Universidade de Brasília
Departamento de Estatística

Modelo semiparamétrico com fração de cura: uma aplicação ao estudo de prognóstico de pacientes com câncer de cabeça e pescoço

Gabriel Couto Schelbauer

Brasília
2024

Gabriel Couto Schelbauer

Modelo semiparamétrico com fração de cura: uma aplicação ao estudo de prognóstico de pacientes com câncer de cabeça e pescoço

Orientador(a): Prof. Frederico Machado Almeida

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Agradecimentos

Agradeço primeiramente à minha mãe, Elisandra, por ter me apoiado e me incentivado ao longo da minha jornada acadêmica. Agradeço também à minha segunda mãe, a Dona Anita, por ter cuidado de mim e me acolhido como seu filho. Eu não chegaria em lugar algum sem a minha família, e sou grato por ter recebido tanto amor e suporte emocional.

Agradeço ao meu orientador, o professor Frederico, por ter aceitado me guiar durante essa jornada tão difícil, sempre me transmitindo valiosos ensinamentos e me ajudando a moldar o resultado final deste trabalho.

Agradeço aos meus queridos amigos, principalmente o Lucas, a Nataly, a Mariana, a Bebel e a Gabriela. Agradeço ao meu maior parceiro, o Natan, e à minha querida irmãzinha, a Bianca. Vocês foram todos responsáveis por me ajudar a manter a sanidade mental durantes esses anos de graduação.

Aos professores do Departamento de Estatística estendo os meus agradecimentos, em especial ao Antônio Eduardo e à Maria Teresa - vocês são todos peças fundamentais no meu desenvolvimento como profissional e ser humano.

Para concluir, muito obrigado a todos que sempre estiveram ao meu redor torcendo por mim e me incentivando a sempre me desafiar. A maior lição que eu aprendi ao longo da minha jornada como pessoa nesse planeta foi essa - ninguém conquista nada sozinho.

Resumo

Apesar de ser um dos tipos mais comuns de câncer no mundo, o carcinoma de células escamosas de cabeça e pescoço carece de um protocolo de diagnóstico eficaz. Nesse contexto, a análise de sobrevivência pode ser uma aliada valiosa na identificação de fatores relacionados ao prognóstico da doença. O presente trabalho aplica o modelo semiparamétrico com fração de cura, para identificar variáveis de interesse que possam ter relação com a distribuição dos tempos de falha (sobrevivência), e/ou a cura dos indivíduos. A base de dados utilizada nesse estudo possui características clínicas e sócio-econômicas de 1.843 pacientes, e foi obtida através do repositório público da Universidade de Michigan. Para o ajuste do modelo foi utilizado o pacote *penPHcure*, do *R*, para a estimação dos parâmetros do modelo por meio de regressão logística e uso do algoritmo de máxima expectativa, o que possibilitou a identificação de 12 covariáveis relevantes no prognóstico do câncer de cabeça e pescoço. O modelo ajustado tem o potencial de ser uma ferramenta relevante no diagnóstico e tratamento de pacientes diagnosticados com a doença, além de contribuir para discussões acadêmicas sobre o assunto.

Palavras-chave: câncer de cabeça e pescoço; modelo de sobrevivência com fração de cura; modelo de Cox; algoritmo de máxima expectativa; regressão logística; taxa de cura; censura.

Abstract

Despite being one of the most common types of cancer worldwide, head and neck squamous cell carcinoma still lacks an effective diagnostic protocol. In that sense, survival analysis can provide valuable tools in identifying factors associated with the prognosis of the disease. This study applies the semiparametric mixture cure model of Sy e Taylor (2000) in order to identify risk factors and their impact in both part of the model, survival and cured. The dataset used in the study contains clinical, demographic, and self-reported behavior characteristics for a cohort of 1,843 patients and was obtained from the public repository of the University of Michigan. The *penPHcure R package* was employed to estimate model parameters through logistic regression and the expectation-maximization algorithm, which measured the effects of some covariates on the probability of being susceptible and on the time until the occurrence of death. The fitted model has the potential to be a relevant tool in the diagnosis and treatment of patients diagnosed with the disease, as well as contributing to academic discussions on the subject.

Keywords: head and neck cancer; mixture cure model; Cox's model; EM algorithm; logistic regression; cure rate; censoring.

Sumário

Lista de Tabelas

Lista de Figuras

1 Introdução

O câncer de cabeça e pescoço é o sétimo tipo de câncer mais comum no mundo, sendo responsável por aproximadamente 890.000 dos novos diagnósticos de câncer e por 450.000 das mortes ocasionadas pela doença no mundo todo (BARSOUK et al., 2023). O carcinoma de células escamosas de cabeça pescoço (CCECP) apresenta taxas de sobrevivência particularmente baixas, o que pode ser atribuído em parte ao diagnóstico tardio e frequente associação de algumas comorbidades (VIRANI et al., 2015), portanto é de extrema importância identificar os potenciais fatores de risco associados a sua progressão.

Mais recentemente, alguns estudos mostraram que a presença do Papilomavírus Humano (HPV, da sigla em inglês) se manifesta como um agente etiológico independente em tipos específicos do CCECP, e que nesses casos o prognóstico é melhor do que aqueles não associados ao vírus (VIRANI et al., 2015; SABATINI; CHIOCCA, 2020). Os autores acrescentam ainda que, mesmo nesses casos, há heterogeneidade nos tempos de sobrevivência, o que pode ser atribuído a marcadores genéticos de metilação específicos associados à recorrência e sobrevivência a tumores.

Dentre os tipos do vírus HPV associados a cânceres anogenitais que pertencem ao grupo de alto risco, se destaca o HPV16, estando presente em aproximadamente 50% dos cânceres cervicais no mundo todo, e sendo responsável por pelo menos 90% dos casos de CCECP associados a vírus. Nos últimos anos, na Europa e na América do Norte, casos de CCECP causados pelo HPV têm aumentado drasticamente. Entretanto, nenhum protocolo de diagnóstico para detecção precoce desses carcinomas foi implementado (GALATI et al., 2022).

Estudos da área de oncologia apontam que alterações epigenéticas impactam diretamente os mecanismos complexos da carcinogênese de cabeça e pescoço. Nas últimas décadas, marcadores epigenéticos têm sido reconhecidos como agentes etiológicos relevantes no desenvolvimento do CCECP, destacando-se os marcadores CD1A e NDN, associados a taxas de sobrevivência maiores em pacientes com o este tipo de carcinoma. Mais especificamente, temos que a hipermetilação de NDN está associada a menores probabilidades gerais de recidiva do CCECP, ao passo que a hipermetilação do marcador CCNA1 no geral representa um prognóstico melhor da doença apenas em pacientes HPV-positivos. Portanto, é evidente que a identificação de marcadores epigenéticos relevantes e o mapeamento do comportamento biológico do tumor têm grande potencial de auxiliar no desenvolvimento de ferramentas de prognóstico mais acuradas e novas alternativas terapêuticas (VIRANI et al., 2015). Adicionalmente, por se tratar de um vírus sexualmente transmissível, é intuitivo pensar que hábitos sexuais como início da atividade sexual, quantidade de parceiros sexuais e incidência prévia de verrugas genitais estão associados

a um risco maior de contágio do HPV e, consequentemente, do desenvolvimento de casos de CCECP. Desse modo, a identificação e controle desses cofatores se manifesta como um instrumento essencial no diagnóstico precoce de quadros de câncer de cabeça e pescoço.

Para entender melhor esses fatores e suas influências, a análise de sobrevivência se mostra essencial. Essa área da estatística se concentra no estudo do tempo até a ocorrência de um evento de interesse, frequentemente denominado tempo de falha. Amplamente utilizada em estudos médicos, a análise de sobrevivência destaca-se pela presença de dados censurados, que ocorrem quando a informação completa sobre o tempo de falha não está disponível para todos os indivíduos (COLOSIMO; GIOLO, 2006). Nesse contexto, a aplicação de modelos de sobrevivência se apresenta como uma ferramenta crucial na compreensão do desenvolvimento do CCECP e na identificação de agentes etiológicos.

