covariável do modelo e respetivos intervalos de confiança (IC) e valores de prova. 62
6.4 Teste para avaliação do pressuposto da proporcionalidade dos riscos. 63
6.5 Razão de risco para cada covariável do modelo e respetivos intervalos de confiança (IC) e valores de prova. 63
6.6 Teste para avaliação do pressuposto da proporcionalidade dos riscos. 64

Capítulo 1

Introdução

O vírus da imunodeficiência humana (HIV, do inglês *Human Immunodeficiency Virus*) é uma infeção que ataca o sistema imunitário tornando o indivíduo mais vulnerável a infeções e alguns tipos de cancro (Organização Mundial de Saúde, 2020).

O HIV foi responsável por quase 33 milhões de mortes até ao momento e, no final de 2019, cerca de 38 milhões de pessoas viviam com HIV (Organização Mundial de Saúde, 2020). Em Portugal, as estimativas realizadas para o ano 2017 revelaram que viviam 39820 pessoas com infeção por HIV, $7,8\%$ das quais não estavam diagnosticadas (Direção Geral de Saúde, 2019). De acordo com as notificações recebidas até 30 de junho de 2019, em 2018 foram diagnosticados 973 novos casos de infeção por HIV em Portugal, o que equivale a uma taxa de 9,5 casos/100 mil habitantes, não ajustada para o atraso da notificação (Direção Geral de Saúde, 2019).

Duas medidas frequentemente usadas para determinar o estádio da infeção HIV são a contagem das células CD4 por milímetro cúbico (células/mm^3) no sangue e a quantidade de moléculas HIV RNA por mililitro (cópias/mL) no plasma, também designada por carga viral (Volberding et al., 2010). A partir do momento em que há uma infeção, as células CD4 vão diminuindo em número com o tempo, de tal modo que, o número de células CD4 de uma pessoa infetada pode ser usado para monitorizar a progressão da doença. Assim, a contagem de CD4 indica o grau de depleção imunitária ou imunodeficiência (Volberding et al., 2010). A restante reserva imunológica, refletida na contagem de CD4, é altamente preditiva da capacidade do risco de doenças oportunistas e mortalidade (Volberding et al., 2010). Por outro lado, a carga viral indica a taxa de produção de viriões de HIV-1 e a taxa esperada da subsequente destruição das células CD4 (Volberding et al., 2010). Embora a contagem de células CD4 é o meio principal para determinar o estádio da doença na infeção

por HIV (Volberding et al., 2010), a carga viral é a medida preferencialmente recomendada para determinar o sucesso ou insucesso da terapia antirretrovírica (Volberding et al., 2010; Organização Mundial de Saúde, 2017). Após o início do tratamento antirretrovírico, espera-se que a carga viral diminua rapidamente e que desça para níveis inferiores a 50 cópias/mL depois de 12 a 24 semanas de tratamento (Volberding et al., 2010).

Um pessoa saudável, que não esteja infetada por HIV, tem uma contagem de células CD4 superior a 500 células/mm^3 (Klimas et al., 2008). Por outro lado, se o valor da contagem das células CD4 para um indivíduo for inferior a 200 células/mm^3, significa que a sua imunidade está severamente comprometida e, de acordo com a Organização Mundial de Saúde (2007), estes indivíduos possuem SIDA (Síndrome da imunodeficiência adquirida). A SIDA é o estádio mais avançado da infeção HIV e pode demorar vários anos para se desenvolver se não for tratada. A SIDA é definida como o desenvolvimento de certos cancros, infeções ou outras manifestações clínicas severas (Organização Mundial de Saúde, 2020).

Atualmente não existe tratamento antirretrovírico capaz de curar a infeção por HIV e o tratamento suprime a replicação viral no corpo do indivíduo e permite a recuperação do sistema imunitário para combater as infeções (Organização Mundial de Saúde, 2020). Desde de 2016, a Organização Mundial de Saúde recomenda que todos os indivíduos infetados por HIV recebam tratamento antirretrovírico durante toda a vida, independentemente do seu estádio clínico ou da contagem de células CD4.

Segundo a Organização Mundial de Saúde (2017), um paciente é considerado estável na terapia antirretrovírica se cumprir os seguintes critérios: segue a terapia há pelo menos um ano, não apresenta doenças, boa compreensão da adesão para toda a vida e evidência do sucesso do tratamento (duas medidas consecutivas da carga viral abaixo de 1000 cópias/mL). Assim, visto que os pacientes têm que receber tratamento antirretrovírico ao longo da vida, é importante avaliar a eficácia dos vários tratamentos disponíveis e identificar se existem tratamentos mais eficazes do que outros.

O objetivo desta investigação é avaliar a dinâmica da carga viral do HIV bem como o número de células CD4 ao longo do tempo em resposta a diferentes regimes de terapia antirretroviral (TARV), numa *coort naïve* de pacientes com HIV e investigar os fatores associados à incapacidade de obter a supressão da carga viral (valores de carga viral iguais a zero). Pretende-se também determinar o risco de um paciente de desenvolver neoplasia. Em particular, tendo em conta que o que se pretende é reduzir a carga viral de forma a ser

zero e manter os CD4 em níveis que evitem as infeções oportunistas ou outros problemas, o presente estudo tem como objetivos:

- Avaliar qual o efeito tratamento retrovírico (TARV) na resposta virológica ao longo do tempo e qual destes fármacos é o mais eficaz;
- Identificar quais os fatores associados à evolução das células CD4 e carga viral ao longo do tempo e os seus respetivos efeitos nas células CD4 e carga viral;
- Determinar a razão pela qual certos pacientes ao longo do tempo nunca atingem valores de carga viral iguais a zero, o que seria de esperar ao fim de 6 meses de tratamento, e quais as diferenças entre os pacientes que atingem os valores zero dos pacientes cujos valores de carga viral são superiores a zero;
- Calcular o risco de neoplasia, usando o modelo de riscos proporcionais de Cox;
- Identificar quais os fatores de risco que estão associados à existência de neoplasias e se algum tratamento em particular aumenta o risco de neoplasia.

Para além do capítulo da introdução, o presente trabalho está organizado em mais seis capítulos.

O segundo capítulo apresenta a teoria sobre a análise de dados longitudinais. Após uma breve apresentação do tipo de modelos que podem ser usados no tratamento destes dado, é abordada a teoria sobre os modelos de efeitos aleatórios assumindo que os erros seguem uma distribuição normal e quando existe um excesso de zeros.

O terceiro capítulo aborda alguns conceitos básicos em análise de sobrevivência e apresenta o modelo de riscos proporcionais de Cox.

O quarto capítulo apresenta as características da base de dados e a análise exploratória.

No quinto capítulo são apresentados os modelos com efeitos aleatórios para os biomarcadores (células CD4 e carga viral).

O sexto capítulo apresenta os modelos de risco proporcionais para determinar o risco de neoplasia.

No último capítulo, encontram-se as principais conclusões do trabalho desenvolvido, indicando alguns comentários acerca dos resultados obtidos e das dificuldades/limitações encontradas ao longo deste estudo.

Capítulo 2

Análise de Dados Longitudinais

Os dados longitudinais surgem quando observações repetidas da variável resposta são obtidas ao longo do tempo para cada indivíduo, num ou mais grupos em estudo (Cabral & Gonçalves, 2011). Este tipo de dados estão frequentemente presentes em investigação clínica em que estudos longitudinais desempenham um papel importante em aumentar a compreensão do desenvolvimento e da persistência da doença (Rizopoulos, 2012).

Os estudos longitudinais permitem avaliar as alterações da variável resposta ao longo do tempo, através da medição repetida dos indivíduos durante o estudo (Rizopoulos, 2012). Para além disso, de acordo com Diggle et al. (2002), permitem distinguir as modificações da variável resposta ao longo do tempo dentro do mesmo indivíduo (efeito logitudinal) das diferenças entre indivíduos ao nível dos valores de *baseline* (efeito de coorte).

Num estudo longitudinal, é esperado que as medidas repetidas obtidas num mesmo indivíduo estejam correlacionadas (Rizopoulos, 2012). Esta característica implica que ferramentas estatísticas, tais como o teste t e regressão linear simples que assumem independência entre as observações, sejam desadequadas na análise de dados longitudinais (Rizopoulos, 2012). A ação de ignorar a correlação dos dados longitudinais pode conduzir a inferências incorretas sobre os coeficientes de regressão, estimativas ineficientes dos coeficientes e proteção sub-óptima contra os enviesamentos causados por dados omissos (Diggle et al., 2002).

2.1 Modelos

De acordo com Diggle et al. (2002) existem três tipos de modelos para lidar com dados longitudinais: modelos marginais, modelos de efeitos aleatórios e modelos de transição.

Num modelo marginal, a relação da resposta com as variáveis explicativas é modelada separadamente da correlação das observações para o mesmo indivíduo. No entanto, embora os modelos marginais permitam modelar as consequências da correlação entre as medidas repetidas, estes modelos não explicam a sua origem (Fitzmaurice et al., 2008).

Os modelos de transição (ou de Markov) são modelos dinâmicos que buscam estudar a transição de um estado para o outro. Neste tipo de modelos, os valores passados são tratados como variáveis preditoras adicionais.

Nos modelos de efeitos mistos aleatórios é assumida uma relação linear entre a resposta e as variáveis explicativas, em que os coeficientes de regressão são diferentes para cada indivíduo. Este método foi o escolhido para analisar a base de dados e que será seguidamente explicado com mais detalhe.

2.2 Conceitos Básicos

Seja Y_{ij} a resposta do indivíduo i, $i = 1, ..., m$, no tempo t_j, $j = 1, ..., n_i$, e seja $N = \sum_{i=1}^{m} n_i$ o número total de medidas da base de dados. Se n_i é igual para todos os indivíduos, ou seja, se todas as medidas forem registadas nos mesmos tempos, $n_i = n$ e, neste caso, o estudo é balanceado (Borges, 2015). Caso contrário, o estudo é não balanceado.

2.2.1 Variograma

Para explorar a estrutura de correlação dos dados longitudinais é frequentemente usado o variograma empírico construído a partir dos resíduos obtidos após a construção de um modelo saturado de regresssão simples para a média da resposta (Diggle et al., 2002).

De acordo com Diggle et al. (2002), a funcão de autocorrelação é mais eficaz para estudar dados igualmente espaçados que são aproximadamente estacionários. No entanto, as autocorrelações são mais difíceis de determinar para dados irregularmente espaçados a não ser que seja arredondado os tempos das observações. Alternativamente, pode ser usado o variograma que descreve a associação entre valores repetidos e é facilmente estimado para tempos de observação irregulares (Diggle, 1990).

Para um processo estocástico $Y(t)$, o variograma é definido por

$$Y(u) = \frac{1}{2}E[\{Y(t) - Y(t-u)\}^2], \quad u \geq 0. \tag{2.2.1}$$

Se $Y(t)$ é estacionário, o variograma está diretamente relacionado com a função de autocorrelação, $\rho(u)$, através da equação

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}, \tag{2.2.2}$$

em que σ^2 é a variância de $Y(t)$. Para além disso, o variograma está também bem definido para uma classe restrita de processos não-estacionários para os quais os incrementos, $Y(t) - Y(t-u)$, são estacionários.

Para dados longitudinais, o variograma empírico é calculado a partir da metade do quadrado das diferenças entre os pares de resíduos,

$$v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2, \tag{2.2.3}$$

e das correspondentes medidas entre os tempos

$$u_{ijk} = t_{ij} - t_{ik}. \tag{2.2.4}$$

Se os tempos t_{ij} não forem completamente irregulares, haverá mais do que uma observação para cada valor de u. Considerando que $\hat{\gamma}(u)$ é a média de todos v_{ijk} correspondentes a um valor particular de u, então para tempos de amostragem irregulares, o variograma pode ser estimado a partir dos dados (u_{ijk}, v_{ijk}), $j < k = 1, ..., n_i$; $i = 1, ..., m$ ajustando uma curva não paramétrica. A variância, σ^2, é estimada determinando a média de todas as metades dos quadrados das diferenças $\frac{1}{2}(y_{ij} - y_{lk})^2$ com $i \neq l$. A função de autocorrelação para o lag u pode ser estimada a partir do variograma empírico através de

$$\hat{\rho}(u) = 1 - \hat{\gamma}(u)/\hat{\sigma^2}. \tag{2.2.5}$$

2.2.2 Fontes de Variabilidade dos Dados Longitudinais

Na construção de modelos para dados longitudinais, é necessário perceber quais são as fontes de variação deste tipo de dados. De acordo com Diggle et al. (2002), para dados

longitudinais, existem três fontes de variação aleatória:

- Efeitos aleatórios: quando as unidades são amostradas de forma aleatória de uma população, vários aspetos do seu comportamento poderão exibir variação estocástica entre as unidades. Por exemplo, algumas unidades podem responder intrinsecamente melhor ao tratamento do que outras unidades.

- Correlação em série: pelo menos algumas medidas observadas da unidade poderão ser a resposta aos processos estocásticos que variam com o tempo dentro de cada unidade. Este tipo de variação estocástica resulta na correlação entre pares de medidas para a mesma unidade. Esta correlação depende do espaçamento temporal entre os pares de medidas e, tipicamente, a correlação torna-se mais baixa quando o espaçamento temporal aumenta.

- Erro de medida: quando as medidas individuais envolvem algum tipo de amostragem dentro das unidades, o processo de medida pode adicionar uma componente de variação em relação aos dados.

2.3 Modelo de Efeitos Aleatórios

Existem muitas formas em que as fontes de variabilidade dos dados longitudinais podem ser incorporadas em modelos específicos.

Primeiro, seja Y_{ij} a resposta do indivíduo i, $i = 1, ..., m$, no tempo t_j, $j = 1, ..., n_i$, e sejam x_{ijk} os valores associados a cada Y_{ij}, em que $k = 1, ..., p$ representa o número de variáveis explicativas p. Assim, um possível modelo para a resposta Y_{ij} é

$$Y_{ij} = \beta_1 x_{ij1} + ... + \beta_p x_{ijp} + \epsilon_{ij}, \tag{2.3.1}$$

em que os erros ϵ_{ij} são sequências aleatórias de comprimento n associadas a cada indivíduo m. Em regressão linear, ϵ_{ij} seriam variáveis aleatórias mutuamente independentes com distribuição $\mathcal{N}(0, \sigma^2)$. No entanto, no contexto de análise de dados longitudinais, é esperado que ϵ_{ij} estejam correlacionados para cada indivíduo (Diggle et al., 2002).

A ideia geral dos modelos de efeitos mistos aleatórios é assumir a estrutura para ϵ_{ij} como em (2.3.1), separando o erro de medida da variabilidade entre indivíduos e da variabilidade dentro indivíduos (Sousa, 2011). Para a base de dados completa $N = \sum_{i=1}^{m} n_i$,

seja $\boldsymbol{Y} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_m)$ a variável aleatória para todas as medidas de todos os indivíduos, então no modelo linear geral para dados longitudinais

$$\boldsymbol{Y} \sim \mathcal{MNV}(X\boldsymbol{\beta}, \sigma^2 V(\psi)), \tag{2.3.2}$$

em que X é a matriz $(N \times p)$ das variáveis explicativas. A matriz V, de dimensão $(N \times N)$ e com parâmetro ψ, é uma matriz diagonal de bloco porque é assumida a independência entre indivíduos, em que a matriz diagonal V_i representa a matriz de covariância para o indivíduo i (Sousa, 2011).

De acordo com Diggle et al. (2002), o modelo de efeitos aleatórios geral é definido por

$$Y_{ij} = \mu_{ij} + \boldsymbol{d}_{ij}^{'}\boldsymbol{U}_i + W_i(t_{ij}) + Z_{ij}, \tag{2.3.3}$$

em que $\boldsymbol{U}_i$ são m realizações i.i.d. de $\mathcal{MNV}(0, G)$, que representa os efeitos aleatórios a nível individual, e $\boldsymbol{d}_{ij}^{'}$ é um vetor de covariáveis para os efeitos aleatórios. $W_i(t_{ij})$ é um processo Gaussiano estacionário contínuo no tempo com média 0, variância σ^2 e função de correlação $\rho(u)$, em que u é o intervalo de tempo. Para além disso, $\boldsymbol{U}_i$, $W_i(t_{ij})$ e Z_{ij} correspondem aos efeitos aleatórios, correlação em série e erro de medida, respetivamente. Ainda, Z_{ij} são N realizações i.i.d com $\mathcal{N}(0, \tau^2)$.

A função de correlação entre $W_i(t_{ij})$, $\rho(u)$, pode ser definida de várias formas. Para um modelo de efeitos aleatórios que tem em conta a estrutura de correlação exponencial dentro de indivíduos, $\rho(u)$ é dada por

$$\rho(u) = exp\left(-\frac{1}{\phi}|u|\right), \tag{2.3.4}$$

e para um modelo de efeitos aleatórios que tem em conta a estrutura de correlação gaussiana dentro de indivíduos, $\rho(u)$ é

$$\rho(u) = exp\left(-\frac{1}{\phi}u^2\right), \tag{2.3.5}$$

em que ϕ é a amplitude, que especifica a distância a que dois pontos deixam de estar correlacionados (Borges, 2015).