No contexto da análise de sobrevivência, quando há evidências empíricas de que o evento de interesse não será observado em parte dos indivíduos, pode-se utilizar o modelo de sobrevivência com fração de cura, que considera uma distribuição binária para a parte da cura, e uma distribuição paramétrica para a modelagem do tempo de falha (SY; TAYLOR, 2000). Entretanto, apesar de terem ajuste e interpretação mais fáceis, modelos paramétricos são sensíveis a especificações incorretas (FANG H.-B.; SUN, 2005). Desta forma, modelos semiparamétricos podem ser utilizados como uma alternativa flexível a estes modelos.

A falta de ferramentas capazes de diagnosticar precocemente o CCECP é um problema de saúde pública com a capacidade de impactar um grande número de indivíduos. Deste modo, o presente trabalho tem por objetivo identificar potenciais marcadores clínicos, sociais e demográficos que possam influenciar na progressão do CCECP por meio do ajuste de um modelo semiparamétrico com fração de cura.

2 Objetivos

2.1 Objetivo Geral

O presente trabalho tem como objetivo geral aplicar os modelos de sobrevivência com fração de cura para identificar potenciais fatores de risco que podem influenciar na progressão do CCECP.

2.2 Objetivos Específicos

- Apresentar a formulação do modelo de fração de cura;
- Propor um estudo de simulação de Monte Carlo para avaliar o desempenho do modelo;
- Fazer uma análise descritiva dos dados;
- Identificar, com base nos dados reais, quais fatores de risco podem influenciar na evolução do CCECP.

3 Revisão de Literatura

3.1 Conceitos Básicos em Análise de Sobrevivência

A análise de sobrevivência é uma área da estatística que se concentra no estudo do tempo decorrido até que um evento de interesse aconteça, denominado tempo de falha. Amplamente utilizada em estudos médicos, onde o tempo de falha é, geralmente, o tempo até o óbito, por exemplo, a análise de sobrevivência pode ter aplicações em diversas outras áreas de estudo, como engenharia e economia, por exemplo. Segundo Colosimo e Giolo (2006), a principal característica de dados de sobrevivência é a presença de censuras, que são observações parciais do tempo de falha de alguns dos indivíduos da amostra. Em situações onde não se observa censura, ou seja, o tempo de falha para todas as observações é completamente observado, técnicas usuais de análise estatística (como a análise de variância, por exemplo) podem ser utilizadas. Entretanto, a probabilidade de haver censura geralmente é alta, o que faz com que a análise de sobrevivência seja a ferramenta mais adequada para o estudo de dados deste tipo.

3.1.1 Tempo de Falha

Em análise de sobrevivência, temos como foco principal um grupo (ou grupos) de indivíduos para os quais ocorre um evento específico de interesse, frequentemente denominado falha, e o tempo decorrido até esse evento (COX; OAKES, 1984). O tempo de falha é constituído de três elementos principais: o tempo inicial, a escala de medida e o evento de interesse (falha). O primeiro desses, o tempo inicial, deve ser definido de forma que todos os indivíduos sejam comparáveis no início do estudo. A escala de medida frequentemente utilizada é o tempo real, mas outras escalas podem ser utilizadas (como o número de ciclos, por exemplo). Por último, o evento de interesse (ou falha) deve ser definido de forma clara e precisa antes do início do estudo (COLOSIMO; GIOLO, 2006).

3.1.2 Censura

Frequentemente, estudos de sobrevivência se encerram antes que todos os participantes desenvolvam o evento de interesse, o que resulta em observações parciais e incompletas da variável resposta. Os mecanimos de censura se manifestam de forma diferente a depender de como o estudo foi estruturado, podendo ser classificados em três tipos principais:

- Censura à direita: ocorre quando o evento de interesse está à direita do tempo de

observação, ou seja, a falha aconteceu após o término do estudo. Pode ser observada em três tipos diferentes:

1. Censura do tipo I: ocorre quando, dado um período pré-estabelecido de tempo, existem indivíduos que não apresentaram o evento de interesse após o término do estudo.
2. Censura do tipo II: presente em estudos onde o término só ocorre após um número pré-estabelecido de indivíduos apresentar o evento de interesse.
3. Censura do tipo aleatório: bastante frequente em estudos médicos, ocorre quando o indivíduo é removido do estudo sem que tenha ocorrido a falha. Ou ainda, se a falha ocorre por um motivo diferente daquele a ser estudado.

- Censura à esquerda: apresenta-se quando a falha ocorre antes do início do estudo ou da coleta dos dados.
- Censura intervalar: ocorre em estudos em que os indivíduos são acompanhados em visitas periódicas e, consequentemente, sabe-se apenas que a falha ocorreu em um intervalo de tempo.

Com exceção dos casos em que o tipo de censura observada foi a censura à esquerda, mecanismos de censura são representados por duas variáveis aleatórias: T, uma variável aleatória correspondente ao tempo de falha, e C, uma variável aleatória independente de T que representa o tempo de censura associado ao indivíduo. Desta forma, temos que o tempo de falha t pode ser expresso por:

$$t = min(T, C)$$

e a variável indicadora da presença de censura δ por:

$$\delta = \begin{cases} 1, & \text{se T} \leq \text{C (a falha ocorre antes da censura)} \\ 0, & \text{se T} > \text{C (a falha ocorre após a censura).} \end{cases} \tag{3.1.1}$$

3.1.3 Representação dos Dados de Sobrevivência

Para um dado indivíduo i ($i = 1, 2, ..., n$), os dados de sobrevivência são representados pelo par (t_i, δ_i), onde t_i representa o tempo de falha ou censura e δ_i corresponde à variável indicadora de falha ou censura, onde:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo censurado,} \end{cases} \tag{3.1.2}$$

quando o tipo de censura observada foi a censura à direita.

Adicionalmente, quando temos covariáveis de interesse para todos os n indivíduos da amostra, representamos nossos dados por $(t_i, \delta_i, \mathbf{x}_i)$, onde $\mathbf{x}_i$ é o vetor aleatório correspondente às covariáveis.

3.1.4 Função de Sobrevivência

Definida como a probabilidade de uma observação i não falhar até um dado tempo t, a função de sobrevivência é uma das funções probabilísticas mais importantes em análise de sobrevivência. Matematicamente, podemos expressá-la por:

$$S(t) = P(T > t). \tag{3.1.3}$$

De forma geral, a função de sobrevivência possui três propriedades principais:

1. S(0) = 1, ou seja, a probabilidade de sobrevivência no tempo inicial é igual a 1 e nenhum indivíduo apresentou o evento de interesse,
2. S(t) é não-crescente no tempo t,
3. $\lim_{t\to\infty} S(t) = 0$.

Podemos definir a função de distribuição acumulada como a probabilidade de uma observação não sobreviver ao tempo t, ou seja, $F(t) = 1 - S(t)$

3.1.5 Função de Risco

Em análise de sobrevivência, definimos a função de risco (ou função de taxa de falha) como sendo a probabilidade do evento de interesse ser observado em um intervalo de tempo $[t_1, t_2)$, dado que o indivíduo sobreviveu até t_1, em unidades de tempo. Podemos expressá-la em termos da função de sobrevivência como:

$$S(t_1) - S(t_2). \tag{3.1.4}$$

Definimos a taxa de falha como a probabilidade de que o evento de interesse ocorra no intervalo $[t_1, t_2)$, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Desta forma, a taxa de falha pode ser escrita como:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{3.1.5}$$

Podemos redefinir o intervalo como $[t,\ t\ +\ \Delta t)$, o que resulta na seguinte expressão:

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{3.1.6}$$

Assim, se tivermos um Δt pequeno, $\lambda(t)$ corresponde à taxa de falha instantânea no tempo t condicional, dada a sobrevivência até este tempo. Portanto, a função de taxa de falha de T é definida como:

$$\lambda(t) = \lim_{\Delta t \to\ 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.1.7}$$

Ainda, temos que:

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3.1.8}$$

que denota a relação existente entre as três funções frequentemente utilizadas em análise de sobrevivência, onde $f(t)$ é a função de densidade de probabilidade obtida por meio da derivação da função $F(t)$.