Seja $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})$ o vetor das variáveis aleatórias, ϵ_{ij}, associado à unidade i. Seja D_i a matriz $n_i \times r$ com a linha j, $\boldsymbol{d}_{ij}$, e H_i a matriz $n_i \times n_i$ com elemento (j,k), $h_{ij} = \rho(|t_{ij} - t_{ik}|)$, isto é, h_{ijk} é a correlação entre $W_i(t_{ij})$ e $W_i(t_{ik})$. Finalmente, seja I_i a matriz identidade $n_i \times n_i$, então a matriz de covariância para ϵ_i é

$$Var(\boldsymbol{\epsilon}_i) = D_i G D_i' + \sigma^2 H_i + \tau^2 I_i. \tag{2.3.6}$$

Ainda ao longo desta secção serão apresentados casos particulares de (2.3.6). Visto que para estes modelos as medidas para diferentes indivíduos são independentes, (2.3.6) pode ser escrita como

$$Var(\boldsymbol{\epsilon}) = D_i G D' + \sigma^2 H + \tau^2 I. \tag{2.3.7}$$

em que $\boldsymbol{\epsilon} = (\epsilon_i, ..., \epsilon_n)$ denota uma sequência genérica de n medidas de um indivíduo.

2.3.1 Alguns Modelos Paramétricos Longitudinais

2.3.1.1 Correlação em Série Pura

Assumindo que não estão presentes efeitos aleatórios ou erros de medida, então

$$\epsilon_j = W(t_j), \tag{2.3.8}$$

e a equação (2.3.7) corresponde a

$$Var(\boldsymbol{\epsilon}) = \sigma^2 H. \tag{2.3.9}$$

2.3.1.2 *Random Intercept* com Correlação em Série e Erro de Medida

O modelo (2.3.3) em que todas as componentes de variação estão presentes considera U como sendo uma variável aleatória Gaussiana univariada, com média zero, variância ν^2, e $d_j = 1$. Assim, o valor da realização de U representa o *random intercept*, ou seja, a quantidade para qual todas as medidas de uma unidade aumentam ou diminuem em relação à média da população. Neste caso, a matriz de covariância (2.3.7) é

$$Var(\boldsymbol{\epsilon}) = \nu^2 J + \sigma^2 H + \tau^2 I, \tag{2.3.10}$$

em que J é a matriz $n \times n$ com todos os seus elementos iguais a 1. O variograma para este tipo de modelo tem a forma

$$\gamma(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}, \tag{2.3.11}$$

exceto que agora a variância para cada ϵ_j é $Var(\epsilon_j) = \nu^2 + \sigma^2 + \tau^2$ e o limite de $\gamma(u)$ quando $u \to \infty$ é menor do que $Var(\epsilon_j)$.

2.3.1.3 Efeitos Aleatórios e Erro de Medida

Embora a presença da correlação em série para um modelo longitudinal aparente ser algo natural, em certas situações o seu efeito pode ser dominado pela combinação dos efeitos aleatórios e do erro de medida. Se σ^2 é muito mais baixa do que τ^2 ou ν^2, torna-se um refinamento desnecessário do modelo. Assim, eliminando a componente da correlação em série, a partir da equação (2.3.3), obtém-se

$$\epsilon_j = \boldsymbol{d}_j'\boldsymbol{U} + Z_j. \tag{2.3.12}$$

O modelo deste tipo mais simples tem *random intercept* escalar, U, com $d_j = 1$ para todos os j, e assim

$$Var(\boldsymbol{\epsilon}) = \nu^2 J + \tau^2 I. \tag{2.3.13}$$

A variância para cada ϵ_j é $\nu^2 + \tau^2$, e a correlação entre duas medidas quaisquer para o mesmo individuo é $\rho = \nu^2/(\nu^2 + \tau^2)$.

2.3.2 Estimação de Parâmetros

A estimação dos parâmetros dos modelos de efeitos aleatórios é usualmente baseada em métodos de máxima verosimilhança. Uma estratégia adoptada para a determinação de parâmetros para este tipo de modelos é considerar simultaneamente os parâmetros de interesse, β, e os parâmetros da covariância, σ^2 e V_0, usando a função de verosimilhança (Diggle et al., 2002). Recorda-se que V é a matriz diagonal de bloco com blocos comuns diferentes de zero V_0. Por uma questão de simplicidade, será considerada a base de dados

completa $N = nm$. Assim, sob o pressuposto da normalidade (2.3.2), a log-verosimilhança para os dados observados $\boldsymbol{y}$ é

$$L(\boldsymbol{\beta}, \sigma^2, V_0) = -0,5 \left\{ nm \log(\sigma^2) + nm \log(|V_0|) + \sigma^{-2}(\boldsymbol{y} - X\boldsymbol{\beta})' V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) \right\}. \tag{2.3.14}$$

De acordo com Diggle et al. (2002), para um dado V_0, o estimador de máxima verosimilhança para $\boldsymbol{\beta}$ é o estimador dos mínimos quadrados ponderados, dado por

$$\hat{\boldsymbol{\beta}}(V_0) = (X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y}. \tag{2.3.15}$$

Substituindo a equação (2.3.15) em (2.3.14), obtêm-se

$$L(\hat{\boldsymbol{\beta}}(V_0), \sigma^2, V_0) = -0,5 \left\{ nm \log(\sigma^2) + m \log(|V_0|) + \sigma^{-2} RSS(V_0) \right\}, \tag{2.3.16}$$

em que

$$RSS(V_0) = \left\{ \boldsymbol{y} - X\hat{\boldsymbol{\beta}}(V_0) \right\}' V^{-1} \left\{ (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(V_0) \right\}. \tag{2.3.17}$$

Diferenciando a equação(2.3.17) com respeito a σ^2 resulta que o estimador de máxima verosimilhança para σ^2, e V_0 fixo, é

$$\hat{\sigma}^2(V_0) = \frac{RSS(V_0)}{nm}. \tag{2.3.18}$$

A substituição de (2.3.15) e (2.3.17) em (2.3.14) resulta numa log-verosimilhança reduzida para V_0 que, para além de um termo constante, é

$$L_r(V_0) = -0,5m \left\{ n \log RSS(V_0) + \log(|V_0|) \right\}. \tag{2.3.19}$$

Finalmente, a maximização de $L_r(V_0)$ conduz a $\hat{V}_0$ e, por substituição em (2.3.15) e (2.3.17), resulta nos estimadores de máxima verosimilhança, $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\hat{V}_0)$ e $\hat{\sigma}^2 \equiv \hat{\sigma}^2(\hat{V}_0)$.

2.3.2.1 Estimador de Máxima Verosimilhança Restrita

Sob certas condições de regularidade, o estimador de máxima verosimlhança para V_0 será assintoticamente não enviesado (Rizopoulos, 2012). No entanto, em amostras pequenas,

o estimador de máxima verosimilhança para V_0 é enviesado. Este viés surge porque a estimativa de máxima verosimilhança para σ^2 não tem em conta que $\boldsymbol{\beta}$ também é estimado a partir dos dados. Para evitar este problema, o estimador de máxima verosimilhança restrita (REML, do inglês *restricted maximum likelihood*) foi desenvolvido por Harville (1974). Neste tipo de estimação para V_0, é eliminado $\boldsymbol{\beta}$ de forma a que a estimação seja apenas definida em termos de V_0. Assim, o estimador REML para σ^2 é

$$\tilde{\sigma}^2(V_0) = RSS(V_0)/(nm - p), \tag{2.3.20}$$

em que p é o número de elementos de $\boldsymbol{\beta}$. O estimador REML para V_0 maximiza a log-verosimilhança reduzida

$$L^*(V_0) = -\frac{1}{2}m\{n\, log\, RSS(V_0) + log\,(|V_0|)\} - \frac{1}{2}\, log\,(|X'V^{-1}X|). \tag{2.3.21}$$

Finalmente, a substituição do estimador resultante $\tilde{V}_0$ nas equações (2.3.15) e (2.3.20) conduzem aos estimadores REML $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\tilde{V}_0)$ e $\tilde{\sigma}^2 = \tilde{\sigma}^2(\tilde{V}_0)$.

2.4 Dados Omissos

Um aspeto importante na análise de dados longitudinais é o problema dos dados omissos. Dados em falta surgem quando uma ou mais sequências de medidas de cada indivíduo estão incompletas, i.e., as medidas que foram planeadas obter não foram todas registadas (Diggle et al., 2002). Os dados em falta colocam diversos desafios ao planeamento e à análise de dados longitudinais. Este tipo de dados levam à perda de eficiência, visto que as estimativas das evoluções longitudinais médias são menos precisas e, assim, seria necessário incluir mais indivíduos no estudo para manter o poder em detetar efeitos importantes (Rizopoulos, 2012). Esta redução na precisão está diretamente relacionada com a quantidade de dados em falta e também é afetada pelos métodos de análise escolhidos (Rizopoulos, 2012). Para além disso, os dados em falta conduzem a bases de dados não-balanceadas, o que impossibilita o uso de métodos de análise que conseguem apenas lidar com dados cujas medidas foram registadas nos mesmos tempos para todos os indivíduos (Diggle et al., 2002). Por último, em certas circunstâncias e indevidamente manuseados, os dados em falta podem introduzir viés e, consequentemente, levar a inferências incorretas (Rizopoulos, 2012).

Os métodos de análise de dados longitudinais incompletos são determinados pelos mecanismos de dados omissos. Em seguida, serão definidos os diferentes tipos de mecanismos de dados omissos.

Adotando a notação em Diggle et al. (2002), seja $\boldsymbol{Y}^* = (\boldsymbol{Y}^{(0)}, \boldsymbol{Y}^m)$ o conjunto completo das medidas que seriam obtidas se não existissem dados omissos, em que $\boldsymbol{Y}^{(0)}$ designa as medidas obtidas e $\boldsymbol{Y}^{(m)}$ as medidas que seriam obtidas se não existissem dados omissos. Para além disso, seja $\boldsymbol{R}$ o conjunto das variáveis aleatórias que indicam quais elementos de $\boldsymbol{Y}^*$ pertencem a $\boldsymbol{Y}^{(0)}$ ou a $\boldsymbol{Y}^{(m)}$. Assim, a taxonomia do mecanismo de dados omissos é baseada na distribuição de probabilidade de $\boldsymbol{R}$ condicionada em $\boldsymbol{Y}^* = (\boldsymbol{Y}^{(0)}, \boldsymbol{Y}^{(m)})$.

Existem três tipos de mecanismos de dados omissos (Rubin, 1976; Little & Rubin, 1987):

- Dados Omissos Completamente Aleatórios (MCAR, do inglês *Missing Completely at Random*) : $\boldsymbol{R}$ é independente de $\boldsymbol{Y}^{(0)}$ e $\boldsymbol{Y}^{(m)}$, i.e., $p(\boldsymbol{R} \mid \boldsymbol{Y}^*) = p(\boldsymbol{R})$. Por exemplo, quando um paciente se esquece de ir à consulta.

- Dados Omissos Aleatórios (MAR, do inglês *Missing at Random*): $\boldsymbol{R}$ é independente $\boldsymbol{Y}^{(m)}$, i.e, $p(\boldsymbol{R} \mid \boldsymbol{Y}^*) = p(\boldsymbol{R} \mid \boldsymbol{Y}^{(0)})$. Por exemplo, os pacientes abandonam o estudo por recomendação dos médicos, tendo em conta as medidas observadas.

- Dados Omissos Não Aleatórios (MNAR, do inglês *Missing Not at Random*), também designados por dados informativos (Diggle et al., 2002): $\boldsymbol{R}$ é dependente de $\boldsymbol{Y}^{(m)}$, i.e, $p(\boldsymbol{R} \mid \boldsymbol{Y}^*) = p(\boldsymbol{R} \mid \boldsymbol{Y}^{(0)}, \boldsymbol{Y}^{(m)})$. Por exemplo, em estudos de dor, quando o paciente pede mais medicamento porque não consegue tolerar mais a dor.

Se o método de análise é baseado na verosimilhança, as inferências são válidas para MAR ou MCAR (Diggle et al., 2002). Para demonstrar isto, seja $f(\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}, \boldsymbol{r})$ a função de densidade de probabilidade conjunta de $(\boldsymbol{Y}^{(0)}, \boldsymbol{Y}^{(m)}, \boldsymbol{R})$ dada por

$$f(\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}, \boldsymbol{r}) = f(\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}) f(\boldsymbol{r} \mid \boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}). \tag{2.4.1}$$

Para aplicar um método de estimação baseado na verosimilhança, é necessário a função de densidade de probabilidade conjunta das variáveis aleatórias observáveis, $(\boldsymbol{Y}^{(0)}, \boldsymbol{R})$, que é obtida integrando a equação (2.4.1) e que conduz ao seguinte resultado

$$f(\boldsymbol{y}^{(0)}, \boldsymbol{r}) = \int f(\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}) f(\boldsymbol{r} \mid \boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}) d\boldsymbol{y}^{(m)}. \tag{2.4.2}$$

Se o mecanismo de dados omissos é aleatório, $f(\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}, \boldsymbol{r})$ não depende de $\boldsymbol{y}^{(m)}$ e (2.4.2) é

$$\begin{aligned} f(\boldsymbol{y}^{(0)}, r) &= f(\boldsymbol{r} \mid \boldsymbol{y}^{(0)}) \int f(\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(m)}) d\boldsymbol{y}^{(m)} \\ &= f(\boldsymbol{r} \mid \boldsymbol{y}^{(0)}) f(\boldsymbol{y}^{(0)}). \end{aligned} \tag{2.4.3}$$

Por fim, logaritmizando (2.4.3), a função log-verosimilhança é

$$L = log f(\boldsymbol{r} \mid \boldsymbol{y}^{(0)}) + log f(\boldsymbol{y}^{(0)}), \tag{2.4.4}$$

que é maximizada pela maximização separada dos dois termos do lado direito. Visto que o primeiro termo não contém informação sobre a distribuição de $\boldsymbol{Y}^{(0)}$, este termo pode ser ignorado para fazer inferências acerca de $\boldsymbol{Y}^{(0)}$. Esta propriedade é conhecida como ignorabilidade (Rizopoulos, 2012).

2.4.1 *Intermittent Missing* e *Dropout*

De acordo com Diggle et al. (2002), considerando que se pretende obter uma sequência de medidas, $Y_1, ..., Y_n$, num determinado indivíduo, os dados omissos ocorrem como *dropout* se quando Y_j está em falta, também está Y_k para todos $k \geq j$, onde $k, j \in \{1, ..., n\}$. Caso contrário, os dados omissos são *intermittent missing*.

Quando surge *intermittent missing* por mecanismo de censura conhecido, por exemplo, se os valores abaixo de um determinado valor estão em falta, o algoritmo EM (do inglês, *Expectation Maximization*) proposto por Dempster et al. (1977) pode ser usado (Laird, 1988; Hughes, 1999). Quando os valores em falta intermitentes não surgem por censura, a razão para a sua falta é usualmente conhecida, visto que os indivíduos permanecem no estudo, e em alguns casos esta informação tornará possível assumir que se trata de MCAR (Diggle et al., 2002; Sousa, 2011). Para além disso, quando o número de medidas repetidas é grande, os mecanismos de dados omissos associados podem ser assumidos como sendo MCAR ou MAR (Belin et al., 2000). Nestes casos, os dados podem ser analisados usando qualquer método adequado a dados não-balanceados. Adicionalmente, como já foi mecionado, se o método de estimação é baseado na verosimilhança, as inferências também

serão válidas para dados MAR (Diggle et al., 2002).

Little and Rubin (1987) propõem uma forma simples de verificar o pressuposto MCAR, em que a informação das unidades incompletas é usada para avaliar se as unidades completas são uma sub-amostra aleatória da amostra inicial. As observações na medida Y_j são divididas em dois grupos: i) os indivíduos cujas observações foram registadas para outra medida ou conjunto de medidas, e ii) os indivíduos cujas outras medidas estão em falta.

Se houver diferenças significativas entre as distribuições de ambos os grupos, então o pressuposto de MCAR é inválido, e a análise de casos completos conduz potencialmente a estimadores enviesados. A não rejeição da igualdade das distribuições aumenta a evidência de MCAR, mas não confirma que de facto seja este o mecanismo de dados omissos (Molenberghs & Verbeke, 2000). Este teste tem poder limitado quando a amostra de unidades incompletas é pequena e não providenciam evidência para o pressuposto MAR (Little & Rubin, 1987). Para além disso, não é possível testar se os dados omissos dependem dos valores previamente registados, ou seja, se o mecanismo de dados omissos é MAR ou MNAR (Molenberghs & Kenward, 2007). Por este motivo, o analista assume o pressuposto MAR ou MNAR para construir modelos, e só depois é que faz uma análise de sensibilidade para avaliar se este pressuposto é válido (Minini & Chavance, 2004).

Outra dificuldade em testar se o mecanismo presente é MCAR para *intermittent missing* é que a co-existência de *dropout* não pode ser ignorada (Yang & Shoptaw, 2005). Quando o número de pacientes que abandonam o estudo é pequeno, estes pacientes podem ser eliminados para fazer o teste de MCAR (Yang & Shoptaw, 2005).

2.5 Análise de Diagnóstico

De acordo com Diggle et al. (2002), visto que o modelo é essencialmente um modelo para a média e a estrutura de covariância dos dados, uma forma simples e eficaz de avaliar o ajustamento do modelo é construir um gráfico com a média ajustada da variável resposta e a sua média observada ao longo do tempo para cada combinação de tempo e de tratamento, e construir o gráfico com os variogramas ajustado e empírico. Assim, podem ser detetadas discrepâncias entre os dados e o modelo ajustado.

O gráfico dos resíduos versus os valores ajustados é usado para avaliar o pressuposto da variância constante de ϵ_{ij} (Pinheiro & Bates, 2000). O pressuposto da normalidade dos erros é avaliado através do gráfico Q-Q dos resíduos.

2.6 Modelos para Excesso de Zeros

2.6.1 Modelos Hurdle

Um modelo hurdle (Heilbron, 1994; Mullahy, 1986) é um modelo com duas componentes: a dos zeros e a das observações que não são zero, que seguem uma distribuição de contagem convencional, como a de Poisson ou binomial negativa.

Seja Y_{ij} a variável resposta para o paciente i, $i = 1, ..., m$, e tempo $j = 1, ..., n$. A estrutura geral de um modelo hurdle é dada por

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & , \text{se } y_{ij} = 0 \\ (1 - \pi_{ij}) \frac{p(y_{ij}; \boldsymbol{\theta}_{ij})}{1 - p(0; \boldsymbol{\theta}_{ij})} & , \text{se } y_{ij} > 0 \end{cases}, \tag{2.6.1}$$

em que $\pi_{ij} = P(Y_{ij} = 0)$ é a probabilidade do indivíduo pertencer à componente dos zeros; $p(y_{ij}; \boldsymbol{\theta}_{ij})$ representa a distribuição de probabilidade para uma distribuição de contagem regular com vetor de parâmetros $\boldsymbol{\theta}_{ij}$ e $p(0; \boldsymbol{\theta}_{ij})$ é a distribuição para a componente dos zeros. Se a distribuição de contagem segue uma distribuição de Poisson, a distribuição de probabilidade para o modelo hurdle de Poisson é

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & , \text{se } y_{ij} = 0 \\ (1 - \pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} / y_{ij}!}{1 - e^{-\mu_{ij}}} & , \text{se } y_{ij} > 0 \end{cases}. \tag{2.6.2}$$

Noutros casos, a componente que não contém os zeros pode seguir outras distribuições que têm em conta a sobredispersão dos dados, como a binomial negativa. O modelo hurdle binomial negativo (hurdle NB) é dado por

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & , \text{se } y_{ij} = 0 \\ \frac{1 - \pi_{ij}}{1 - \left(\frac{r}{\mu_{ij} + 1}\right)} \frac{\Gamma(y_{ij} + r)}{\Gamma(r) y_{ij}!} \left(\frac{\mu_{ij}}{\mu_{ij} + r}\right)^{y_{ij}} \left(\frac{r}{\mu_{ij} + r}\right)^{r} & , \text{se } y_{ij} > 0 \end{cases}, \tag{2.6.3}$$

em que $(1 + \mu_{ij}/r)$ é uma medida de sobredispersão. À medida que $r \to \infty$, a binomial negativa converge para a distribuição de Poisson.

Ambos os modelos hurdle de Poisson e binomial negativo podem ser usados em regressão, modelando cada componente como uma função de covariáveis. Nestes modelos, as covariáveis que surgem nas duas componentes não são necessariamente iguais.

O modelo de regressão logística é usado para π_{ij} e o modelo log linear é usado para a média μ_{ij} da função $p(y_{ij};\boldsymbol{\theta}_{ij})$. Para além disso, também podem ser adicionados efeitos aleatórios que têm em conta a correlação das medidas do mesmo indivíduo.

Sejam $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})'$ os efeitos aleatórios do modelo, então

$$logit(\pi_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i} \tag{2.6.4}$$

$$log(\mu_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i} \tag{2.6.5}$$

em que $\boldsymbol{x}_{kij}$ e $\boldsymbol{z}_{kij}$ são vetores de covariáveis associados aos efeitos fixos $\boldsymbol{\beta}_k$ e aos efeitos aleatórios $\boldsymbol{b}_{ki}$, respetivamente. Na prática, os modelos com um simples *random intercept* são frequentemente adequados e, neste tipo de modelos, $\boldsymbol{b}_{1i} = b_{1i}$ e $\boldsymbol{b}_{2i} = b_{2i}$ são univariados e $z_{1ij} = z_{2ij} = 1$.

Tendo em conta a potencial associação de medidas, pode ser assumido que os efeitos aleatórios seguem uma distribuição multivariada normal conjunta,

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_{i1} \\ \boldsymbol{b}_{i2} \end{pmatrix} \sim \mathcal{MNV}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \right), \tag{2.6.6}$$

em que Σ_{11} e Σ_{22} representam a variância condicionada de $\boldsymbol{b}_1 = (b_{1i}, ..., b_{1n})^T$ e $\boldsymbol{b}_2 = (b_{2i}, ..., b_{2n})^T$ respetivamente, e Σ_{12} é a covariância entre $\boldsymbol{b}_1$ e $\boldsymbol{b}_2$. A correlação entre $\boldsymbol{b}_1$ e $\boldsymbol{b}_2$ é $\rho = \Sigma_{12}/\Sigma_{11}\Sigma_{22}$ e, quando $\rho = 0$, significa que as duas componentes do modelo hurdle não estão correlaciandas.

Seja $\boldsymbol{\Psi}$ o vetor de parâmetros, $\boldsymbol{\Psi} = (\beta_1, \beta_2, \Sigma)$. A log-verosimilhança marginal para o modelo hurdle de efeitos aleatórios é dada por

$$l(\boldsymbol{\Psi}) = \Sigma_{i=1}^{n}\, log\, L_i(\boldsymbol{\Psi}), \tag{2.6.7}$$

em que

$$\begin{aligned} L_i(\boldsymbol{\Psi}) &= \int \left[\prod_{j=1}^{t_i} \pi_{ij}^{(1-u_{ij})} \left((1-\pi_{ij}) \frac{p_{ij}}{1-p(0)} \right)^{u_{ij}} \right] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \int \left[\prod_{j=1}^{t_i} f_1(u_{ij}|\boldsymbol{b}_{1i}) f_2(y_{ij}, u_{ij}|\boldsymbol{b}_{2i}) \right] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned} \tag{2.6.8}$$

ϕ denota a função de densidade normal para os efeitos aleatórios e

$$u_{ij} = \begin{cases} 0 & \text{, se } y_{ij} = 0 \\ 1 & \text{, se } y_{ij} > 0 \end{cases} . \tag{2.6.9}$$

2.6.2 Modelo de Zeros Inflacionados

O modelo com inflação de zeros assume que as observações de valor zero têm duas origens distintas: estrutural e de amostragem. Os zeros estruturais ocorrem devido à distribuição de Poisson ou binomial negativa, que assumem que que essas observações dos zeros ocorreram por acaso. Os modelos de zero inflacionados assumem que algumas destas observações dos zeros ocorreram devido à estrutura específica dos dados. Os modelos de zeros inflacionados têm a seguinte estrutura geral

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} + (1 - \pi_{ij})p(0; \boldsymbol{\theta}_{ij}) & \text{, se } y_{ij} = 0 \\ (1 - \pi_{ij})p(y_{ij}; \boldsymbol{\theta}_{ij}) & \text{, se } y_{ij} > 0 \end{cases} , \tag{2.6.10}$$

que consiste de uma distribuição degenerada em zero e distribuição de contagem não truncada com vetor de parâmetros $\boldsymbol{\theta_{ij}}$. Se a distribuição de contagem segue uma distribuição de Poisson, o modelo de Poisson inflacionado de zeros (ZIP, do inglês *zero-inflated Poisson model*) é dado por

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} + (1 - \pi_{ij})e^{-\mu_{ij}} & \text{, se } y_{ij} = 0 \\ (1 - \pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} & \text{, se } y_{ij} > 0 \end{cases} , \tag{2.6.11}$$

em que μ é a média da distribuição de Poisson. Tal como nos modelos hurdle, se houver sobredispersão, a distribuição de contagem pode seguir uma distribuição binomial negativa. O modelo binomial negativo inflacionado de zeros (ZINB, do inglês *zero-inflated negative binomial model*) é dado por

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} + (1 - \pi_{ij})\left[\left(\frac{r}{\mu_{ij}+r}\right)^r\right] & \text{, se } y_{ij} = 0 \\ (1 - \pi_{ij})\frac{\Gamma(y_{ij}+r)}{\Gamma(r)y_{ij}!}\left(\frac{\mu_{ij}}{\mu_{ij}+r}\right)^{y_{ij}}\left(\frac{r}{\mu_{ij}+r}\right)^r & \text{, se } y_{ij} > 0 \end{cases} , \tag{2.6.12}$$

Tal como nos modelos hurdle, também podem ser adicionados efeitos aleatórios que têm em conta a correlação das medidas do mesmo indivíduo.

Sejam $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})'$ os efeitos aleatórios do modelo, então

$$logit(\pi_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i} \tag{2.6.13}$$

$$log(\mu_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i} \tag{2.6.14}$$

em que $\boldsymbol{x}_{kij}$ e $\boldsymbol{z}_{kij}$ são vetores de covariáveis associados aos efeitos fixos $\boldsymbol{\beta}_k$ e aos efeitos aleatórios $\boldsymbol{b}_{ki}$, respetivamente.

2.6.3 Análise de Diagnóstico

A análise de diagnóstico é feita através da construção do gráfico Q-Q dos valores observados *versus* os ajustados (Rizopoulos, 2020). Se o modelo estiver bem especificado, os resíduos seguirão uma distribuição uniforme no intervalo (0,1). Também é feito o teste de Kolmogorov-Smirnov para testar a uniformidade da distribuição (Rizopoulos, 2020).

Para além disso, o gráfico de dispersão dos resíduos *versus* dos valores ajustados ajudam a detetar desvios da uniformidade na direção dos eixo dos y (Rizopoulos, 2020). Este gráfico efetua uma regressão quantílica, em que são providenciadas linhas para os quantis 0,25, 0,5 e 0,75. Estas linhas devem ser retas, horizontais e para os valores de y 0,25, 0,5 e 0,75. Alguns desvios são esperados resultantes do acaso, mesmo para um modelo perfeito, especialmente se a amostra é pequena.

Capítulo 3

Análise de Sobrevivência

Em análise de sobrevivência, o objetivo é determinar o tempo até que determinado evento de interesse ocorra, i.e. o tempo de falha (Cox & Oakes, 1984). O evento de interesse pode ser, por exemplo, a falha de determinada componente de uma máquina, o divórcio em estudos sociológicos, o aparecimento de determinada doença, ou o deixar de fumar.

3.1 Funções Básicas em Análise de Sobrevivência

Seja T uma variável aleatória não-negativa e absolutamente contínua que denota o tempo para o evento. A função de distribuição, $F(t)$, que representa a probabilidade de ocorrer o acontecimento de interesse até ao tempo t, é definida como

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du, \tag{3.1.1}$$

em que $f(u)$ denota a função densidade de probabilidade correspondente.

A probabilidade de determinado evento ocorrer após o instante t, i.e. a função de sobrevivência, é

$$S(t) = P(T > t) = \int_t^\infty f(u)\, du. \tag{3.1.2}$$

A função de sobrevivência é monótona não-crescente, com $S(0) = 1$ e $\lim_{t\to\infty} S(t) = 0$.

Outra função frequentemente usada em análise de sobrevivência é a função de risco, que descreve o risco instantâneo para o evento no intervalo de tempo $[t, t + dt)$, tendo em conta que o individuo sobreviveu até ao instante t. A função de risco é definida como

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}, \quad t > 0. \tag{3.1.3}$$

Tendo em conta a relação entre as funções de sobrevivência e de risco, a sobrevivência também pode ser expressa da seguinte forma:

$$S(t) = exp\{-H(t)\} = exp\{-\int_0^t h(u)\, du\}, \tag{3.1.4}$$

em que $H(.)$ é a função de risco cumulativa. A função $H(t)$ também pode ser interpretada como o número esperado de eventos que serão observados até ao tempo t.

3.2 Censura

Uma fonte especial de dificuldade na análise de dados de sobrevivência é que para alguns indivíduos o tempo até à ocorrência do evento de interesse não é observado (Cox & Oakes, 1984). Este fenómeno é conhecido por censura. Rizopoulos (2012) aponta duas implicações da presença de censura. Em primeiro lugar, ferramentas estatísticas como a média amostral e o erro padrão, o teste t e a regressão linear não podem ser usados porque assumem informação completa e, por isso, conduzem a estimadores enviesados da distribuição dos tempo de evento e das quantidades relacionadas. Em segundo lugar, as inferências podem ser mais sensíveis à má especificação da distribuição do tempo de sobrevivência em relação aos dados completos.

A análise dos dados censurados depende do tipo de mecanismo de censura presente. Seguidamente, é apresentada uma classificação dos mecanismos de censura baseada na posição relativa no eixo do tempo da censura e dos verdadeiros tempos de falha:

- Censura à direita: ocorre quando um indivíduo abandona o estudo antes da ocorrência do evento, o estudo termina antes da ocorrência do evento (Censura Tipo I), ou o estudo termina depois de um pré-especificado número de eventos ter sido registado e, consequentemente, o evento de interesse não ocorreu em todos os indivíduos.

- Censura à esquerda: quando o evento de interesse ocorreu antes do indivíduo iniciar o estudo. Por exemplo, antes de frequentarem a primeira classe, algumas crianças já são capazes de soletrar o seu nome .

- Censura intervalar: quando apenas se sabe que o evento de interesse ocorreu num intervalo de tempo. Por exemplo, os pacientes infetados por HIV são frequentemente testados para a presença de SIDA. Quando um paciente obtém um resultado positivo, sabe-se que o paciente contraiu a doença entre as duas últimas visitas.

Outra forma de classificar o mecanismo de censura é ter em conta se a probabilidade de o indivíduo ser censurado depende ou não do processo de falha (Rizopoulos, 2012). Adotando este tipo de classificação, são definidos os seguintes tipos de censura:

- Censura informativa: ocorre quando o indivíduo abandona o estudo por razões diretamente relacionadas com o tempo de falha esperado, por exemplo, quando o seu prognóstico piora. Formalmente, o mecanismo de censura é informativo se para qualquer tempo t, as taxas de falha para os indivíduos que ainda permanecem no estudo são diferentes daqueles que abandonaram o estudo. De certa forma, este mecanismo de censura é similar ao mecanismo MNAR dos estudos longitudinais (Rizopoulos, 2012), mencionado no capítulo anterior.

- Censura não-informativa (ou aleatória): o indivíduo abandona o estudo por razões não relacionadas com o seu prognóstico, mas pode depender das covariáveis. Este mecanismo de censura corresponde ao mecanismo MCAR (Rizopoulos, 2012).

Em relação à primeira categorização, a maior parte da literatura foca-se nos métodos para lidar com censura à direita (Rizopoulos, 2012). Na segunda categorização, quando o mecanismo de censura é informativo pouco pode ser feito no tratamento de dados (Rizopoulos, 2012). Identicamente ao mecanismo MNAR nos estudos longitudinais, o problema é que os dados observados não contém informação suficiente para modelar o mecanismo de censura (Rizopoulos, 2012).

Quando se pretende estimar as funções apresentadas na seção 3.1 ou outra característica da função de distribuição do tempo de falha, a partir de uma amostra aleatória, deve-se ter em conta a censura (Rizopoulos, 2012). Considerando T_i o tempo de falha para o indivíduo i e C_i o tempo de censura para o indivíduo i, então o tempo de vida observado para um indivíduo é uma realização da variável $S_i = min(T_i, C_i)$ (Borges, 2015). Para

além disso, o indicador do evento $\delta_i = I(T_i \leq C_i)$ tem valor 1 se o tempo de falha é observado e 0 caso contrário, em que $I(.)$ é a função indicatriz.

Para realizar uma inferência sobre o modelo paramétrico indexado por um vetor de parâmetros $\boldsymbol{\theta}$ em que o tempo de vida T segue uma distribuição específica, é frequentemente usado o método da máxima verosimilhança. De acordo com Borges (2015), a função de verosimilhança de um vetor de parâmetros, $\boldsymbol{\theta}$, tendo em conta os dados $Y = (T_i, \delta_i)$, $i = 1, ..., N$, em que N é a dimensão da amostra, é

$$L(\boldsymbol{\theta}|Y) = \prod_{i=1}^{N} f(t_i|\boldsymbol{\theta})^{\delta_i} S(t_i|\boldsymbol{\theta})^{1-\delta_i}. \tag{3.2.1}$$

3.3 Modelo de Riscos Proporcionais de Cox

O modelo de riscos proporcionais de Cox (Cox, 1972) é um modelo frequentemente usado em análise sobrevivência. Este modelo assume que as covariáveis têm um efeito multiplicativo no risco de evento, e é formulado como

$$h_i(t|\boldsymbol{z}_i) = \lim_{dt \to 0} P(t \leq T < t + dt | T \geq t, \boldsymbol{z}_i)/dt = h_0(t) exp(\boldsymbol{\beta}^T \boldsymbol{z}_i), \tag{3.3.1}$$

em que $\boldsymbol{z}_i^T = (z_{i1}, ..., z_{ip})$ é o vetor de p covariáveis que estão associadas com o risco de cada indivíduo i, $i = 1, ..., N$, e $\boldsymbol{\beta}$ representa o vetor de coeficientes de regressão correspondente. A função $h_0(t)$ é designada por risco de base e corresponde à função de risco de um indivíduo com $\boldsymbol{\beta}^T \boldsymbol{z}_i = 0$.

Neste tipo de modelo, a função de risco de base não é especificada, i.e. a especificação da distribuição de T_i não é necessária, e o efeito das covariáveis é modelado parametricamente (Klein & Moeschberger, 2006). Assim, o modelo de riscos proporcionais de Cox é não-paramétrico.

O modelo de Cox é designado por modelo de riscos proporcionais porque, se for considerado o vetor de covariáveis $\boldsymbol{z}_i$ para o indivíduo i e o vetor de covariáveis $\boldsymbol{z}_j$ para o indivíduo j, a razão das suas taxas de risco é

$$\frac{h_i(t|\boldsymbol{z}_i)}{h_j(t|\boldsymbol{z}_j)} = exp\{\boldsymbol{\beta}^T(\boldsymbol{z}_i - \boldsymbol{z}_j)\}, \tag{3.3.2}$$

que é uma constante e, assim, as taxas de risco são proporcionais (Klein & Moeschberger, 2006).