3.1.6 Função de Risco Acumulado

A função de risco acumulado é outra função relevante em análise de sobrevivência, podendo ser expressa por:

$$\Lambda(t) = \int_0^t \lambda(u)\, du. \tag{3.1.9}$$

Apesar de não termos uma interpretação direta para a função $\Lambda(t)$, ela pode ser útil na análise da função de risco, $\lambda(t)$. Com base na expressão (3.1.9), a função de sobrevivência pode ser reescrita como $S(t) = exp(-\Lambda(t))$.

3.2 Modelos de Sobrevivência com Fração de Cura

No contexto da análise de sobrevivência, há situações em que uma parcela dos indivíduos em observação não desenvolverão o evento de interesse, mesmo após um tempo de acompanhamento longo. Nesses casos, a população é uma mistura de dois grupos: um contendo indivíduos que apresentam o evento de interesse (denominados "não-curados"), e o outro abrangendo aqueles que não apresentaram o evento em estudo (denominados

"curados") (BOAG, 1949). Desta forma, consideramos o modelo com fração de cura como sendo uma combinação de dois modelos: o "modelo da incidência", que se refere à probabilidade de cura, e o "modelo da latência", correspondente à modelagem do tempo de falha (CORBIÈRE; JOLY, 2007).

Segundo Klein et al. (2014), se tomarmos T como sendo o tempo até o evento de interesse, observamos que, em um modelo com fração de cura, o limite de P(T > t) é diferente de zero à medida que t tende ao infinito. Esta característica pode ser observada obtendo um gráfico da estimativa de Kaplan-Meier, onde uma assíntota não-nula sugere que o uso de modelos com fração de cura pode ser mais apropriada.

Dentre os modelos utilizados em análise de sobrevivência, o modelo de riscos proporcionais e o modelo de tempo de falha acelerada se destacam como os mais difundidos na área. Se o modelo de riscos proporcionais é utilizado para modelar S(t) (a *latência*), o modelo de mistura é denominado modelo de riscos proporcionais com fração de cura. Por outro lado, se um modelo de tempo de falha acelerada é utilizado, o modelo é então chamado de modelo de tempo de falha acelerado com fração de cura (CAI et al., 2012).

3.2.1 Formulação do Modelo

Utilizamos a variável indicadora de cura Y, onde Y = 1 quando o indivíduo não é curado e Y = 0, caso contrário. Consequentemente, a probabilidade de um indivíduo não ser curado é denotada por P(Y = 1) = π, e assumimos que a probabilidade para pacientes curados é 1 - π. Definimos S_u(t) e S_c(t) como as funções de sobrevivência para os indivíduos não-curados e curados, respectivamente, onde S_u(t) = $P(T > t|Y = 1)$ e S_c(t) = $P(T > t|Y = 0)$. É intuitivo pensarmos que S_c(t) $\equiv$ 1, e portanto esta é uma função de sobrevivência degenerada. Portanto, o modelo resultante pode ser escrito como S_p(t) = 1 - π + πS_u(t) (KLEIN et al., 2014).

Adicionalmente, podemos tomar X e Z como sendo os vetores de covariáveis de interesse que possam ter impacto na taxa de cura ou na probabilidade de sobrevivência de pacientes não curados, de modo que o modelo resultante pode ser escrito na forma S_p(t|X, Z) = 1 - π(Z) + π(Z)S_u(t|X), onde π(Z) é a proporção de não curados, ou *incidência*, e S_u(t|X) é a probabilidade de sobrevivência de pacientes não curados, ou *latência* (BERKSON; GAGE, 1952).

Se tivermos um vetor z de covariáveis de interesse, seu efeito no tempo de falha de não-curados (representado aqui por π) é usualmente modelado utilizando uma função de ligação logit, log-log ou probit. A mais utilizada das três é a função de ligação logit:

$$logit[\pi(z_i)] = z_i'\gamma, \tag{3.2.1}$$

onde $logit[\pi(z_i)] = log(\pi/(1-\pi))$ e γ é o vetor de coeficientes das covariáveis em z, incluindo um intercepto (KLEIN et al., 2014). No presente trabalho a modelagem de π será feita desta forma.

Para a modelagem da parte dos incidentes, ou seja, do efeito de um vetor de covariáveis x em S_u(t), diversas abordagens podem ser adotadas. Dentre elas, se destacam o modelo de riscos proporcionais com fração de cura (ou PHMC, da sigla em inglês) e o modelo de tempo de falha acelerada com fração de cura (AFTMC, também do termo em inglês). Amplamete utilizado, o PHMC assemelha-se bastante ao modelo de Cox por também se basear no pressuposto de riscos proporcionais, e portanto é o modelo de mais fácil uso e interpretação (KLEIN et al., 2014). Nele, a função de sobrevivência é representada da seguinte forma:

$$S_u(t) = S_u(t|x) = S_{u0}(t)^{exp(x'\beta)}, \tag{3.2.2}$$

onde $\boldsymbol{\beta}$ é um vetor dos coeficientes das covariáveis em x e S_{u0} é a função de sobrevivência basal quando $x = 0$.

3.2.2 Estimação

Os dados de um modelo com fração de cura podem ser representados pelo vetor $(t_i,\ \delta_i,\ \mathbf{x}_i,\ \mathbf{z}_i)$, i = 1,...,n, em que δ_i é o indicador de censura com $\delta_i = 0$, caso o tempo de falha t_i seja censurado, e $\delta_i = 1$ caso contrário, ao passo que $\mathbf{x}_i$ e $\mathbf{z}_i$ são os vetores de covariáveis de interesse vistos acima (CORBIÈRE; JOLY, 2007).

Conforme demonstrado por Corbière e Joly (2007), a contribuição de um indivíduo i para a verossimilhança é $\pi_i(z_i)f(t_i|\text{Y=1}, x_i)$, para $\delta_i = 1$ e (1-$\pi_i(z_i)$) + $\pi_i(z_i)S(t_i|\text{Y=1}, x_i)$ quando $\delta_i = 0$, onde $f(.) = S(.)\lambda(.)$ é a função de densidade condicional de T. Sendo assim, a função de verossimilhança completa é dada por:

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{\pi_i(z_i)f(t_i|Y=1, x_i)\}^{\delta_i} \times \{(1-\pi_i(z_i)) + \pi_i(z_i)S(t_i|Y=1, x_i)\}^{1-\delta_i}. \tag{3.2.3}$$

A função de verossimilhança em (3.2.3) se reduz à função de verossimilhança de um modelo de sobrevivência padrão quando não há a presunção de uma fração de cura, ou seja, quando $\pi(z_i) = 1$ para todo z_i.

O estimador de máxima verossimilhança (EMV) $\{\hat{\boldsymbol{\gamma}},\ \hat{\boldsymbol{\beta}}\}$ é obtido por meio da maximização da função de log-verossimilhança $\ell(\boldsymbol{\gamma},\ \boldsymbol{\beta})$. Para isso, é necessário especificar a distribuição do tempo de falha para os indivíduos não curados, uma vez que a estimativa

de $\boldsymbol{\gamma}$ depende dessa distribuição. Em modelos de mistura com fração de cura paramétricos, essa distribuição é obtida a partir das funções $f(.|\ Y = 1)$ e S(.|\ $Y = 1$), que representam, respectivamente, a densidade e a função de sobrevivência condicionadas à população não curada. Essas funções, por sua vez, são definidas por parâmetros desconhecidos presentes na equação (3.2.3) (CORBIÈRE; JOLY, 2007). Além disso, a função $S_0(t|Y = 1)$ não pode ser eliminada do modelo sem que se perca informação a respeito de β. Por fim, as estimativas de máxima verossimilhança são obtidas utilizando métodos de otimização numérica, como o método de Newton-Raphson, que busca iterativamente os valores dos parâmetros que maximizam a função de verossimilhança, porém essa abordagem depende de como a função S_{u0} é parametrizada (KLEIN et al., 2014).

Outra forma de obter as estimativas da função de máxima verossimilhança é utilizando o algoritmo de maximização de expectativa (*EM*, da sigla em inglês), um método iterativo de estimação de parâmetros proposto por Dempster et al. (1977) que consiste em duas fases: a etapa das expectativas (E) e a da maximização (M).