Cox (1972) apresentou um método para a estimação dos principais parâmetros de interesse, β, baseado na maximização da função da log-verosimilhança parcial,

$$PL(\boldsymbol{\beta}) = \sum_{i=1}^{N} \delta_i \Big[\boldsymbol{\beta}^T \boldsymbol{z}_i - log\{ \sum_{T_i \geq T_j} exp(\boldsymbol{\beta}^T \boldsymbol{z}_j) \} \Big]. \tag{3.3.3}$$

Para d observações do tempo de falha, sejam $t_1 < t_2 < ... < t_d$ as observações dos tempos ordenadas e $R(t)$ o conjunto de indivíduos em risco para o tempo t, então a função da log-verosimilhança parcial é

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{exp(\boldsymbol{\beta}^T \boldsymbol{z}_i)}{\sum_{k \in R(t_i)} exp(\boldsymbol{\beta}^T \boldsymbol{z}_k)} \,. \tag{3.3.4}$$

A função da log-verosimilhança parcial não depende de $h_0(.)$ e, por isso, não é necessário especificar a sua distribuição. Nesta função, o numerador da verosimilhança depende apenas da informação do indivíduo em que o evento ocorreu, enquanto que o denominador utiliza informação de todos os indivíduos que não sofreram o evento, o que inclui alguns indivíduos que serão censurados mais tarde (Klein & Moeschberger, 2006).

3.4 Testes de Hipóteses para os Coeficientes de Regressão

De acordo com Therneau e Grambsch (2000), os teste de Razão de Verosimilhança, de Wald e de score também estão disponíveis para a verosimilhança parcial de Cox para testar a hipótese nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$. As estatísticas de teste são calculadas da seguinte forma:

- Para o teste de Razão de Verosimilhança:

$$2\{PL(\hat{\boldsymbol{\beta}}) - PL(\boldsymbol{\beta}^{(0)})\}, \tag{3.4.1}$$

  ou seja, duas vezes a diferença da função log-verosimilhança parcial para as estimativas incial, $\boldsymbol{\beta}^{(0)}$, e final, $\hat{\boldsymbol{\beta}}$.

- Para o teste de Wald:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})' \hat{I} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}), \tag{3.4.2}$$

em que $\hat{I} = I(\hat{\boldsymbol{\beta}})$ é a matriz de informação estimada para a equação (3.6) em Therneau e Grambsch (2000), ou o seu equivalente estratificado.

- Para o teste de score:

$$\boldsymbol{U}'(\boldsymbol{\beta}^{(0)})I(\boldsymbol{\beta}^{(0)})^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(0)}), \tag{3.4.3}$$

em que $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} PL(\hat{\boldsymbol{\beta}})$ é o vetor das funções score.

A distribuição da hipótese nula para cada um destes testes é uma qui-quadrado com p graus de liberdade.

3.5 Diagnósticos do Modelo

Para avaliar o pressuposto da proporcionalidade dos ricos pode ser usado um teste proposto por Gambasch & Therneau (1994), que é baseado no cálculo dos resíduos de Shoenfeld.

Também pode ser usado o gráfico dos resíduos de Cox-Snell (Cox & Snell, 1968) para avaliar o pressuposto da proporcionalidade dos riscos. De acordo com Klein and Moeschberger (2006), considerando o modelo de Cox (3.3.1) ajustado aos dados $(T_j, \delta_j, \boldsymbol{z}_j), j = 1, ...n$, e assumindo que $\boldsymbol{z}_j = (z_{j1}, ..., z_{jp})'$ são covariáveis que não variam com o tempo. Então, se o modelo estiver correto e se for considerada a transformação da probabilidade integral no tempo verdadeiro de falha T, a variável aleatória resultante tem distribuição uniforme no intervalo da unidade ou a variável aleatória $U = H(T_j \mid \boldsymbol{z}_j)$ tem distribuição exponencial com taxa de risco 1. Neste caso, $H(T_j \mid \boldsymbol{z}_j)$ é a verdadeira taxa de risco acumulada para um indivíduo com vetor de covariáveis $\boldsymbol{z}_j$.

Se as estimativas de $\boldsymbol{\beta}$ para o modelo (3.3.1) são o vetor $\boldsymbol{b} = (b_1, ..., b_p)'$, então os resíduos de Cox-Snell são definidos por

$$r_j = \hat{H}_0(T_j)\, exp\Big(\sum_{k=1}^{p}(z_{jk}, b_k)\Big), \qquad j = 1, ..., n. \tag{3.5.1}$$

Se o modelo está correto e as estimativas estão próximas dos valores verdadeiros, então os r_j são semelhantes a uma amostra censurada da distribuição exponencial de parâmetro 1. Para verificar isto, é computado o estimador Nelson-Aalen (Nelson, 1972; Aalen, 1976) para a taxa de risco acumulada dos r_j. Se a distribuição exponencial de parâmetro 1 se ajustar aos dados, o gráfico da taxa de risco acumulada de r_j, $\hat{H}_r(r_j)$, contra r_j deve ser aproximadamente uma linha reta que passa na origem com declive 1.

Capítulo 4

Descrição e Análise Exploratória da Base de Dados

4.1 Descrição da Base de Dados

A base de dados usada neste trabalho contém uma *coort naïve* de 1044 pacientes adultos infetados com HIV que iniciaram terapia antirretroviral entre janeiro 2008 e dezembro 2017 num hospital de cuidado terciário e de ensino localizado na região Norte de Portugal. Neste estudo, foram incluídos 642 pacientes da base de dados com HIV do tipo 1 e cuja diferença entre a data de diagnóstico da infeção e a data de início de tratamento foi inferior ou igual a 366 dias.

Este estudo é um estudo longitudinal não balanceado pois os valores das variáveis não foram registadas nos mesmos tempos para todos os pacientes. Na Tabela 4.1, encontram-se as variáveis registadas para cada paciente e posteriormente usadas neste estudo.

Tabela 4.1: Variáveis registadas e usadas neste estudo.

Variáveis Categóricas	**Níveis**
Sexo	Feminino, Masculino
Naturalidade	Portugal, Outros
Tipo de admissão	Consulta, Hospital
Modo de transmissão	Heterossexuais, HSH, Toxicodependentes
Resistências	Sim, Não
Presença de sífilis	Sim, Não
Presença de hepatite	Sim, Não
Presença de infeções oportunistas	Sim, Não
Presença de neoplasia	Sim, Não
Adesão ao tratamento na primeira consulta	Sim, Não
Tipo de tratamento	PIs, NNRTIs, INSTIs

Variáveis Contínuas
Valores de hemoglobina (g/dL)
Valores de leucócitos ($\times 10^9$/L)
Valores de neutrófilos ($\times 10^9$/L)
Valores de linfócitos ($\times 10^9$/L)
Valores de plaquetas ($\times 10^9$/L)
Valores de albumina ($\times 10^9$/L)
Tempo de follow-up (meses)
Idade na primeira consulta (anos)
Idade ao início do tratamento (anos)
Idade em cada consulta de follow-up (anos)
Níveis de células CD4 (células/mm^3)
Níveis de carga viral (cópias/mL)

A amostra analisada contém 155 pacientes do sexo feminino e 487 do sexo masculino. A maior parte dos pacientes era de naturalidade portuguesa (584) e apenas 58 pacientes tinha outra nacionalidade. Dos 642 pacientes, 170 foram admitidos por consulta e 478 pelo

hospital. Também foram registados três modos de transmissão do vírus HIV: heterossexual, homens que têm sexo com outros homens (HSH) e toxicodependentes. Pelo modo de transmissão heterossexual foram registados 382 pacientes, os casos de homens que têm sexo com outros homens foram 213 e 47 pacientes toxicodependentes (Figura 4.1).

Figura 4.1: Número de pacientes por modo de transmissão: heterossexuais, HSH e toxicodependentes.

Foram registados 198 pacientes com sífilis, não havendo registo desta variável para 6 indivíduos, e 69 pacientes com hepatite.

Os pacientes foram sujeitos a três tratamentos: PIs (do inglês, *protease inhibitors*), NNRTIs (do inglês, *non-nucleoside reverse transcriptase inhibitors*) e INSTIs (do inglês, *integrase strand transfer inhibitors*). A maior parte dos indivíduos foram submetidos ao tratamento NNRTIs (284), 144 foram submetidos a PIs e 214 a INSTIs (Figura 4.2). Foi verificado que 549 pacientes adeririam ao tratamento na primeira consulta, enquanto que 90 não adeririam a nenhum tratamento na primeira consulta. A adesão ao tratamento ou não na primeira consulta não foi registada para três indivíduos.

Após a avaliação genérica efetuada pelo médico, foi constatado que a maior parte dos pacientes não apresentaram resistências ao tratamento (525), sendo verificado apenas 117 pacientes com resistências.

Houve 77 indivíduos que desenvolveram infeções oportunistas e 35 que desenvolveram neoplasia. Para a variável neoplasia, não existe registo para 4 indivíduos.

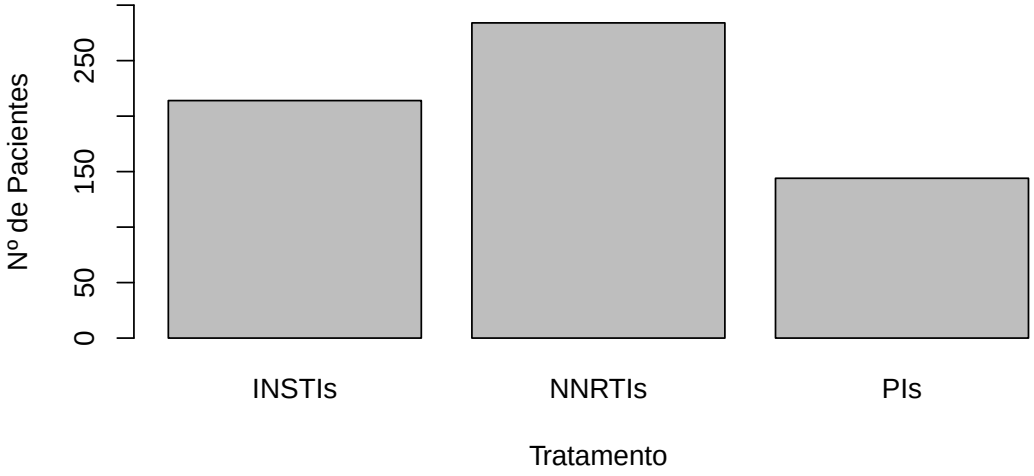


Figura 4.2: Número de pacientes que adeririam aos diferentes tratamentos.

O número de pacientes que no total e em cada grupo de tratamento seguiram sempre o mesmo tratamento, morreram, mudaram de tratamento e perderam-se de follow-up (FU) está representado na Tabela 4.2. Dos 642 pacientes, 444 (69,16%) seguiram sempre o mesmo tratamento, 154 (23,99%) mudaram de tratamento, 28 (4,36%) perderam-se durante o follow-up e 15 (2,34%) faleceram, e não existe este tipo de informação para 1 paciente. O grupo sujeito ao tratamento INSTIs apresenta a percentagem mais baixa de pacientes que mudaram de tratamento (9,81%) em relação aos grupos PIs (37,50%) e NNRTIs (27,82%). A percentagem mais alta de pacientes que seguiram sempre o mesmo tratamento também foi para o grupo de tratamento INSTIs (82,24%).

Tabela 4.2: Número de pacientes que seguiram sempre o mesmo tratamento, morreram, mudaram de tratamento e perderam-se de follow-up (FU) globalmente e por tratamento.

		Tratamento		
	Global	INSTIs	NNRTIs	PIs
Mesmo	444	176	187	81
Mudou	154	21	79	54
Morreu	15	6	4	5
Perdeu-se de FU	28	11	13	4

A distribuição das idades na primeira consulta está representada na Figura 4.3. A idade mínima na primeira consulta em anos é 15,10 e a máxima é 82,6, sendo a média de 42,06.

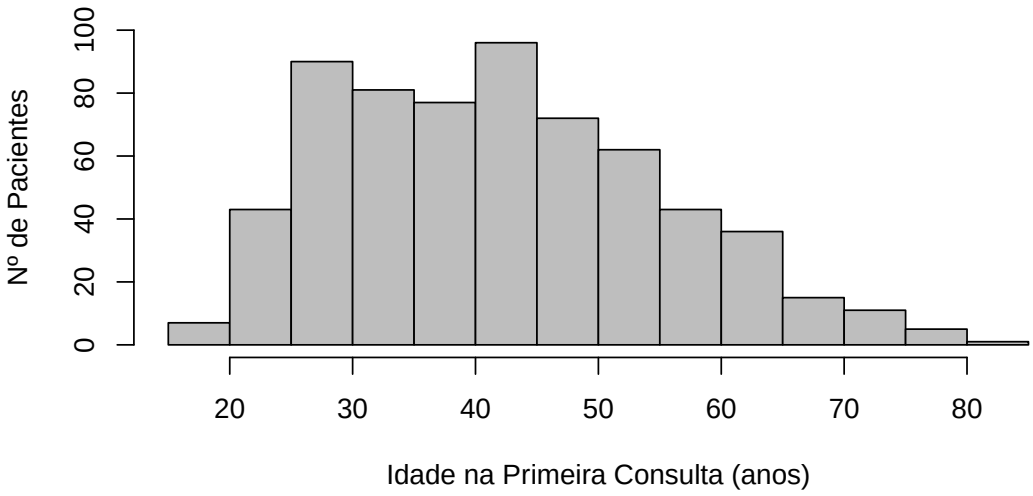


Figura 4.3: Distribuição das idades dos pacientes na primeira consulta.

A distribuição das idades ao início do tratamento está representada na Figura 4.4. A idade mínima em anos ao início de tratamento é 15,90 e a máxima é 82,6, sendo a média de 42,27.

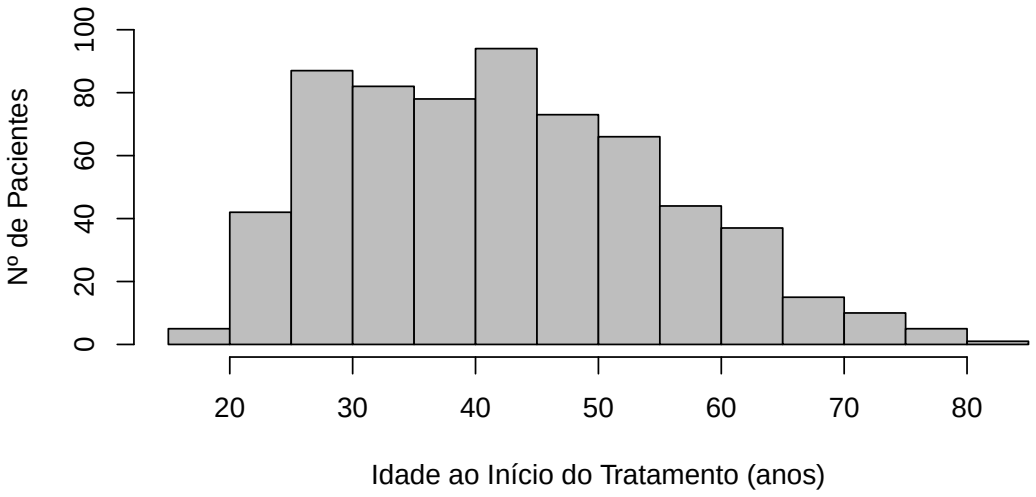


Figura 4.4: Distribuição das idades dos pacientes ao início do tratamento.

A Tabela 4.3 apresenta um breve sumário de algumas estatísticas descritivas para as variáveis contínuas.

Tabela 4.3: Estatísticas descritivas para as variáveis contínuas registadas e usadas neste estudo.

			Tratamento		
		Global	INSTIs	NNRTIs	PIs
Valores de hemoglobina (g/dL) na data do diagnóstico	Mediana	13,60	13,75	13,90	12,50
	Média	13,21	13,24	13,61	12,39
	Mínimo	7,30	7,30	8,10	7,50
	Máxima	19,10	17,30	19,10	16,70
	Variância	4,90	5,05	4,09	5,36
Valores de leucócitos ($\times 10^9$/L) na data do diagnóstico	Mediana	5,06	5,32	5,14	4,50
	Média	5,42	5,68	5,37	5,14
	Mínimo	0,19	1,72	1,38	0,19
	Máxima	21,84	16,50	13,90	21,84
	Variância	5,46	5,07	4,22	8,40
Valores de neutrócitos ($\times 10^9$/L) na data do diagnóstico	Mediana	2,89	2,94	2,83	2,85
	Média	3,34	3,40	3,26	3,44
	Mínimo	0,29	0,50	0,45	0,29
	Máxima	19,33	14,34	12,75	19,33
	Variância	4,00	3,56	2,91	6,87
Valores de linfócitos ($\times 10^9$/L) na data do diagnóstico	Mediana	1,36	1,43	1,43	1,11
	Média	1,60	1,52	1,48	1,96
	Mínimo	0,03	0,18	0,17	0,03
	Máxima	103,00	4,10	4,67	103,00
	Variância	17,10	0,58	0,57	74,84
Valores de plaquetas ($\times 10^9$/L) na data do diagnóstico	Mediana	186,00	188,00	188,50	177,00
	Média	190,60	194,70	193,50	178,50
	Mínimo	10,00	10,00	10,00	10,00
	Máxima	587,00	587,00	546,00	485,00
	Variância	6205,57	6882,48	6186,93	5126,22
Valores de albumina ($\times 10^9$/L) na data do diagnóstico	Mediana	40,10	40,00	41,00	36,80
	Média	38,25	38,63	39,48	35,23
	Mínimo	13,90	16,80	14,60	13,90
	Máxima	57,40	49,50	57,40	50,10
	Variância	51,95	42,20	43,18	71,77

Tabela 4.3: Estatísticas descritivas para as variáveis contínuas registadas e usadas neste estudo (continuação).