Podemos reescrever a função de verossimilhança completa (3.2.3) como a soma de dois componentes: ℓ_I, que depende apenas de $\boldsymbol{\beta}$, e ℓ_S, que depende somente de $\boldsymbol{\gamma}$ e Λ_0:

$$\ell_I(\beta; \mathbf{y}) = log \prod_{i=1}^{n} \pi(z_i)^{y_i}(1 - \pi(z_i))^{1-y_i}, \tag{3.2.4}$$

$$\ell_S(\gamma, \Lambda_0; \mathbf{y}) = log \prod_{i=1}^{n} \lambda(t|Y = 1, x_i)^{\delta_i y_i} S(t|Y = 1, x_i)^{y_i}, \tag{3.2.5}$$

onde y é o vetor de valores de y_i e Λ_0 é o log de S_{u0} (CORBIÈRE; JOLY, 2007). Considerando a variável indicadora de que um individuo é suscetível ao evento de interesse (Y = 1) ou não (Y = 0), e o vetor correspondente $y = (y_1, y_2, ..., y_n)^T$, a etapa E na r-ésima iteração computa a expectativa condicional da função de log-verossimilhança completa $\ell(\boldsymbol{\gamma}, \boldsymbol{\beta})$ com respeito aos $y_i^{'}$s, dado os dados observados e os valores inicias de $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ e $S_0^{(0)}$(t|Y = 1) (CAI et al., 2012). Uma vez que (3.2.4) e (3.2.5) são funções lineares de y_i, o cálculo da expectativa condicional é suficiente para completar a etapa das expectativas (E). A expectativa condicional pode ser expressa da seguinte forma:

$$\begin{aligned} y_i^{(r)} &= E\{y_i|\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, S_0^{(r)}(t_i|Y_i = 1)\} \\ &= \delta_i + (1 - \delta_i)\frac{\pi^{(r)}(z_i)S^{(r)}(t_i|Y_i = 1, x_i)}{1 - \pi^{(r)}(z_i) + \pi^{(r)}(z_i)S^{(r)}(t_i|Y_i = 1, x_i)}, \end{aligned} \tag{3.2.6}$$

que é r-ésimo estimador da probabilidade do i-ésimo indivíduo ser suscetível ao evento de

interesse.

A etapa de maximização (M) na $(r+1)$-ésima iteração toma então as expectativas de ℓ_I e ℓ_S e as maximiza com respeito aos parâmetros desconhecidos $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, obtendo assim $\boldsymbol{\beta}^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$, $S_0^{(r+1)}(t_i|Y_i = 1)$. A estimativa da função de sobrevivência basal $S_{u0}(t)$ pode ser obtida na etapa de maximização (M), no entanto, é importante destacar que esse método produz apenas uma estimativa não paramétrica e não suavizada da função de sobrevivência basal. (KLEIN et al., 2014).

4 Metodologia

A modelagem das partes da latência e dos incidentes será feita separadamente com o uso do pacote penPHcure, do R, desenvolvido e proposto por Beretta e Heuchenne (2021). Para a primeira parte, utilizamos regressão logística por meio da função de ligação logit para a estimação dos parâmetros. Em seguida, o modelo utilizado na parte dos incidentes foi o modelo de riscos proporcionais com fração de cura (PHMC), cujos parâmetros foram estimados com o uso do algoritmo de máxima expectativa (EM). Finalmente, utilizamos uma técnica de seleção de variáveis de penalização do desvio absoluto suavemente recortado (*SCAD*, da sigla em inglês), que consiste na maximização de uma versão penalizada da função de log-verossimilhança com base nos dados completos:

$$\ell_C^P(\theta;\lambda_1,\lambda_2) = \underbrace{\ell_1(\gamma) - n\sum_{j=2}^{\mathbf{x}+1} p_{\lambda_1}(|\gamma_j|)}_{\ell_1^P(\gamma;\lambda_1)} + \underbrace{\ell_2(\beta,h_0) - n\sum_{l=1}^{\mathbf{z}} p_{\lambda_2}(|\beta_l|)}_{\ell_2^P(\beta,h_0;\lambda_2)}. \tag{4.0.1}$$

onde ℓ representa a função de log-verossimilhança, $p_\lambda(.)$ a função de penalização *SCAD*, h_0 é a função de risco condicional basal, γ é o vetor de coeficientes das covariáveis em $\mathbf{x}$ e β é o vetor dos coeficientes das covariáveis em $\mathbf{z}$ (BERETTA; HEUCHENNE, 2019). A função *SCAD* completa, proposta por Fan e Li (2002), é dada por:

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{se } |\beta_j| \leq \lambda \\ \frac{(a^2-1)\lambda^2-(|\beta_j|-a\lambda)^2}{2(a-1)}, & \text{se } \lambda < |\beta_j| \leq \mathrm{a}\lambda \\ \frac{(a+1)\lambda^2}{2}, & \text{se } |\beta_j| > \mathrm{a}\lambda, \end{cases} \tag{4.0.2}$$

para algum a > 2 e $\lambda > 0$, onde (a, λ) são os parâmetros de ajuste.

O pacote penPHcure, do R, opta por seguir a sugestão de Fan e Li (2002) e atribuir um valor fixo para o parâmetro de ajuste a, a saber $a_1 = a_2 = 3{,}7$. Em seguida, seleciona valores para o conjunto (λ_1,λ_2) que minimizem os critérios de informação de Akaike (AIC) e bayesiano (BIC):

$$AIC(\lambda_1,\lambda_2) = -2\ell(\hat{\theta}_{\lambda_1,\lambda_2}) + 2v, \tag{4.0.3}$$

$$BIC(\lambda_1,\lambda_2) = -2\ell(\hat{\theta}_{\lambda_1,\lambda_2}) + ln(n)v, \tag{4.0.4}$$

onde $\ell(\hat{\theta}_{\lambda_1,\lambda_2})$ é a função de log-verossimilhança com base nos dados observados na estima-

tiva de máxima verossimilhança penalizada $\hat{\theta}_{\lambda_1,\lambda_2}$, e v é o número de coeficientes não-nulos, ou seja, coeficientes com valores absolutos maiores que um limite pré-determinado (10^{-6}, por padrão) (BERETTA; HEUCHENNE, 2021).

5 Resultados

5.1 Conjunto de dados

O banco de dados utilizado neste trabalho foi obtido através do repositório público da Universidade de Michigan. Os dados foram coletados de 2003 à 2014 e contêm características clínicas e sócio-demográficas de 1843 pacientes com CCECP.

Coletado pelo Programa de Excelência em Pesquisa de Cabeça e Pescoço (SPORE) da Universidade de Michigan, os indivíduos da amostra representam 28% dos casos de câncer de cabeça e pescoço do estado do Michigan, nos Estados Unidos. Os pacientes foram recrutados em clínicas especializadas e eram incluídos no estudo caso tivessem sido diagnosticados com carcinoma de células escamosas de cabeça e pescoço previamente não tratado. Após o diagnóstico, os participantes do estudo eram abordados por um entrevistador treinado para dar início a uma avaliação inicial completa. Em seguida, os indivíduos respondiam anulmente a questionários que abordavam questões demográficas, características epidemiológicas, perfil comportamental e consumo alimentar por cinco anós após o diagnóstico inicial. Também foram coletadas informações adicionais que abordavam questões referentes ao regime alimentar dos participantes por dois anos após o primeiro diagnóstico.

Informações sobre histórico de câncer pessoal e familiar também foram coletadas pelos entrevistadores. Por fim, anualmente durante os cinco anos em que os participantes foram acompanhados, questionários avaliando o bem-estar físico e emocional, hábitos comportamentais e a saúde como um todo foram obtidos. No que diz respeito ao estágio clínico do câncer, foram utilizadas diretrizes contidas na sétima edição do manual de classificação de estágios do câncer elaborado pelo *American Joint Committee on Cancer* (AJCC), embora um pequeno grupo de participantes tenha sido submetido a métodos de classificação contidos na oitava edição do mesmo manual. Nesse contexto, duas variáveis correspondentes ao estadiamento do câncer foram coletadas: a variável T, que corresponde à extensão anatômica do tumor primário; e a variável N, que se refere ao número de linfonodos onde o câncer foi observado (SOBIN; WITTEKIND, 2002).

Comorbidades prévias foram avaliadas utilizando a Avaliação de Comorbidades em Adultos 27 (*Adult Comorbidity Evaluation 27*, em inglês). A ACE-27 é uma escala validada com escores de comorbidades em pacientes com câncer, sendo dividida em ausente, leve, moderada e severa.

5.2 Variáveis

A base de dados possui 17 variáveis, a saber:

1. Idade no momento do diagnóstico;
2. Sexo;
3. Raça;
4. Maior grau de instrução;
5. Estado civil;
6. Estadiamento (T);
7. Estadiamento (N);
8. Presença de câncer prévio que não o de cabeça e pescoço;
9. Escore geral de comorbidades ACE-27;
10. Índice de massa corporal (IMC);
11. Tabagismo;
12. Consumo de bebida alcóolica;
13. Local do câncer;
14. Tipo do tratamento recebido;
15. Status de infecção por HPV determinado por PCR ou índice patológico clínico p16;
16. Óbito (evento em estudo);
17. Tempo total de sobrevivência em meses (resposta de interesse).

5.3 Análise Descritiva

Inicialmente, foi realizada a análise descritiva das variáveis categóricas com gráficos de barras para a frequência e para a frequência em relação à variável indicadora de falha ou censura. Além disso, foram feitos gráficos de Kaplan-Meier com curvas de sobrevivência para cada um dos grupos. Foram utilizados medidas-resumo, histogramas e gráficos *box-plot* para a frequência pelos níveis da variável indicadora de falha ou censura. O teste de *logRank* foi aplicado para comparar a distribuição de sobrevivência de diferentes grupos sob as seguintes hipóteses:

$$\begin{cases} H_0: & \text{Não existe diferença entre as curvas de sobrevivência} \\ H_1: & \text{Pelo menos uma das curvas de sobrevivência é diferente das demais} \end{cases} \tag{5.3.1}$$

5.3.1 Gráfico de Kaplan-Meier para a base de dados completa

Ao realizar o gráfico de Kaplan-Meier com base no nosso conjunto de dados completo, observamos que a estimativa de P(T > t) obtida sugere uma assíntota diferente de zero e, portanto, o uso de um modelo com fração de cura é apropriado:

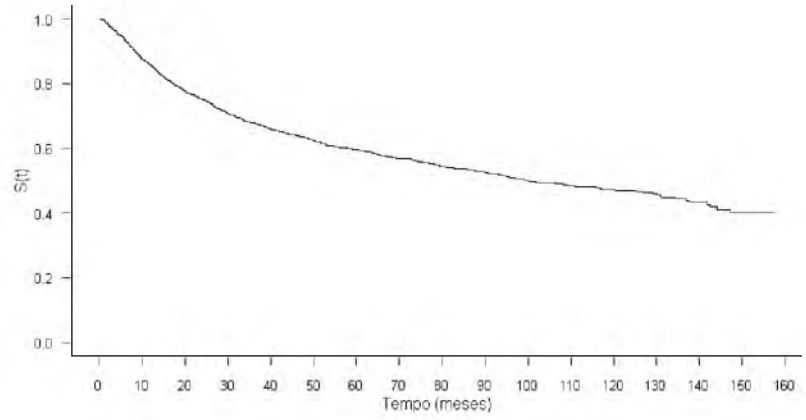

Figura 1: Gráfico de Kaplan-Meier para a base de dados completa

5.3.2 Idade

No nosso banco de dados, esta é uma variável quantitativa discreta que representa a idade (em anos completos) no momento do diagnóstico. Medidas-resumo (Tabela 1) e histograma (Figura 2) da variável idade são apresentados a seguir,

Tabela 1: Medidas-resumo para a variável idade em anos completos

Mínimo	1° quartil	Mediana	Média	3° quartil	Máximo
21,0	52,0	59,0	59,9	68,0	88,0

a partir dos quais podemos observar que aproximadamente 75% dos pacientes em estudo têm entre 52 e 88 anos.

Adicionalmente, realizamos o gráfico *boxplot* para a idade e medidas-resumo agrupadas por status (apresentadas na Figura 3 e Tabela 2, respectivamente).

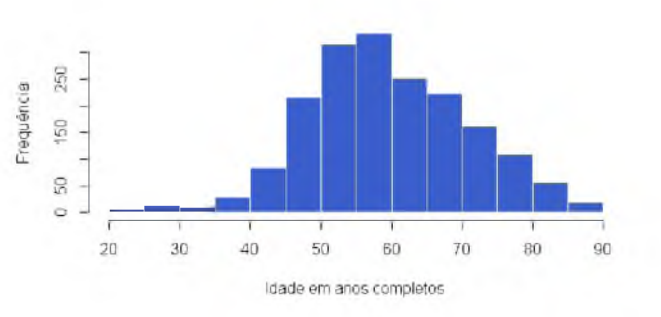

Figura 2: Histograma para a variável idade

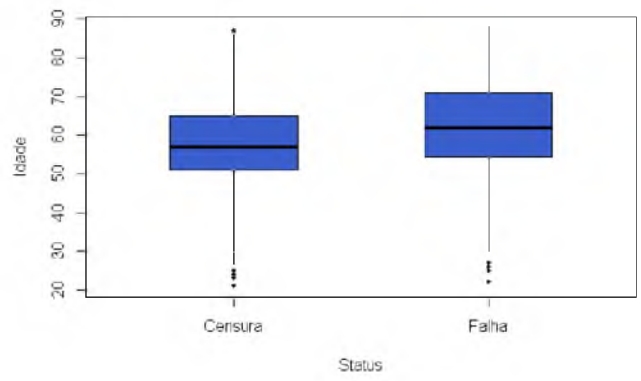

Figura 3: Gráfico *boxplot* para a variável idade vs status

Tabela 2: Medidas-resumo agrupadas por status de censura para a variável idade em anos completos

Status	Mínimo	1° quartil	Mediana	Média	3° quartil	Máximo
Censura	21,0	51,0	57,0	58,0	65,0	87,0
Falha	22,0	54,5	62,0	62,6	71,0	88,0

O gráfico *boxplot* indica a relação da idade com a variável status. Primeiramente, notamos que a mediana, o primeiro e o terceiro quantis são superiores para o caso das falhas. A caixa que corresponde às censuras é pouco mais achatada, o que indica uma distribuição de idades levemente mais homogênea. Adicionalmente, nota-se a presença de mais valores atípicos para o grupo das censuras.

5.3.3 Sexo

Dividida em masculino e feminino, esta é uma variável binária que indica o sexo do paciente em estudo. Como indicado pelo gráfico de barras abaixo, o banco de dados utilizado neste estudo consiste em aproximadamente três vezes mais indivíduos do sexo masculino do que feminino.

Em relação à proporção de falha e censura, o gráfico de barras representado na

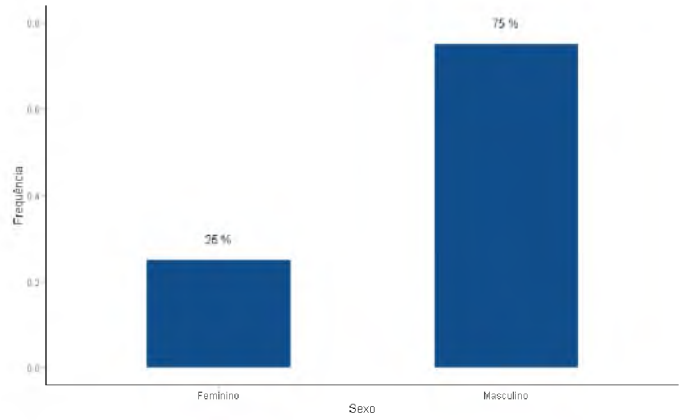

Figura 4: Gráfico de barras para a variável sexo

Figura 5 nos mostra que, à primeira vista, não existe diferença quando agrupamos os indivíduos em estudo por sexo.