		Tratamento			
		Global	INSTIs	NNRTIs	PIs
Idade na primeira consulta (anos)	Mediana	41,20	42,05	40,70	41,70
	Média	42,06	42,48	41,34	42,88
	Mínimo	15,10	19,10	15,10	19,30
	Máxima	82,60	82,60	77,80	73,60
	Variância	173,37	180,16	171,23	183,10
Idade ao início do tratamento (anos)	Mediana	41,40	42,55	41,10	41,60
	Média	42,27	42,65	41,58	43,07
	Mínimo	15,90	19,30	15,90	19,50
	Máxima	82,60	82,60	78,10	73,70
	Variância	170,53	180,18	167,14	181,28
Idade em cada consulta de follow-up (anos)	Mediana	42,72	43,47	39,39	43,11
	Média	43,68	44,11	42,97	44,46
	Mínimo	16,06	19,41	16,06	19,59
	Máxima	83,65	51,91	80,14	74,71
	Variância	173,55	169,93	172,68	177,85
Tempo de follow-up (meses)	Mediana	7,21	6,02	7,66	7,61
	Média	9,06	8,47	9,36	9,21
	Mínimo	0,00	0,00	0,00	0,00
	Máxima	62,00	62,00	43,98	44,54
	Variância	68,09	70,69	67,65	65,48
Níveis de células CD4 (células/mm^3)	Mediana	368,00	467,00	367,00	273,00
	Média	418,60	505,30	411,00	336,60
	Mínimo	1,00	5,00	3,00	1,00
	Máxima	2.037,00	2.037,00	1.598,00	1.953,00
	Variância	77.659,78	94.879,82	62.412,69	70.403,71
Níveis de carga viral (cópias/mL)	Mediana	0	0	0	35
	Média	8123,3	1952,4	7190	16701
	Mínimo	0	0	0	0
	Máxima	2461078	676000	2461078	1670000
	Variância	7,29E+09	595645098	7641222829	1,4094E+10

4.2 Análise Exploratória

A contagem das células CD4 usualmente aumenta após o início do tratamento (Volberding et al., 2010). Analisando a trajetória dos níveis de CD4 para cada paciente, o número de células parece geralmente aumentar ao longo do tempo para todos os tratamentos (Figura 4.5). No entanto, este aumento aparenta ser menos pronunciado para o grupo PIs.

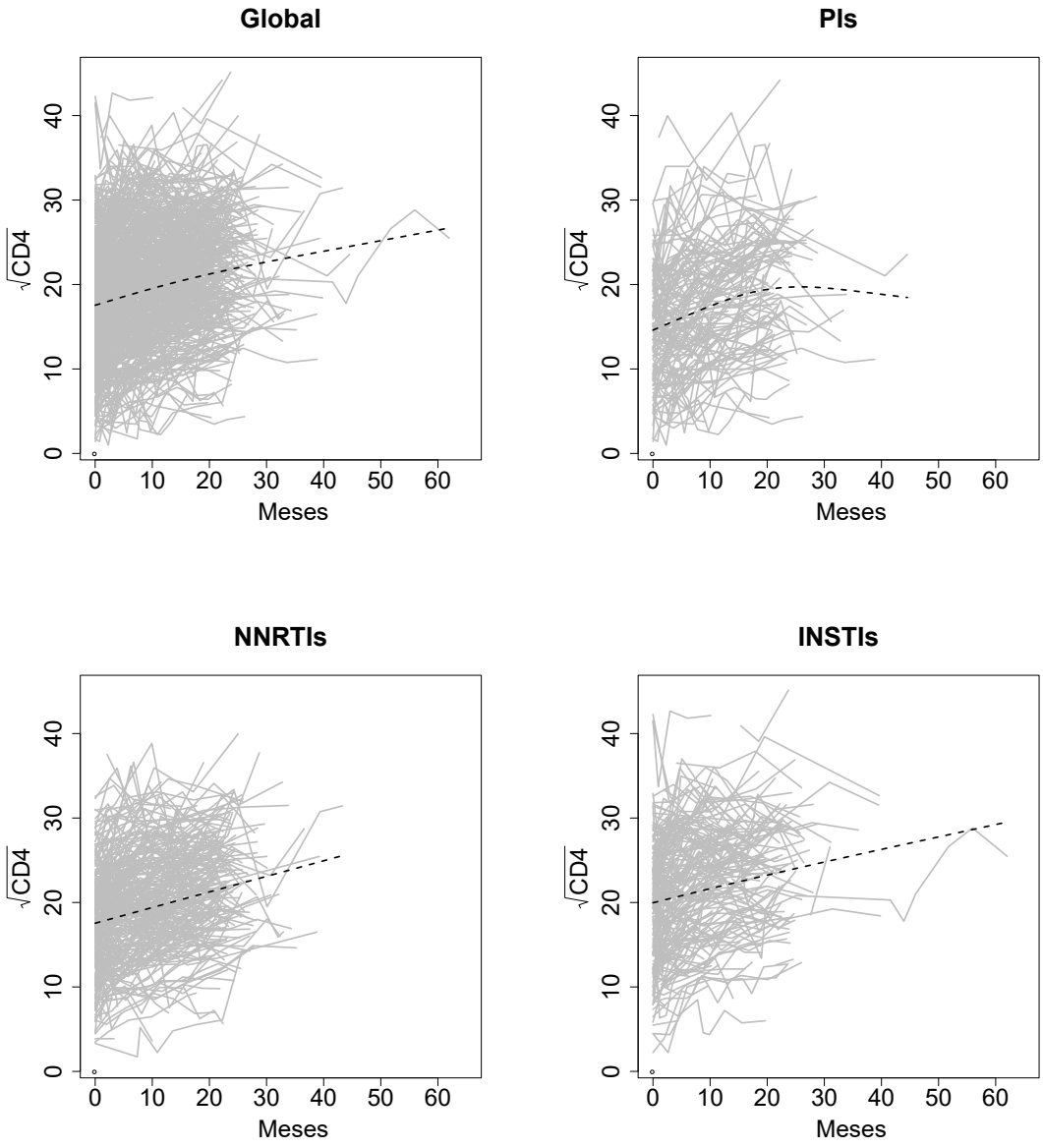


Figura 4.5: Progressão dos níveis das células $\sqrt{CD4}$ por mm^3 ao longo do tempo para cada paciente. A linha a tracejado é uma *spline* cúbica.

Após o início do tratamento, espera-se que a carga viral diminua rapidamente e que desça para níveis inferiores a 50 cópias/mL depois de 12 a 24 semanas de tratamento

(Volberding et al., 2010). A Figura 4.6 sugere que os níveis de carga viral diminuem ao longo do tratamento e que estabilizam, aproximadamente, entre o mês 20 e 30. No entanto, alguns pacientes nunca atingem níveis baixos de carga viral. Quanto às diferenças entre os diferentes tratamentos, o tratamento INSTIs apresenta os níveis mais baixos e mais constantes de carga de viral, o que poderá indicar que este tratamento seja o mais eficaz.

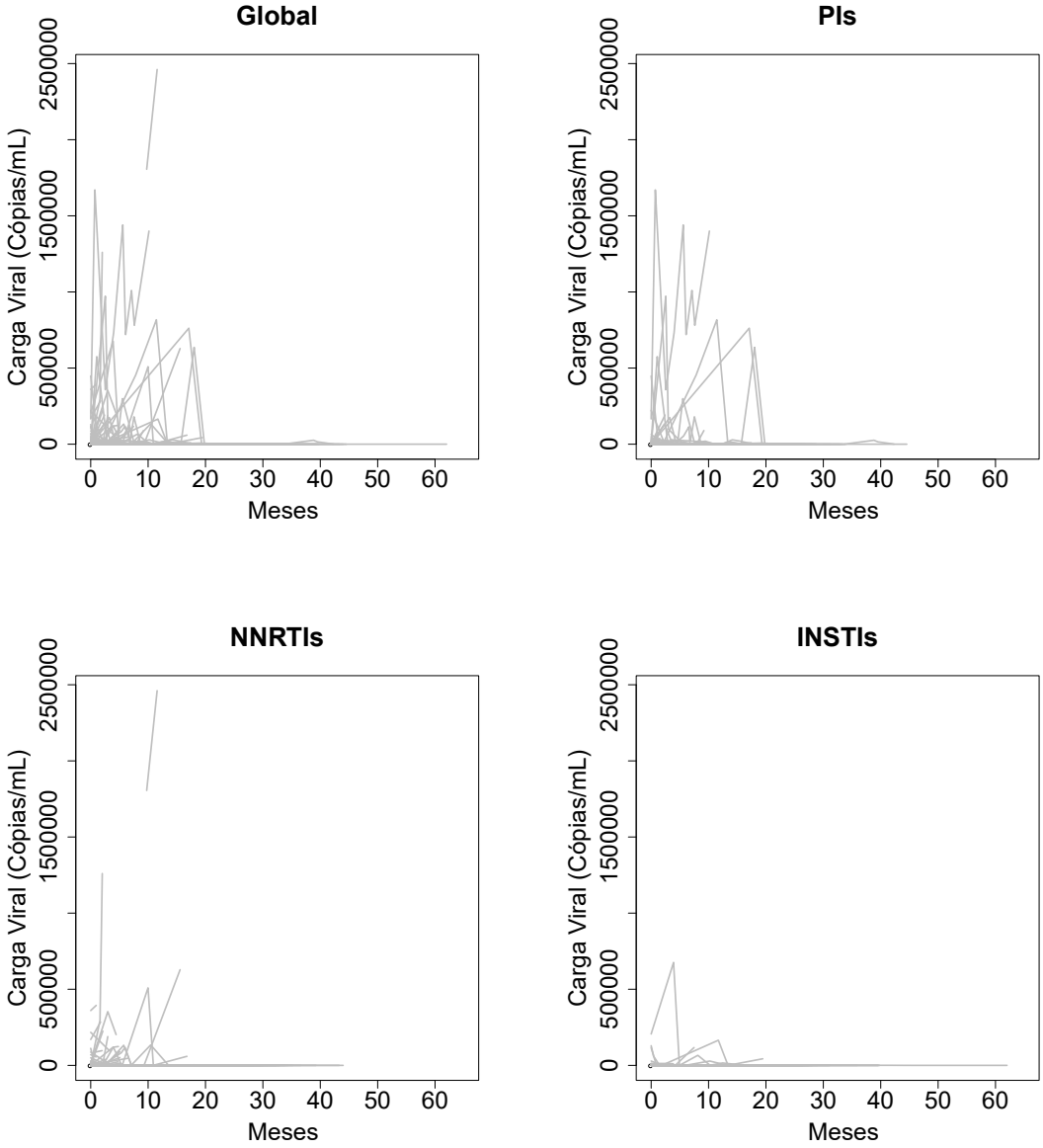


Figura 4.6: Progressão dos níveis da carga viral (cópias/mL) ao longo do tempo para cada paciente.

Existem 444 indivíduos que completaram o tratamento e 154 indivíduos que mudaram de tratamento, embora não tivessem abandonado o estudo (Tabela 4.2). Antes de avançar com a análise de dados, é importante averiguar se estes dois grupos devem ser tratados

como grupos distintos. Analisando a Figura 4.7, parece haver poucas diferenças entre as medianas para quase todas as variáveis contínuas. No entanto, essas diferenças aparentam ser um pouco mais acentuadas para as variáveis Hemoglobina, Albumina, Idade na primeira consulta e Idade ao início do tratamento. Para além disso, existe um grande número de outliers para as variáveis Leucócitos, Neutrófilos, Linfócitos, Plaquetas e Albumina.

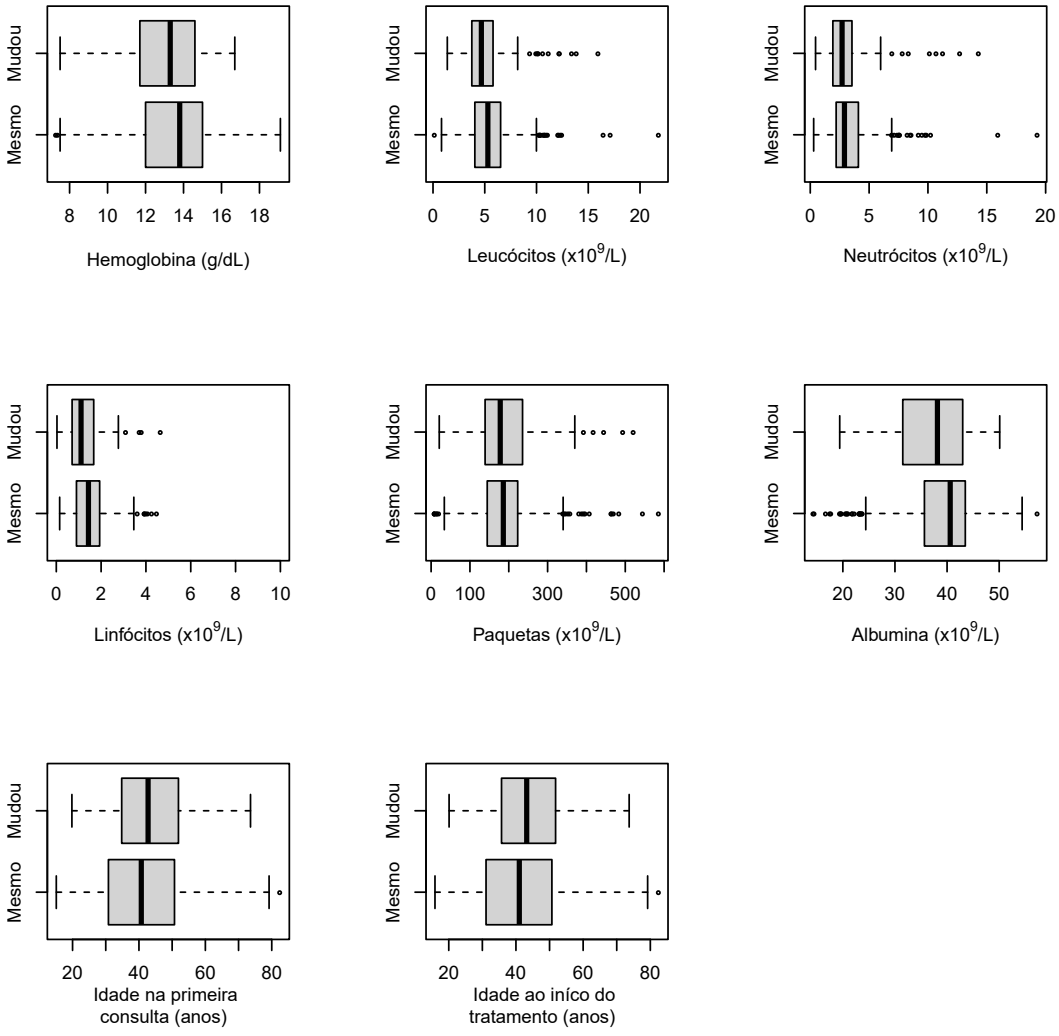


Figura 4.7: Gráficos de caixa de bigodes para as variáveis contínuas que não variam ao longo do tempo.

Seguidamente, foram efetuados testes de Mann-Whitney para verificar se existem diferenças estatisticamente significativas entre as medianas destes dois grupos para cada variável contínua invariável com o tempo. Este teste requer que as distribuições sejam simétricas, o que é verificado (Figura 4.7). Para um nível de significância de 10%, rejeita-

se a hipótese de que as medianas entre os dois grupos sejam iguais para todas as variáveis, exceto para os valores de plaquetas (Tabela 4.4).

Tabela 4.4: Resumo dos testes de Mann-Whitney para as variáveis contínuas que não variam com o tempo.

Variável	Estatística de Teste	Valor de Prova
Valores de hemoglobina	37290	0,0383
Valores de leucócitos	39209	0,0018
Valores de neutrófilos	37020	0,0490
Valores de linfócitos	40182	0,0002
Valores de plaquetas	33682	0,7663
Valores de albumina	37020	0,0092
Idade na primeira consulta	30986	0,0977
Idade ao início do tratamento	30946	0,0934

Para comparar as características dos indivíduos dos dois grupos, foram usados testes de Qui-Quadrado. Para um nível de significância de 10%, as diferenças entre os dois grupos não são estatisticamente significativas para as variáveis sexo, Naturalidade, Modo de transmissão, Resistências e Presença de sífilis (Tabela 4.5). Para as restantes variáveis, existem diferenças significativas entre os dois grupos.

Tendo em conta os resultados obtidos com os testes de Mann-Whitney e de Qui-Quadrado, os grupos que completaram o estudo seguindo sempre o mesmo tratamento e que mudaram de tratamento foram considerados como grupos distintos. Para além disso, visto que o número de indivíduos que abandonaram o estudo foi pequeno, estes também foram eliminados. Assim, todas as análises que se seguem foram feitas usando apenas o grupo que completou o tratamento.

Existem dados em falta do tipo *intermittent missing*, que foi assumidos como sendo MCAR.

Tabela 4.5: Características dos indivíduos que completaram o tratamento e dos que mudaram de tratamento e respetivos testes de Qui-Quadrado.