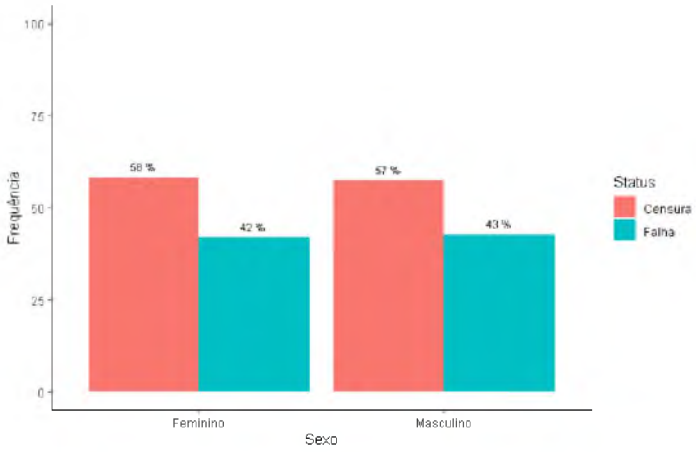

Figura 5: Gráfico de barras para a variável status agrupada por sexo

Reforçando a hipótese de que não há diferença entre a distribuição de probabilidade de sobrevivência entre os dois grupos, o gráfico de Kaplan-Meier nos traz duas curvas bastante semelhantes, como evidenciado na Figura 6.

O teste de *logRank* foi realizado com o intuito de avaliar a diferença entre as curvas de sobrevivência,e os resultados são apresentados na Tabela 3.

Tabela 3: Teste de *logRank* para a variável sexo

Variável	Estatística do teste	g.l.	p-valor
Sexo	0,02	1	0,89

Considerando um nível de significância de 5%, o p-valor obtido de 0,89 não nos dá indícios de que a hipótese nula deve ser rejeitada, ou seja, não há evidências estatísticas de que há diferença entre as duas curvas de sobrevivência.

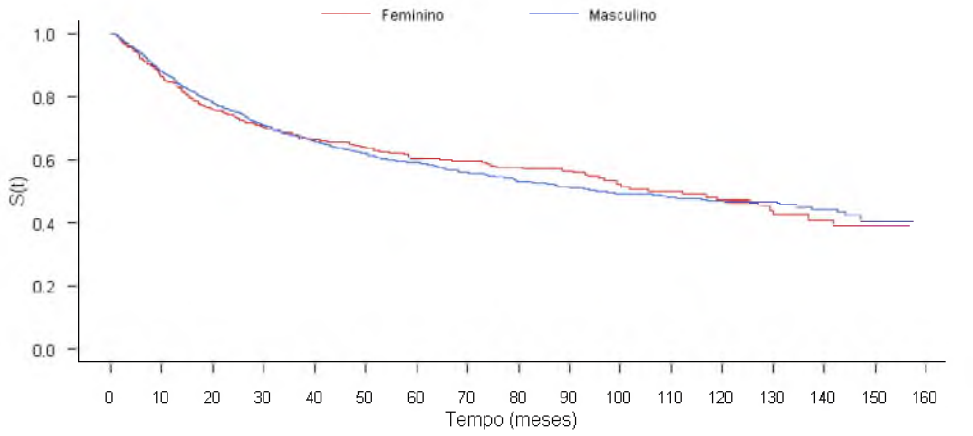

Figura 6: Gráfico das curvas de sobrevivência por sexo

5.3.4 Raça

A variável qualitativa categórica raça está dividida em três categorias: branco, não declarado e outros. Aproximadamente 90% dos indivíduos em estudo se declararam pessoas brancas, com o restante sendo composto pelas demais categorias.

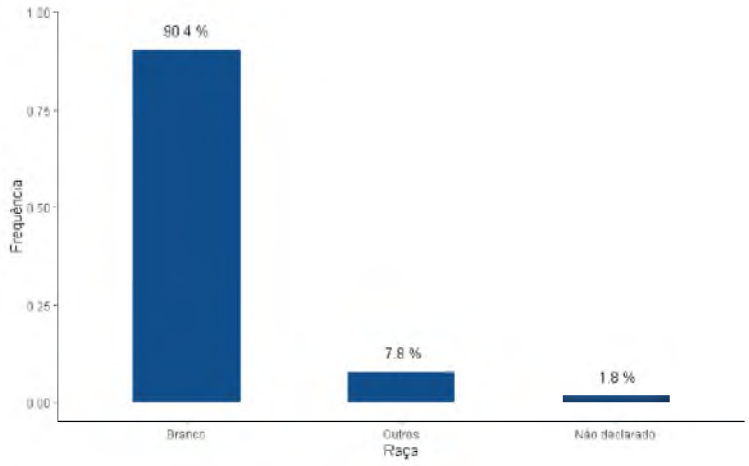

Figura 7: Gráfico de barras para a variável raça

Quando analisamos a proporção de falha e censura agrupada por raça, notamos uma disparidade entre as categorias representadas em nossa base de dados, como evidenciado na Figura 7. Por exemplo, quando ignoramos o grupo que não declarou raça, a proporção de indivíduos brancos que apresentaram o evento de interesse é inferior a outras raças.

As curvas de sobrevivência representadas na Figura 9 evidenciam mais ainda esta disparidade, uma vez que a curva correspondente a pessoas brancas apresenta uma queda

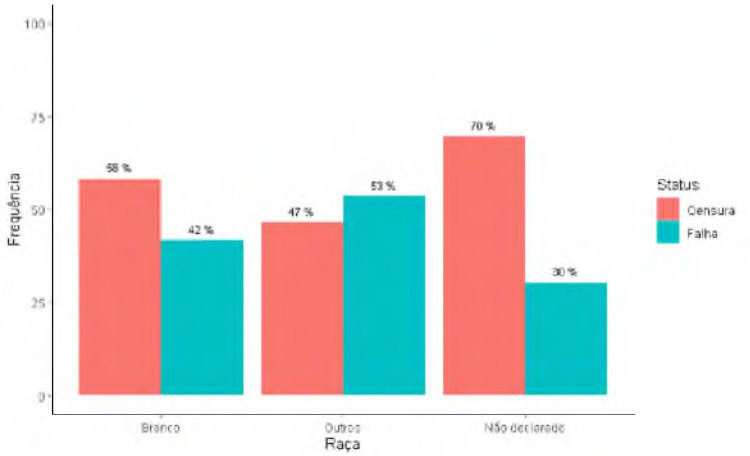

Figura 8: Gráfico de barras para a variável status agrupada por raça

menos acelerada e assíntota superior às demais curvas.

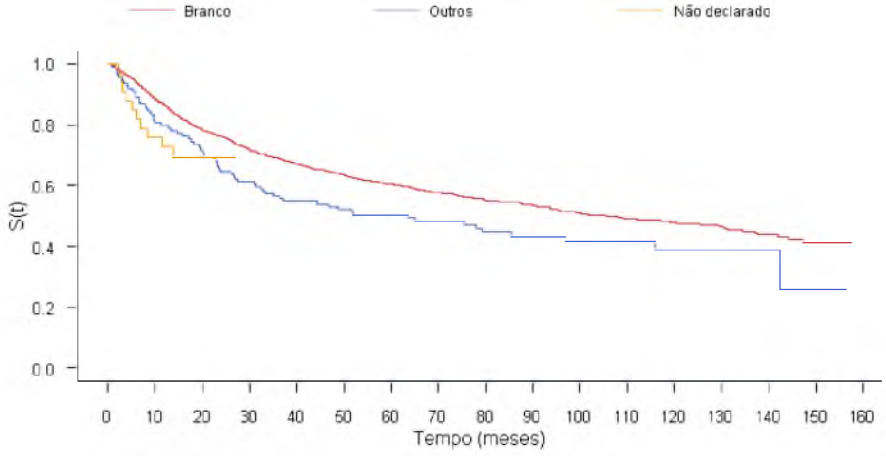

Figura 9: Gráfico das curvas de sobrevivência por raça

Para obtermos uma conclusão mais precisa a respeito desta disparidade, realizmaos um teste de *logRank*. Os resultados são apresentados na Tabela 4.

Tabela 4: Teste de *logRank* para a variável raça

Variável	Estatística do teste	g.l.	p-valor
Raça	9,4	2	0,009

O p-valor de 0,001 nos indica que, a um nível de significância de 5%, temos evidências estatísticas o suficiente para rejeitarmos a hipótese nula de que não existe diferença entre as curvas de sobrevivência.

5.3.5 Maior Grau de Instrução

Temos seis categorias para a variável maior grau de instrução: inferior a ensino médio completo, ensino médio completo, ensino superior incompleto, ensino superior completo, pós-graduação e NAs. As categorias com maior representação na nossa base são ensino superior incompleto e ensino médio completo, ambas com aproximadamente 26% de indivíduos, seguidas por não informado, com 20,9% dos indivíduos e, por ultimo, as categorias pós-graduação, inferior a ensino médio completo e ensino superior completo, todas com pouco menos de 10%.

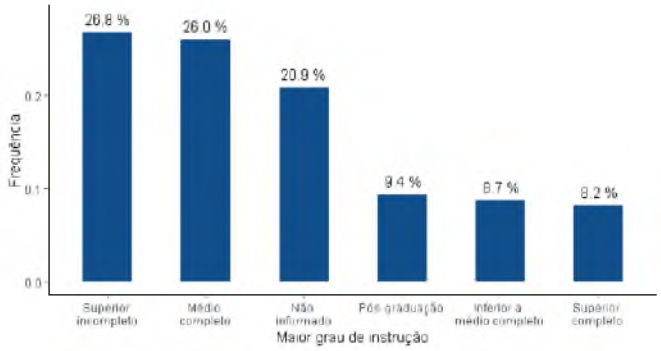

Figura 10: Gráfico de barras para a variável grau de instrução

Analisando a proporção de falha e censura segundo grau de instrução, verificamos que a proporção de falhas diminui à medida que o grau de instrução aumenta, como representado na Figura 11.

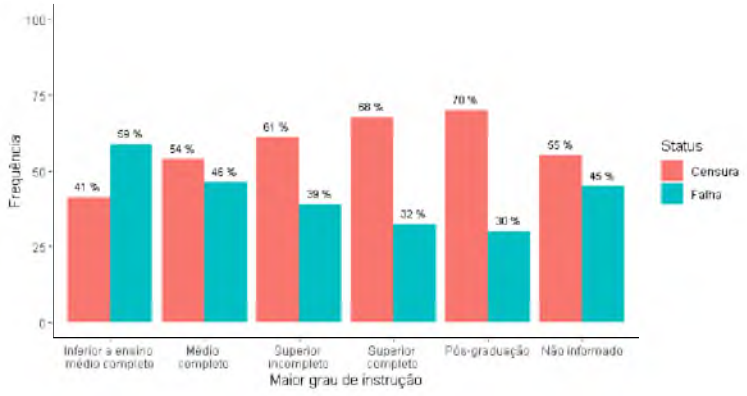

Figura 11: Gráfico de barras para a variável status agrupada por grau de instrução

Adicionalmente, as curvas de sobrevivência corroboram a hipótese de que a probabilidade de sobrevivência aumenta à medida que aumenta o grau de instrução (Figura 12).

Realizamos o teste de *logRank* para avaliar a hipótese de diferença entre as curvas. Os resultados são apresentados na Tabela 5.

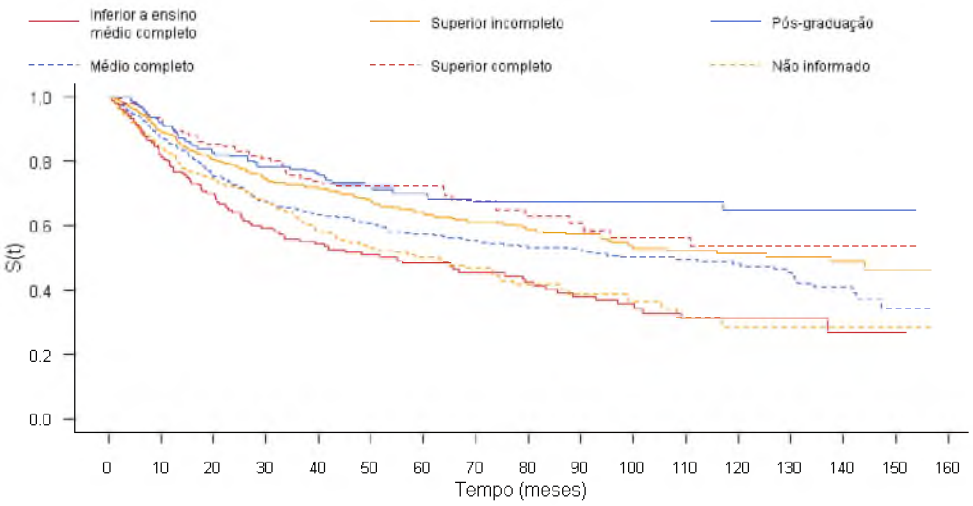

Figura 12: Gráfico das curvas de sobrevivência por grau de instrução

Tabela 5: Teste de *logRank* para a variável maior grau de instrução

Variável	Estatística do teste	g.l.	p-valor
Maior Grau de Instrução	50,0	5	$1,4 \times 10^{-9}$

O p-valor se aproxima bastante de zero e, portanto, a um nível de significância de 5%, há evidências estatísticas significativas para rejeitarmos a hipótese nula de que não há diferença entre as curvas de sobrevivência.

5.3.6 Estado Civil

A variável estado civil é dividida em cinco categorias: casado, separado, divorciado, viúvo e solteiro. Dentre elas, a categoria que corresponde a pessoas casadas possui o maior percentual de indivíduos em nossa base (aproximadamente 60%), como apresentado na Figura 13.

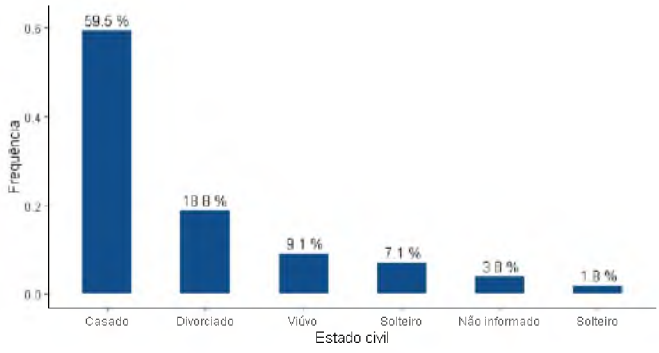

Figura 13: Gráfico de barras para a variável estado civil

Quando analisamos a proporção de falha e censura segundo o estado civil, pode-

mos perceber que, em pessoas casadas, a proporção de falhas é a menor dentre todas as categorias quando excluímos os indivíduos que não informaram o estado civil. Analogamente, esta proporção é maior em pessoas viúvas, como evidenciado na Figura 14.

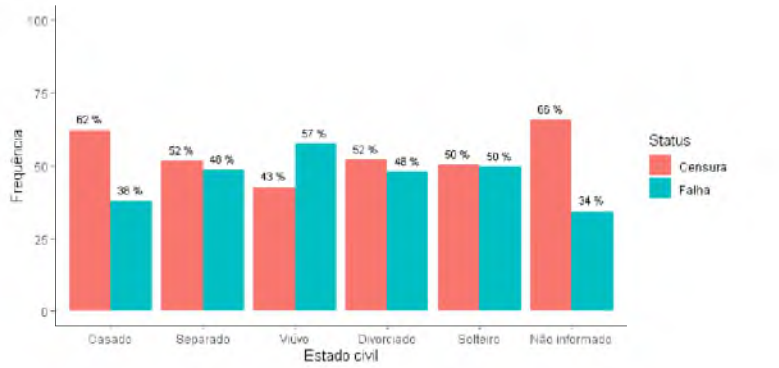

Figura 14: Gráfico de barras para a variável status agrupada por estado civil

Em seguida, realizamos o gráfico das estimativas de Kaplan-Meier para as curvas de sobrevivência (Figura 15). As curvas apresentam um comportamento semelhante até aproximadamente 90 meses, que é quando a curva que corresponde a pessoas viúvas começa a decair mais rapidamente do que as demais.