		Número de Indivíduos		
Variável	Categoria	Mesmo	Mudou	Valor de Prova
Sexo	Feminino	107	37	
	Masculino	337	117	1
Naturalidade	Portugal	408	140	
	Outros	36	14	0,8331
Tipo de admissão	Consulta	98	56	
	Hospital	346	98	0,0007
Modo de transmissão	Heterossexuais	258	100	
	HSH	159	44	
	Toxicodependentes	27	10	0,2609
Resistências	Sim	75	34	
	Não	369	120	0,1884
Presença de sífilis	Sim	130	55	
	Não	310	98	0,1704
Presença de hepatite	Sim	41	13	
	Não	403	141	5,12E-10
Presença de infeções	Sim	42	27	
oportunistas	Não	402	127	0,0106
Presença de neoplasia	Sim	19	13	
	Não	422	141	0,0801
Adesão ao tratamento	Sim	404	121	
na primeira consulta	Não	39	32	0,0001
Tipo de tratamento	PIs	81	54	
	NNRTIs	18	79	
	INSTIs	17	21	2,09E-06

4.2.1 Análise Exploratória para os Indivíduos que Completaram o Tratamento

Para um indivíduo havia dois valores para células CD4 e carga viral registados no mesmo dia. Foi assumido que se tratava de um erro e, por isso, estas duas observações foram

retiradas.

Após a visualização do histograma para a variável células CD4, foi constatado que a sua distribuição não seguia a normalidade. Por esta razão, foi usada a transformação da raiz quadrada para CD4. Esta transformação é bastante comum em estudos com CD4, por exemplo, em Temesgen & Kebede (2016) ou Mcneil & Gore (1996).

A multicolinearidade é definida como a presença de um alto grau de correlação entre as variáveis explicativas (Freund et al., 2006). Na análise exploratória foram identificadas variáveis bastante correlacionadas e, por isso, após a geração do primeiro modelo (de regressão linear simples ou linear misto de efeitos aleatório) foi usado o Fator de Inflação de Variância (VIF) para selecionar as variáveis. De acordo com James et al. (2013), um valor de VIF que exceda 5 ou 10 indica uma quantidade problemática de colinearidade. Assim, todas as variáveis que apresentaram um valor VIF superior a 5 foram retiradas. Na construção do modelo de regressão linear simples assumindo erros independentes foram retiradas as variáveis idade ao início do tratamento e idade na primeira consulta. Os resíduos deste modelo completo foram usados para construir o variograma empírico (Figura 4.8).

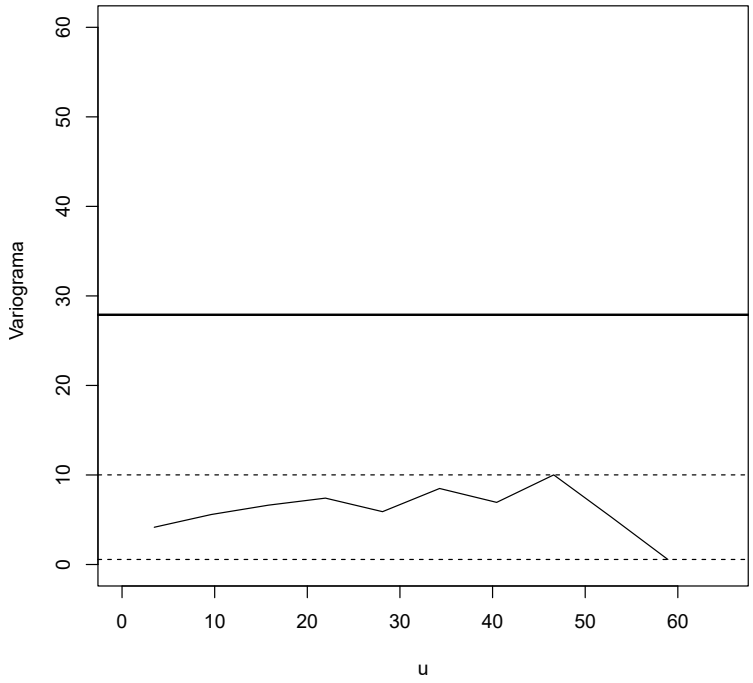


Figura 4.8: Variograma empírico sem pontos para as células CD4.

Como já foi referido na subsecção 2.2.1, o variograma descreve a associacão entre medidas e pode ser decomposto em três componentes de variabilidade: do erro de medida (representada pelo espaço entre as abcissas e a primeira linha horizontal a tracejado), intraindividual (espaço entre as duas linhas horizontais a tracejado) e entre indivíduos (entre a segunda linha a tracejado e linha a negrito). Observando a Figura 4.8, pode ser constatado que a variabilidade dentro do mesmo indivíduo geralmente aumenta com a diferença entre os tempos, u. Isto sugere que as medidas de um indivíduo em tempos mais próximos estão mais correlacionadas entre si. A diminuição da variabilidade dentro do mesmo indivíduo quando o u se aproxima de 50 pode ser explicada pela diminuição do número de dados e por o indivíduo atingir valores mais constantes. Através do variograma empírico, ainda pode ser constatado que a variabilidade entre indivíduos é grande e que é maior do que a variabilidade intraindividual e a variabilidade do erro de medida.

Capítulo 5

Modelos de Efeitos Aleatórios para os Biomarcadores

5.1 Modelo para as Células CD4

Na construção dos diversos modelos de efeitos aleatórios foram retiradas as variáveis idade ao início do tratamento e idade em cada consulta de follow-up. Seguidamente, foi feita a análise longitudinal com as restantes variáveis explicativas. Foram ajustadas diferentes estruturas de correlação para o modelo longitudinal tendo em conta que os dados são não balanceados e o variograma empírico. Assim, os seguintes modelos foram ajustados: com um fator aleatório a nível individual, com dois efeitos aleatórios correlacionados ao nível da especificidade individual (que têm em conta a variação aleatória entre indivíduos e a variância para cada indivíduo ao longo do tempo), um efeito aleatório a nível individual e uma estrutura de correlação temporal contínua no tempo Gaussiana e Exponencial.

Em seguida, foi usado o teste de Razão de Verosimilhança para comparar os modelos com um fator aleatório a nível individual e com dois efeitos aleatórios correlacionados ao nível da especificidade individual. Para um nível de significância 5%, estes dois modelos são estatisticamente diferentes (valor de prova $< 0,0001$), então escolhe-se o modelo com o maior número de parâmetros: com dois efeitos aleatórios correlacionados ao nível da especificidade individual. Os restantes modelos foram comparados usando o Critério de Informação de Akaike (AIC, do inglês *Akaike Information Criterion*) . O modelo escolhido para modelar CD4 foi o modelo com um efeito aleatório a nível individual e uma estrutura de correlação temporal contínua no tempo Exponencial, pois foi aquele que apresentou

menor AIC (Tabela 5.1).

Tabela 5.1: AIC para os modelos ajustados.

Modelo	Graus de liberdade	AIC
1- Com dois efeitos aleatórios correlacionados ao nível da especificidade individual	25	13041,19
2- Com um efeito aleatório e uma estrutura de correlação exponencial contínua no tempo	25	12988,73
3- Com um efeito aleatório e uma estrutura de correlação gaussiana contínua no tempo	25	12990,15

O modelo escolhido tem coeficientes que não são significativos. Seguindo o método descrito em Zuur et al. (2009), as variáveis explicativas foram retiradas uma a uma até que foi obtido um modelo só com coeficientes significativos, começando com aquela cujo teste t para a estimativa do coeficiente tem maior valor de prova. Também pode ser usado o método *stepwise backward* baseado no AIC para a seleção de variáveis e, neste caso, o modelo após a seleção é idêntico ao modelo obtido pelo método anterior. Seguidamente, o modelo com todas as variáveis explicativas foi comparado com o modelo só com coeficientes significativos usando o teste de Razão de Verosimilhança, e não se rejeitou a hipótese de que os dois modelos sejam iguais (valor de prova =0,9085, nível de significância de 5%). Por fim, o modelo escolhido foi novamente ajustado usando o método da máxima verosimilhança restrita. Assim, o modelo final obtido foi

$$\begin{aligned}\sqrt{}(Y_{ij}) = {} & 5,9766 + 2,3390 \cdot se(Sexo = Feminino) + 0,3430 \cdot Hemoglobina \\ & + 1,8759 \cdot Leucócitos - 1,6064 \cdot Neutrófilos + 0,1844 \cdot Albumina \\ & - 2,2410 \cdot se(Tratamento = NNRTIs) - 2,7281 \cdot se(Tratamento = PIs) \\ & - 0,0764 \cdot Idade \quad na \quad 1^{a} \quad consulta - 0,194 \cdot Tempo_{ij} + U_i^1 + W_i(t_{ij}) + Z_{ij},\end{aligned} \tag{5.1.1}$$

em que $U_i^1 \sim \mathcal{N}(0; 2,2426^2)$, $Z_{ij} \sim \mathcal{N}(0; 3,1404)$ e $W_i(t_{ij})$ é um processo Exponencial contínuo no tempo, onde $i = 1, 2, ..., 444$ e $j \in \mathbb{R}_{\geq 0}$.

A estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente estão na Tabela 5.2.

Tabela 5.2: Estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente do modelo para CD4.

	Estimativa	Erro Padrão	Estatística t	P-valor
(*Intercept*)	5,9776	2,4236	2,4664	0,0137
Sexo Feminino	2,3390	0,6346	3,6856	0,0000
Hemoglobina	0,3430	0,1847	1,8571	0,0640
Leucócitos	1,8759	0,2400	7,8176	0,0000
Neutrófilos	-1,6064	0,3005	-5,3464	0,0000
Albumina	0,1844	0,0587	3,1423	0,0018
Tratamento NNRTIs	-2,2410	0,5573	-4,0209	0,0001
Tratamento PIs	-2,7281	0,7348	-3,7130	0,0002
Idade na 1ª consulta	-0,0764	0,0202	-3,7817	0,0002
Tempo	0,1941	0,0090	21,6064	0,0000

O coeficiente de interseção significa que o paciente do sexo masculino, que segue o tratamento INSTIs e cujos valores de hemoglobina, leucócitos, neutrófilos, albumina, idade na primeira consulta e tempo sejam zero teria um valor de células CD4 de $5,9776^2$.

De acordo com o modelo, o tratamento mais eficaz é INSTIs. Um paciente que siga o tratamento NNRTIs teria um valor de células $\sqrt{CD4}$ 2,410 inferior em relação ao tratamento INSTIs, mantendo as restantes variáveis constantes. Para o tratamento PIs, esta diferença seria 2,7281 inferior em relação ao tratamento INSTIs.

Os valores de células $\sqrt{CD4}$ diminuem 1,6064 por cada unidade de aumento de neutrófilos, mantendo as restantes variáveis constantes. Enquanto que os valores de células $\sqrt{CD4}$ aumentam 0,3430 por cada unidade de aumento de hemoglobina, mantendo as restantes variáveis constantes. Em relação à variável leucócitos, os valores de células $\sqrt{CD4}$ aumentariam 1,8759, por cada unidade de aumento de leucócitos, mantendo as restantes variáveis constantes.

Os valores de células $\sqrt{CD4}$ aumentam 0,1844 por cada unidade de aumento de albumina, mantendo as restantes variáveis constantes. Os valores de células $\sqrt{CD4}$ diminuem 0,0764 por cada unidade de aumento da idade na primeira consulta, mantendo as restantes variáveis constantes. Os valores de células $\sqrt{CD4}$ aumentam 0,1941 por cada mês que o paciente continua o tratamento, mantendo as restantes variáveis constantes.

O modelo de efeitos mistos escolhido tem um efeito aleatório a nível individual e uma estrutura de correlação exponencial para descrever a variabilidade ao longo do tempo

para cada paciente, em que a variabilidade entre indivíduos, $\hat{\nu}^2$, é $2,2426^2$, a estrutura de correlação exponencial para descrever a variabilidade ao longo do tempo para cada paciente tem $\hat{\rho} = exp(-\frac{1}{\phi}.|u|) = exp(-\frac{1}{81,9653}.|u|)$, a variabilidade dentro do mesmo indivíduo, $\hat{\sigma}^2$, é $22,9065$ e a variância do erro de medida,$\hat{\tau}^2$, é $3,1404$.

Analisando os variogramas empírico e teórico (Figura 5.1), pode ser constatado que para distâncias inferiores a 20 meses o modelo escolhido ajusta-se bem aos dados. A escassez de dados para distâncias temporais longas pode explicar a discrepância entre o variograma teórico e o empírico nestes pontos.

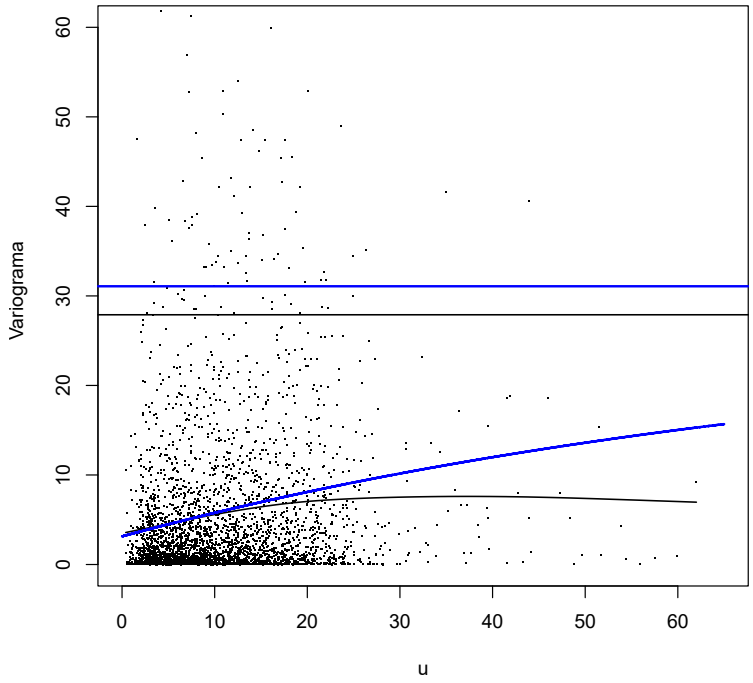


Figura 5.1: Variogramas empírico e teórico para as células CD4. A linha azul representa o variograma teórico para o modelo com um efeito aleatório a nível individual e uma estrutura de correlação exponencial contínua no tempo.

Os pressupostos do modelo longitudinal são verificados. Os erros seguem uma distribuição normal e verifica-se a homogeneidade das variâncias dos erros (Figura 5.2). Também foi feito o teste de Kolmogarov-Smirnov para avaliar a normalidade dos erros. Para um nível de significância de 5%, não é rejeitada a hipótese de que os erros sigam uma distribuição normal (valor de prova = 0,1066).

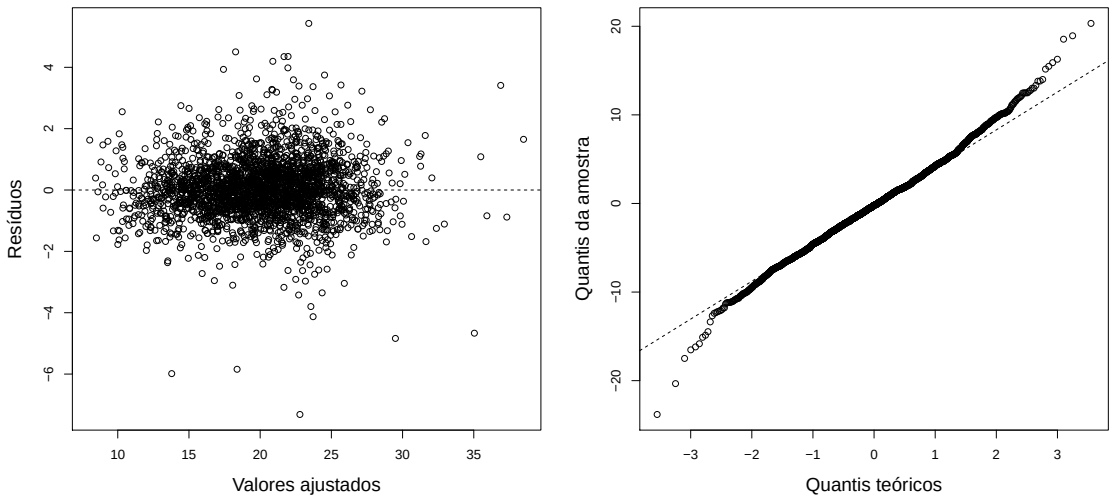


Figura 5.2: Análise de resíduos para o modelo das células CD4.

5.2 Modelos para a Carga Viral

Através do histograma para a Carga viral, foi constatado que a sua distribuição não seguia a normalidade e que há um elevado número de zeros (1617 em 2760 observações). Assim, mesmo após a aplicação de transformação, a Carga viral não segue uma distribuição normal.

Existem vários modelos de efeitos aleatórios que têm em conta a existência de excesso de zeros nos dados, tais como o Modelo Poisson Inflacionado de Zeros (ZIP) de efeitos aleatórios, o Modelo Binomial Negativo Inflacionado de Zeros (ZINB) de efeitos aleatórios, o Modelo Hurdle Poisson de efeitos aleatórios e Modelo Hurdle binomial negativo de efeitos aleatórios. Estes modelos poderão ser uma alternativa viável na análise destes dados.

5.2.1 Modelo de Poisson com Efeitos Aleatórios

Primeiro, os seguintes modelos de Poisson foram ajustados: com um fator aleatório a nível individual e com dois efeitos aleatórios correlacionados ao nível da especificidade individual (que têm em conta a variação aleatória entre indivíduos e a variância para cada indivíduo ao longo do tempo). A Carga viral foi considerada como variável resposta e as restantes variáveis da Tabela 4.1 (exceto as Células CD4) foram usadas como variáveis explicativas.

Para evitar a elevada correlação entre as variáveis explicativas, foi novamente usado o VIF para selecionar as variáveis a incluir no modelo. Assim, as variáveis explicativas Idade na primeira consulta e Idade em cada consulta de follow-up para o modelo com um efeito aleatório foram retiradas e para o modelo com dois efeitos aleatórios foram retiradas as variáveis Idade ao início de tratamento e Idade na primeira consulta.

Seguidamente, foi usado o Critério de Informação de Akaike (AIC) para comparar os dois modelos não-aninhados. Assim, escolhe-se o modelo que apresenta um menor AIC: o modelo de Poisson com dois efeitos aleatórios correlacionados ao nível da especificidade individual (Tabela 5.3).

Para verificar se existe sobredispersão dos dados pode ser calculada a soma dos resíduos de Pearson ao quadrado e comparar esta soma com os graus de liberdade dos resíduos (Zuur et al., 2013; Bolker, 2020). Neste caso, o valor obtido foi $2,3914 \times 10^{15}$, o que indica que há sobredispersão. Caso contrário, o valor seria próximo de 1.

5.2.2 Modelo Binomial Negativo com Efeitos Aleatórios

Para lidar com a sobredispersão dos dados, foi seguidamente usado o modelo binomial negativo com efeitos aleatórios. Na aplicação deste modelo, foi necessária a transformação da raiz quadrada para a variável resposta Carga viral. Os seguintes modelos binomial negativos foram ajustados: com um fator aleatório a nível individual e com dois efeitos aleatórios correlacionados ao nível da especificidade individual (que têm em conta a variação aleatória entre indivíduos e a variância para cada individuo ao longo do tempo).

De modo a evitar a multicolinearidade, foi novamente usado o VIF para selecionar as variáveis explicativas a incluir no modelo. Assim, as variáveis Idade ao início de tratamento e Idade em cada consulta de follow-up para o modelo com um efeito aleatório a nível individual e as variáveis Idade ao início de tratamento e Idade na primeira consulta para o segundo modelo foram retiradas.

Em seguida, foi usado o AIC para comparar os dois modelos anteriores. Então, escolhe-se o modelo com o menor AIC: o modelo binomial negativo com dois efeitos aleatórios correlacionados ao nível da especificidade individual (Tabela 5.3).

5.2.3 Modelos para Excesso de Zeros

Os modelos de Poisson e binomial negativos mencionados até agora foram construídos usando o pacote *lme4* do R. Seguidamente, vários modelos de efeitos aleatórios que têm em conta a existência de excesso de zeros foram ajustados usando o pacote *GLMMadaptative* do R. A lista dos modelos construídos está na Tabela 5.3, com os seus respetivos AIC, indicação do pacote usado, as variáveis explicativas e as transformações da variável resposta usadas. Alguns modelos construídos previamente com o pacote *lme4* foram novamente construídos usando *GLMMadaptative*. Isto foi feito porque, para comparar os modelos usando o critério AIC, é necessário que a variável resposta esteja transformada da mesma forma ou não sofra transformação (Akaike, 1974) e os algoritmos dos dois pacotes são diferentes, o que permite a construção de modelos diferentes com diferentes transformações para a Carga viral.

Ainda na Tabela 5.3, pode ser constatado que AIC do modelo 3 (de Poisson) é inferior ao do modelo 5 (binomial negativo) e o AIC do modelo 7 (binomial negativo) é superior aos dos modelos 8 (binomial negativo inflacionado de zeros) e 9 (hurdle binomial negativo), sendo o valor deste último o mais baixo. Assim, o modelo escolhido para modelar a Carga

viral é o modelo 9, o modelo hurdle binomial negativo de efeitos aleatórios.

Tabela 5.3: Modelos para a Carga viral.

Modelo	Distribuição	Transformação da Carga viral	Pacote	Efeitos aleatórios	AIC	Varriáveis Explicativas
1	Poisson	Sem	*lme4*	Um efeito aleatório ao nível da especificidade individual	14411306	Sexo , Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade ao início de tratamento, Tempo, Infeções oportunistas
2	Poisson	Sem	*lme4*	Dois efeitos aleatórios que têm em conta a variação aleatória entre indivíduos e a variância para cada indivíduo ao longo do tempo	9576127	Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade em cada consulta de follow-up, Tempo, Infeções oportunistas
3	Poisson	Raiz quadrada	*GLMMadaptative*	Dois efeitos aleatórios que têm em conta a variação aleatória entre indivíduos e a variância para cada indivíduo ao longo do tempo	30902,63	Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade em cada consulta de follow-up, Tempo, Infeções oportunistas

Tabela 5.3: Modelos para a Carga viral (continuação).

Modelo	Distribuição	Transformação da Carga viral	Pacote	Efeitos aleatórios	AIC	Varriáveis Explicativas
4	Binomial negativo	Raiz quadrada	*lme4*	Um efeito aleatório ao nível da especificidade individual	11916,28	Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade à primeira consulta, Tempo, Infeções oportunistas
5	Binomial negativo	Raiz quadrada	*lme4*	Dois efeitos aleatórios que têm em conta a variação aleatória entre indivíduos e a variância para cada indivíduo ao longo do tempo	11496,26	Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade em cada consulta de follow-up, Tempo, Infeções oportunistas
6	Binomial negativo	Raiz quadrada	*GLMMadaptative*	Dois efeitos aleatórios correlacionados ao nível da especificidade individual	11464,78	Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade em cada consulta de follow-up, Tempo, Infeções oportunistas

Tabela 5.3: Modelos para a Carga viral (continuação).

Modelo	Distribuição	Transformação da Carga viral	Pacote	Efeitos aleatórios	AIC	Variáveis Explicativas
7	Binomial negativo	Sem	*GLMMadaptative*	Dois efeitos aleatórios que têm em conta a variação aleatória entre indivíduos e a variância para cada indivíduo ao longo do tempo	18142,02	Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade em cada consulta de follow-up, Tempo, Infeções oportunistas
8	ZIBN	sem	*GLMMadaptative*	Parte dos zeros: Um efeito aleatório ao nível da especificidade individual Parte dos valores positivos: Um efeito aleatório ao nível da especificidade individual	17920,13	Para as duas componentes: Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade à primeira consulta, Tempo, Infeções oportunistas
9	Hurdle binomial negativo	sem	*GLMMadaptative*	Parte dos zeros: Um efeito aleatório ao nível da especificidade individual Parte dos valores positivos: Um efeito aleatório ao nível da especificidade individual	17917,71	Para as duas componentes: Sexo, Naturalidade, Admissão, Transmissão, Resistências, Hemoglobina, Neutrófilos, Leucócitos, Linfócitos, Plaquetas, Albumina, Sífilis, Hepatite, Adesão, Tratamento, Idade na primeira consulta, Tempo, Infeções oportunistas

5.2.4 Modelo Escolhido para a Carga Viral

Como já foi referido anteriormente o modelo escolhido para a Carga viral foi o modelo hurdle binomial negativo de efeitos aleatórios.

A seleção das variáveis do modelo escolhido foi feita retirando as variáveis uma a uma, começando com aquelas cujo teste para avaliar a sua significância apresenta maior valor de prova, até ser obtido um modelo só com variáveis significativas. Seguidamente, foi feito um teste de Razão de Verosimilhança para avaliar se existem diferenças significativas entre o modelo antes da seleção e o modelo após seleção. Para um nível de significância de 10%, os dois modelos não apresentam diferenças significativas (valor de prova $> 0,10$). Assim, o modelo final obtido está nas Tabelas 5.5, 5.4 e 5.6. Este modelo pode ser escrito como

$$\begin{aligned} logit(\pi_{ij}) = \beta_{10} + \beta_{11}\, se(Sexo = Feminino) + \beta_{12}\, se(Transmissão = HSH) \\ +\beta_{13}\, se(Transmissão = Toxicopendentes) - \beta_{14}\, Neutrófilos + \beta_{15}\, Albumina \\ -\beta_{16}\, se(Adesão\ na\ 1^a\ consulta = Não) - \beta_{17} se(Tratamento = NNRTIs) \\ -\beta_{18}\, se(Tratamento = PIs) - \beta_{10}\, Tempo_{ij} + b_{1i}, \end{aligned} \tag{5.2.1}$$

$$\begin{aligned} log(\mu_{ij}) = \beta_{20} + \beta_{21}\, Hemoglobina + \beta_{22}\, Albumina \\ -\beta_{23}\, se(Adesão\ na\ 1^a\ consulta = Não) - \beta_{24}\, se(Tratamento = NNRTIs) \\ -\beta_{25}\, se(Tratamento = PIs) - \beta_{26}\, Tempo_{ij} + b_{2i}, \end{aligned} \tag{5.2.2}$$

em que os betas indicam os coeficientes da regressão e (b_{1i}, b_{2i}) são os efeitos aleatórios para o paciente $i = 1, ..., 422$ e para o instante $j \in \mathbb{R}_{\geq 0}$.

Tabela 5.4: Estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente para a componente dos zeros do modelo para a Carga viral.

	Estimativa	Erro Padrão	Valor Z	P-valor
$logit(\pi_{ij})$				
(*Intercept*)	-3,3219	0,5439	-6,1079	$< 1 \times 10^{-4}$
Sexo Feminino	0,4421	0,1848	2,3919	0,0168
Transmissão HSH	0,1300	0,1770	0,7342	0,4628
Transmissão Toxicopendentes	1,1203	0,3195	3,5062	0,0005
Neutrófilos	0,1495	0,0424	3,5229	0,0004
Albumina	0,0461	0,0120	3,8393	0,0001
Adesão na 1ª consulta - Não	-2,0682	0,2910	-7,1070	$< 1 \times 10^{-4}$
Tratamento NNRTIs	0,2640	0,1674	1,5765	0,1149
Tratamento PIs	-0,5269	0,2185	-2,4117	0,0159
Tempo	0,1693	0,0089	18,9874	$< 1 \times 10^{-4}$

Tabela 5.5: Estimativa dos coeficientes do modelo e os testes de significância para cada coeficiente para a componente dos valores positivos do modelo para a Carga viral.

	Estimativa	Erro Padrão	Valor z	P-valor
$log(\mu_{ij})$				
(*Intercept*)	7,4872	0,5765	13,6128	$< 1 \times 10^{-4}$
Hemoglobina	-0,1118	0,0565	-1,9801	0,0477
Albumina	-0,0318	0,0183	-1,7425	0,0814
Adesão na 1ª consulta - Não	2,2333	0,2974	7,5096	$< 1 \times 10^{-4}$
Tratamento NNRTIs	0,975	0,2078	4,6919	$< 1 \times 10^{-4}$
Tratamento PIs	1,6629	0,2505	6,6387	$< 1 \times 10^{-4}$
Tempo	-0,1281	0,0075	-16,9875	$< 1 \times 10^{-4}$

Tabela 5.6: Componentes da variância para o modelo hurdle binomial negativo de efeitos aleatórios.

Componentes da variância	
var(b_{1i})	1,8225
var(b_{2i})	1,1617
cov(b_{1i}, b_{2i})	0,7903
corr(b_{1i}, b_{2i})	-0,5432

De acordo com o modelo apresentado, os fatores associados a valores de carga viral igual a zero são Sexo, Modo de transmissão, Neutrófilos, Albumina, Adesão, Tipo de tratamento e Tempo (Tabela 5.4).

Mantendo as restantes variáveis constantes, a chance de um paciente atingir valores de carga viral iguais a zero é exp(0,4421)=1,5560 vezes superior para um paciente do sexo feminino em relação a um do sexo masculino (Tabela 5.4). Também a chance de um paciente atingir valores de carga viral iguais a zero é exp(1,1203)=3,0658 vezes superior para um paciente cuja a transmissão foi por via toxicodependente em relação à transmissão heterossexual. A chance de o paciente atingir níveis zero de carga viral aumenta $exp(1,1495) = 1,1613$ vezes por cada aumento de uma unidade de Neutrófilos. A chance de o paciente atingir níveis zero de carga viral aumenta $exp(0,0461) = 1,0472$ vezes por cada aumento de uma unidade de Albumina. A chance de o paciente atingir níveis zero de carga viral aumenta $exp(0,1693) = 1,1845$ vezes por cada aumento de uma unidade de Tempo.

A chance de o paciente atingir níveis zero de carga viral diminui $[1 - exp(-2,0682)] \times 100\% = 87,36\%$ se o paciente não aderir ao tratamento na primeira consulta em relação àqueles que aderiram na primeira consulta.

A chance de o paciente atingir níveis zero de carga viral diminui $[1 - exp(-0,5269)] \times 100\% = 40,96\%$ se o paciente seguir o tratamento PIs em relação aos pacientes que seguem o tratamento INSTIs.

A chance de o paciente atingir valores de carga viral zero diminui $[1 - exp(-3,3219)] \times 100\% = 96,39\%$ quando o paciente é do sexo masculino, o modo de transmissão foi por via heterossexual, aderiu ao tratamento na primeira consulta, segue o tratamento INSTIs e os valores de neutrófilos, albumina e o tempo são zero.

O aumento de Albumina e de Tempo de tratamento aumentam as chances de o paciente atingir valores de carga viral zero (Tabela 5.4), e a diminuição de Albumina e de Tempo de tratamento aumentam as chances de o paciente ter valores de carga viral superior a zero (Tabela 5.5). A diminuição da Hemoglobina aumenta a chance de o paciente ter valores de carga viral superiores a zero (Tabela 5.5).

A não adesão ao tratamento na primeira consulta e os tratamentos NNRTIs e PIs (em relação ao INSTIs) aumentam as chances de o doente apresentar cargas virais superiores a zero (Tabela 5.5).

A não adesão ao tratamento na primeira consulta e o tratamento PIs (em relação ao INSTIs) diminuem as chances de o paciente ter cargas virais iguais a zero (Tabela 5.5).

As componentes de variação dos efeitos aleatórios indicam que a correlação das duas componentes (dos zeros e dos valores positivos) é $-0,5432$ (Tabela 5.6). A variância para os pacientes i que apresentam Carga viral zero é $1,8225$ e para os pacientes i que apresentam Carga viral superior a zero é $1,1617$.

Observando a figura 5.3, pode ser constatado que se verificam os pressupostos do modelo: os resíduos seguem uma distribuição uniforme no intervalo (0,1) e não se verificam desvios da uniformidade na direção dos y.

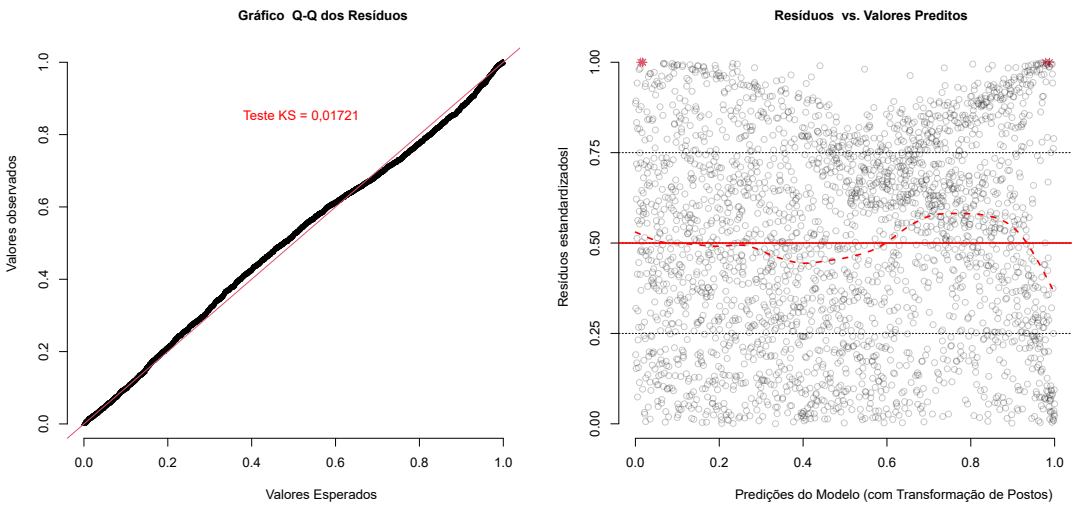


Figura 5.3: Avaliação dos pressupostos do modelo para a Carga viral.

Capítulo 6

Modelos de Riscos Proporcionais de Cox para a Neoplasia

Para determinar o risco de o paciente desenvolver neoplasia, foram construídos modelos de riscos proporcionais de Cox.

Este tipo de modelos não permite a inclusão de variáveis endógenas que variam com o tempo (Rizopoulos, 2012). Por este motivo, para as variáveis que variam com o tempo, foram apenas incluídos os valores registados ou ajustados ao início do estudo.

Quando se ajusta o modelo de riscos proporcionais de Cox é também importante ter em conta a censura dos dados. Os dados analisados apresentam censura à direita.

Os modelos apresentados incluem apenas os indivíduos que apresentaram observações completas, em que a amostra tinha a dimensão de 419 indivíduos e foram registados 18 eventos (i.e., 18 indivíduos desenvolveram neoplasia).

6.1 Seleção das Variáveis Explicativas

Para determinar quais as variáveis explicativas a serem incorporadas nos modelos de riscos proporcionais de Cox que incluirão os valores ajustados para as Células CD4 e Carga viral (usando os modelos anteriormente construídos no capítulo 5), foram ajustados modelos de Cox em que a Neoplasia foi considerada como a variável resposta e as restantes variáveis da Tabela 4.1 (exceto a Carga viral e as Células CD4) foram utilizadas como covariáveis. Para avaliar a existência de multicolinearidade, foi utilizado o fator de inflação de variância (VIF) para o modelo de Cox, disponível no pacote *rms* do R. Seguidamente, foi usado o

método de seleção *stepwise backward* baseado no critério de informação de Akaike (AIC).

O modelo obtido possuía uma variável, a Naturalidade, cujo teste para a significância do coeficiente da estimativa tinha um valor de prova demasiado alto (superior a 0,99). Tendo em conta que a literatura não suporta a inclusão desta variável no modelo, a Naturalidade foi retirada.

A razão de risco para cada covariável do modelo final e respetivos intervalos de confiança e valores de prova estão na Tabela 6.1.

Os valores de prova para os testes de significância (Verosimilhança, Wald, score) do modelo são inferiores a 0,01, o que indica que o modelo é significativo para um nível de significância de 1%.

Tabela 6.1: Razão de risco para cada covariável do modelo final e respetivos intervalos de confiança (IC) e valores de prova.

Covariáveis	Razão de Risco	95% IC	Valor de Prova
Linfócitos	0,4783	[0,204; 1,124]	0,0906
Plaquetas	0,9944	[0,988; 1,001]	0,1021
Infeções oportunistas	6,1100	[2,1737; 17,174]	0,0006

Segundo o modelo, o risco dos pacientes desenvolverem neoplasia é significativamente maior para aqueles que apresentam infeções oportunistas (Razão de Risco = 6,11), mantendo as restantes covariáveis constantes. Adicionalmente se os valores dos linfócitos aumentarem uma unidade, e mantendo as restantes covariáveis inalteradas, o risco do indivíduo desenvolver neoplasia diminui 52,17%. Finalmente, o risco de o indivíduo desenvolver neoplasia diminui 0,56% por cada unidade de aumento dos valores das plaquetas e mantendo as restante variáveis explicativas constantes.

A Figura 6.1 representa a probabilidade de o indivíduo desenvolver neoplasia. Pode ser constatado, segundo o modelo, que aproximadamente após os dez meses de follow-up, o risco de o paciente desenvolver neoplasia mantém-se constante.

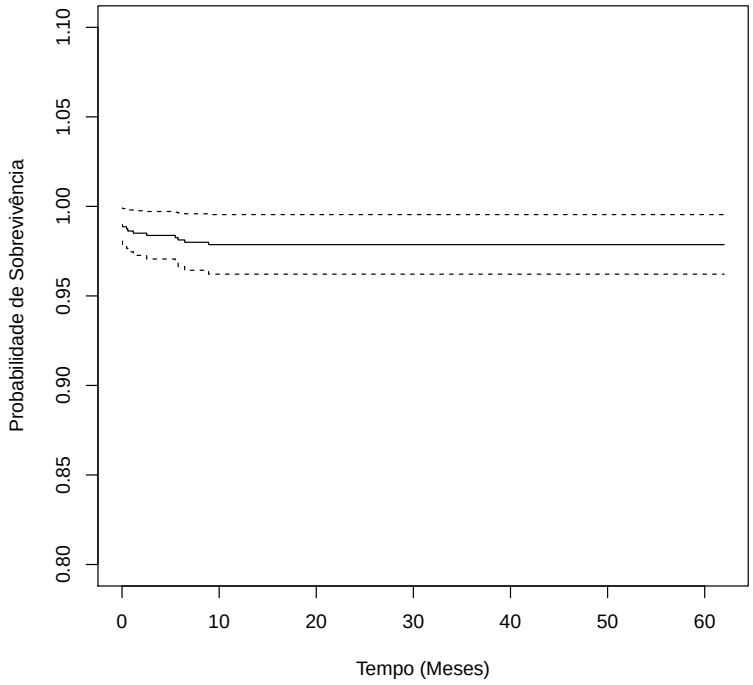


Figura 6.1: Curva de sobrevivência.

6.2 Diagnósticos do Modelo

Na Tabela 6.2, para um nível de confiança de 10%, pode ser constatado que o pressuposto da proporcionalidade dos riscos não é rejeitado (valor de prova $>$ 0,10).

Tabela 6.2: Teste para avaliação do pressuposto da proporcionalidade dos riscos.

Covariáveis	Estatística Qui-quadrado	Valor de Prova
Linfócitos	0,8517	0,36
Plaquetas	0,0071	0,93
Infeções oportunistas	0,9081	0,34
Global	1,2906	0,73

6.3 Modelo com as Células CD4

Após a construção do modelo anterior foi constatado que as covariáveis Plaquetas, Linfócitos e Infeções oportunistas afetam significativamente a ocorrência de neoplasia. Seguidamente foi construído um modelo com estas covariáveis conjuntamente com os valores de células CD4 ajustados usando o modelo construído no capítulo 5 como uma covariável adicional. Após a verificação da ausência de multicolinearidade, foi feita a seleção de variáveis usando o método *stepwise backward* baseado no AIC. Assim, foi obtido um modelo cujas covariáveis são as Células CD4 e as Infeções oportunistas.

A razão de risco para cada covariável do modelo obtido e respetivos intervalos de confiança e valores de prova estão na Tabela 6.3.

Os valores de prova para os testes de significância (Verosimilhança, Wald, score) do modelo são inferiores a 0,01, o que indica que o modelo é significativo para um nível de significância de 1%.

Tabela 6.3: Razão de risco para cada covariável do modelo e respetivos intervalos de confiança (IC) e valores de prova.

Covariáveis	Razão de Risco	95% IC	Valor de Prova
CD4	0,9911	[0,9862; 0,9960]	0,0004
Infeções oportunistas	2,7338	[0,9601; 7,7840]	0,0596

Segundo o modelo, o risco dos pacientes desenvolverem neoplasia é significativamente maior para aqueles que apresentam infeções oportunistas (Razão de Risco = 2,7338), mantendo as restantes covariáveis constantes. Para além disso, o risco do indivíduo desenvolver neoplasia diminui 0,89% por cada unidade de aumento dos valores das células CD4 e mantendo as restante variáveis explicativas constantes.

Na Tabela 6.2, para um nível de confiança de 10%, pode ser constatado que o pressuposto da proporcionalidade dos riscos não é rejeitado (valor de prova $> 0{,}10$).

Tabela 6.4: Teste para avaliação do pressuposto da proporcionalidade dos riscos.

Covariáveis	Estatística Qui-quadrado	Valor de Prova
Células CD4	0,219	0,7
Infeções oportunistas	1,088	0,30
Global	1,089	0,58

6.4 Modelo com a Carga Viral

Após a construção do primeiro modelo de riscos proporcionais de Cox foi constatado que as covariáveis Plaquetas, Linfócitos e Infeções oportunistas afetam significativamente a ocorrência de neoplasia. Foi construído um modelo com as covariáveis Plaquetas, Linfócitos, Infeções oportunistas e os valores de Carga viral ajustados usando o modelo construído no capítulo 5 como uma covariável adicional. Neste modelo, não foram consideradas interações entre as covariáveis. Seguidamente foi feita a seleção de variáveis usando o método *stepwise backward* baseado no AIC. Assim, foi constatado que a variável Carga viral foi eliminada do modelo, o que significa esta covariável não está associada com a presença de neoplasia.

Considerando o modelo com a interação entre as variáveis Carga viral e Infeções oportunistas, as razões de risco para cada cada covariável e respetivos intervalos de confiança e valores de prova estão na Tabela 6.5.

Tabela 6.5: Razão de risco para cada covariável do modelo e respetivos intervalos de confiança (IC) e valores de prova.

Covariáveis	Razão de Risco	95% IC	Valor de Prova
CV	1,0001	[1,0000; 1,0002]	0,1081
Infeções oportunistas	28,8700	[6,4345; 129,5325]	$1,13 \times 10^{-5}$
Plaquetas	0,9930	[0,9859; 1,0002]	0,0552
Linfócitos	0,3810	[0,1563;0,9286]	0,0338
CV*Infeções oportunistas	0,9980	[0,9961; 0,9999]	0,0429

Segundo o modelo, o risco de os pacientes desenvolverem neoplasia é significativamente

maior para aqueles que apresentam infeções oportunistas (Razão de Risco = 28,87), considerando as restantes covariáveis como zero. Existe uma interação entre a carga viral e as infeções oportunistas, ou seja, o efeito da carga viral na neoplasia muda consoante a presença ou ausência de neoplasia.

Para além disso, o risco de o indivíduo desenvolver neoplasia diminui 0,7% por cada unidade de aumento dos valores de plaquetas e mantendo as restante variáveis explicativas constantes. O risco de o indivíduo desenvolver neoplasia diminui 61,9% por cada unidade de aumento dos valores de linfócitos e mantendo as restante variáveis explicativas constantes.

Na Tabela 6.6, para um nível de significância de 5%, pode ser constatado que o pressuposto da proporcionalidade dos riscos não é rejeitado (valor de prova $> 0,05$).

Tabela 6.6: Teste para avaliação do pressuposto da proporcionalidade dos riscos.

Covariáveis	Estatística Qui-quadrado	Valor de Prova
CV	0,9331	0,334
Infeções oportunistas	0,7078	0,400
Plaquetas	0,0036	0,952
Linfócitos	0,9143	0,339
CV*Infeções oportunistas	3,2538	0,071
Global	5,5462	0,353

6.5 Limitações dos Modelos Apresentados

Os dados apresentam um baixo número de pacientes que desenvolveram neoplasia (18 em 419 pacientes). Para efetuar a estimação correta dos coeficientes de regressão do modelo de Cox, Peduzzi et al. (1995) recomendam que existam pelo menos 10 eventos de interesse por cada covariável incluída no modelo, o que não é verificado para os modelos construídos. Para além disso, existem 9 pacientes em que a neoplasia foi registada ao início do follow-up, i.e., o tempo em que ocorre o evento é zero meses. Esta situação pode indicar a presença de censura à esquerda, i.e., o evento foi registado antes do início do estudo.

Capítulo 7

Conclusão

O objetivo deste projeto foi avaliar a evolução da carga viral e contagem das células CD4 ao longo do tempo em resposta a diferentes regimes de terapia antirretroviral (ARV), numa *coort naïve* de pacientes infetados com HIV e investigar os fatores associados à incapacidade de obter a supressão da carga viral (valores de carga viral iguais a zero). Para além disso, foi determinado o risco neoplasia, aplicando o modelo de riscos proporcionais de Cox.

Após a aplicação do modelo de efeitos aleatórios para modelar a evolução das células CD4 ao longo do tempo foi constatado que os seguintes fatores estavam associados com a dinâmica deste tipo de células: Sexo, Hemoglobina, Leucócitos, Neutrófilos, Albumina, Tipo de tratamento, Idade na primeira consulta e o Tempo de follow-up. Os valores de células CD4 aumentam com o aumento dos valores de Hemoglobina, Leucócitos, Albumina e Tempo e para os pacientes do sexo feminino. Por outro lado, os valores de CD4 diminuem com os valores de Neutrófilos, Idade na primeira consulta e se o paciente seguir o tratamento NNRTIs ou PIs, sendo os valores de CD4 ainda mais baixos se seguir o tratamento PIs. Assim, de acordo com o modelo, o tratamento mais eficaz é INSTIs e o menos eficaz é PIs.

Para modelar a resposta da carga viral ao longo do tempo, e após a aplicação de vários modelos, foi escolhido o modelo hurdle binomial negativo com efeitos aleatórios. De acordo com este modelo, os fatores associados a valores de carga viral igual a zero são Sexo, Modo de transmissão, Neutrófilos, Albumina, Adesão, Tipo de tratamento e Tempo.

Um paciente tem a maior probabilidade de atingir valores de carga viral iguais a zero se for do sexo feminino em relação a um do sexo masculino. Também a chance de um paciente atingir valores de carga viral iguais a zero vezes é maior para um paciente cuja a

transmissão foi por via toxicodependente em relação à transmissão heterossexual. A não adesão ao tratamento na primeira consulta e o tratamento PIs (em relação ao INSTIs) diminuem as chances de o paciente ter cargas virais iguais a zero. A chance de o paciente atingir níveis zero de carga viral aumenta com os valores de Neutrófilos, Albumina e Tempo.

O aumento de Albumina e de Tempo de tratamento aumentam as chances de o paciente atingir valores de carga viral zero, e a diminuição de Albumina e de Tempo de tratamento aumentam as chances de o paciente ter valores de carga viral superior a zero. A diminuição da Hemoglobina aumenta a chance de o paciente ter valores de carga viral superiores a zero.

A não adesão ao tratamento na primeira consulta e os tratamentos NNRTIs e PIs (em relação ao INSTIs) aumentam as chances de o doente apresentar cargas virais superiores a zero.

A não adesão ao tratamento na primeira consulta e o tratamento PIs (em relação ao INSTIs) diminuem as chances de o paciente ter cargas virais iguais a zero. Assim, de acordo com o modelo construído para a Carga viral, o tratamento mais eficaz é INSTIs e o menos eficaz é PIs, o que é semelhante ao modelo de efeitos aleatórios construído para as células CD4.

Usando os valores ajustados para as células CD4 e carga viral (obtidos a partir dos modelos de efeitos aleatórios anteriormente construídos) como covariáveis nos modelos de Cox para calcular o risco de neoplasia, o risco de neoplasia está associado com as células CD4 e Infeções oportunistas (para o modelo com as células CD4) e associado com a carga viral, Infeções oportunistas, Plaquetas e Linfócitos, existindo um efeito de interação entre a carga viral e infeções oportunistas (para o modelo com a carga viral). No primeiro modelo de Cox mencionado, o risco de os pacientes desenvolverem neoplasia é maior para aqueles que apresentam infeções oportunistas e o risco de neoplasia diminui com o aumento das células CD4. No segundo modelo, o risco de os pacientes desenvolverem neoplasia é maior para aqueles que apresentam infeções oportunistas e o risco de neoplasia diminui com o aumento de plaquetas e linfócitos.

Ainda, em relação aos modelos de riscos proporcionais apresentados, os coeficientes de regressão podem não estar corretamente estimados, dado que não é registado pelo menos 10 eventos de interesse por cada covariável incluída no modelo. Para além disso, existem 9 pacientes em que a neoplasia foi registada ao início do follow-up, o pode indicar a presença

de censura à esquerda (i.e., o evento foi registado antes do início do estudo). Esta possível presença de censura à esquerda não foi tida em conta nos modelos de riscos proporcionais de Cox construídos.

Bibliografia

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, 15–27.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Belin, T. R., Hu, M.-Y., Young, A. S., & Grusky, O. (2000). Using multiple imputation to incorporate cases with missing items in a mental health services study. *Health Services and Outcomes Research Methodology*, *1*(1), 7–22.

Bolker, B. (2020). *Glmm faq.* `http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html`, Acedido em 19 de novembro de 2020.

Borges, A. I. C. (2015). *Joint modelling of longitudinal and survival data on breast cancer.* `http://hdl.handle.net/1822/40426`.

Cabral, M. S., & Gonçalves, M. H. (2011). Análise de dados longitudinais. *Sociedade Portuguesa de Estatística*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data* (Vol. 21). CRC Press.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(2), 248–265.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Diggle, P. J. (1990). *Time series; a biostatistical introduction* (Tech. Rep.).

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press.

Direção Geral de Saúde. (2019). *Infeção vih e sida em portugal - 2019* (Tech. Rep.).

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. CRC press.

Freund, R. J., Wilson, W. J., & Sa, P. (2006). Multiple linear regression. *Regression Analysis: Statistical Modeling of a Response Variable, Oxford, UK: Elsevier,*, 73–115.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, *61*(2), 383–385.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*(5), 531–547.

Hughes, J. P. (1999). Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, *55*(2), 625–629.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Klein, J. P., & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Klimas, N., Koneru, A. O., & Fletcher, M. A. (2008). Overview of hiv. *Psychosomatic medicine*, *70*(5), 523–530.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in medicine*, *7*(1-2), 305–315.

Little, R. J., & Rubin, D. B. (1987). Statistical analysis with missing data. *Hoboken, NJ: Wiley*, *65*.

Mcneil, A. J., & Gore, S. M. (1996). Statistical analysis of zidovudine (azt) effect on cd4 cell counts in hiv disease. *Statistics in medicine*, *15*(1), 75–92.

Minini, P., & Chavance, M. (2004). Sensitivity analysis of longitudinal normal data with drop-outs. *Statistics in medicine*, *23*(7), 1039–1054.

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies* (Vol. 61). John Wiley & Sons.

Molenberghs, G., & Verbeke, G. (2000). *Linear mixed models for longitudinal data*.

Springer.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, *33*(3), 341–365.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945–966.

Organização Mundial de Saúde. (2007). *Who case definitions of hiv for surveillance and revised clinical staging and immunological classification of hiv-related disease in adults and children*. World Health Organization.

Organização Mundial de Saúde. (2017). What's new in treatment monitoring: viral load and cd4 testing. *Update*.

Organização Mundial de Saúde. (2020). *Hiv/aids.* `https://www.who.int/news-room/fact-sheets/detail/hiv-aids`, Acedido em 10 de dezembro de 2020.

Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis ii. accuracy and precision of regression estimates. *Journal of clinical epidemiology*, *48*(12), 1503–1510.

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in r*. CRC press.

Rizopoulos, D. (2020). *Goodness of fit for mixmod objects.* `https://drizopoulos.github.io/GLMMadaptive/articles/Goodness_of_Fit.html`, Acedido em 14 de dezembro de 2020.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sousa, I. (2011). A review on joint modelling of longitudinal measurements and time-to-event. *Revstat Stat J*, *9*, 57–81.

Temesgen, A., & Kebede, T. (2016). Joint modeling of longitudinal cd4 count and weight measurements of hiv/tuberculosis co-infected patients at jimma university specialized hospital. *Annals of Data Science*, *3*(3), 321–338.

Therneau, T. M., & Grambsch, P. M. (2000). The cox model. In *Modeling survival*

data: extending the cox model (pp. 39–77). Springer.

Volberding, P., Bartlett, J., Del Rio, C., Flynn, P., Gant, L., Grant, I., & Williams, A. (2010). Hiv and disability: Updating the social security listings. *Washington, DC: Institute of Medicine*.

Yang, X., & Shoptaw, S. (2005). Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. *Drug and Alcohol Dependence*, *77*(3), 213–225.

Zuur, A., Hilbe, J. M., & Ieno, E. N. (2013). *A beginner's guide to glm and glmm with r: A frequentist and bayesian perspective for ecologists*. Highland Statistics Limited.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with r*. Springer Science & Business Media.