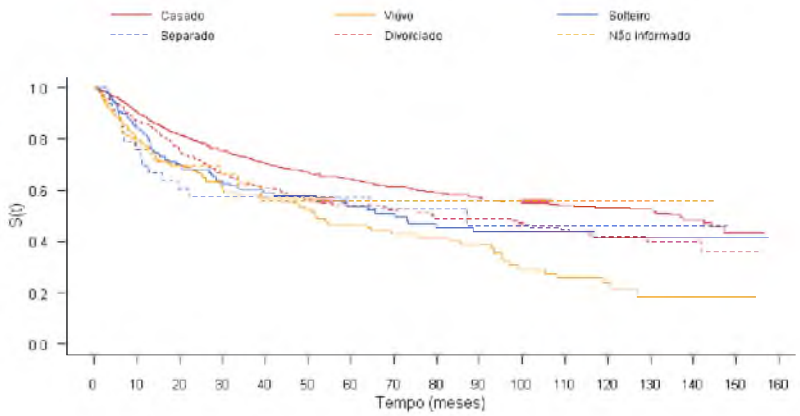

Figura 15: Gráfico das curvas de sobrevivência por estado civil

O teste de *logRank* é realizado para mensurar com mais precisão as diferenças entre as curvas de sobrevivência observadas no gráfico de Kaplan-Meier. Os resultados são apresentados na Tabela 6.

Tabela 6: Teste de *logRank* para a variável estado civil

Variável	Estatística do teste	g.l.	p-valor
Estado Civil	44,46	5	$1,9 \times 10^{-8}$

A um nível de significância de 5%, o p-valor obtido nos dá evidências estatísticas

suficientes para a rejeição da hipótese nula de que não há diferença entre as curvas de sobrevivência.

5.3.7 Estadiamento (T)

A variável de estadiamento T, que corresponde à extensão anatômica do tumor primário, é agrupada em parâmetros graduais de T0 a T4, onde T0 corresponde a características mais brandas em comparação à T4, que se refere a observações mais severas (SOBIN; WITTEKIND, 2002). As distribuições de cada uma das categorias são apresentada na Figura 16.

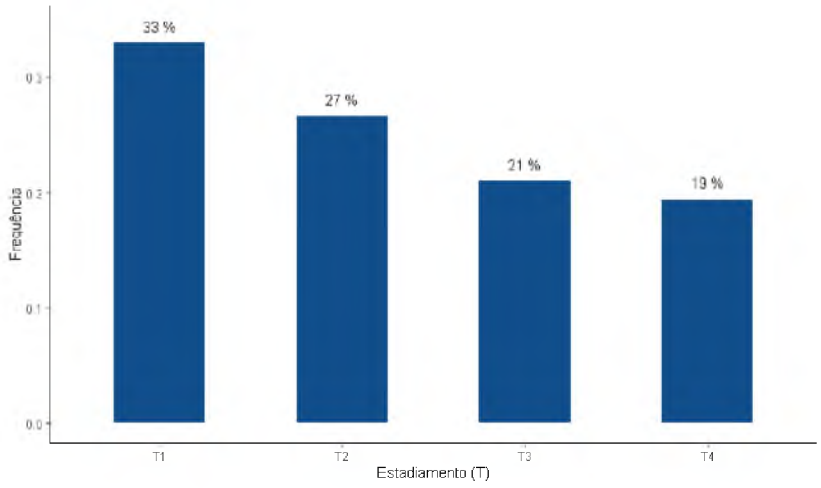

Figura 16: Gráfico de barras para a variável estadiamento (T)

A proporção de falha e censura de acordo com o estadiamento (T) variou bastante entre as categorias, onde a proporção de falhas foi maior no estádio T4 (Figura 17).

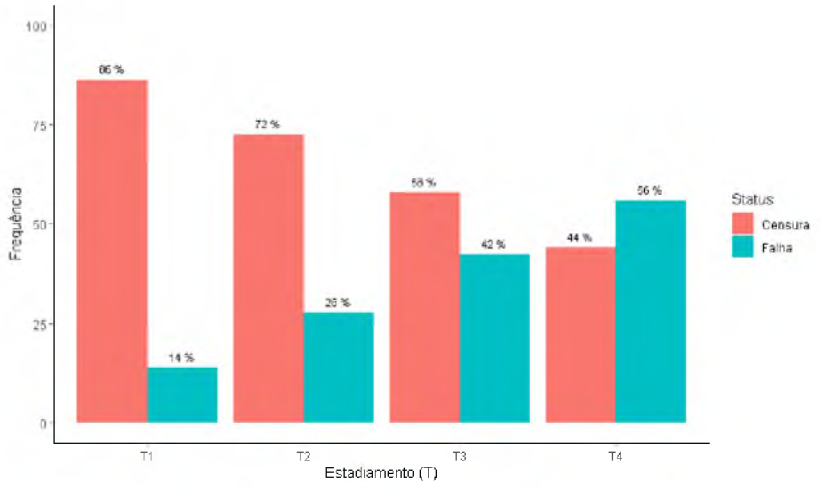

Figura 17: Gráfico de barras para a variável status agrupada por estadiamento (T)

A análise gráfico das curvas de sobrevivência reforçou a diferença entre as cate-

gorias de estadiamento (T): à medida que o estádio aumenta, a probabilidade de sobrevivência diminui para aproximadamente todos os tempos de observação (Figura 18).

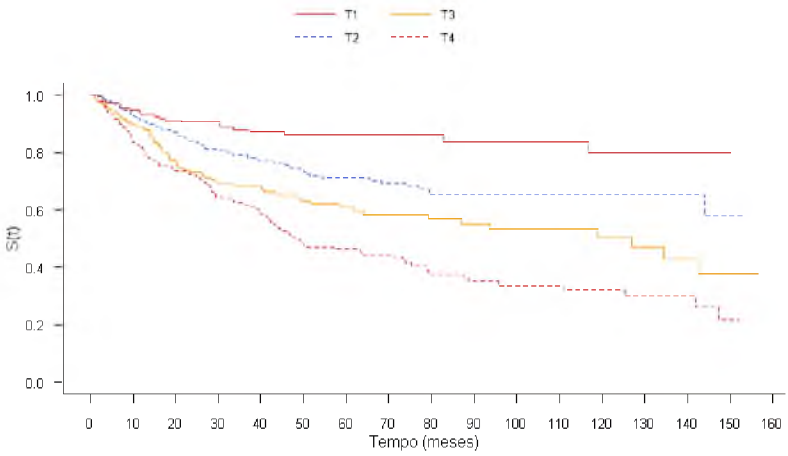

Figura 18: Gráfico das curvas de sobrevivência por estadiamento (T)

Um teste de *logRank* foi feito para as diferenças entre as curvas de sobrevivência (Tabela 7).

Tabela 7: Teste de *logRank* para a variável estadiamento (T)

Variável	Estatística do teste	g.l.	p-valor
Estadiamento (T)	69,1	3	7×10^{-15}

A um nível de significância de 5%, o p-valor obtido de 5×10^{-16} nos dá evidências estatísticas suficentes para a rejeição da hipótese nula.

5.3.8 Estadiamento (N)

A variável de estadiamento N, que corresponde ao número de linfonodos em que o câncer foi observado, é dividida em parâmetros graduais de N0 a N3 (SOBIN; WITTEKIND, 2002). As distribuições de cada uma das categorias são apresentada na Figura 19.

Analisando a proporção de falha e censura de acordo com o estadiamento (N), observou-se uma diferença considerável entre as categorias, onde a proporção de falhas foi maior no estádio N3 (Figura 20).

O gráfico das curvas de sobrevivência reforçou a diferença entre os parâmetros de estadiamento (N). Contraintuitivamente, percebe-se que a probabilidade de sobrevivência é maior em N1, e não em N0. Em relação às demais categorias, à medida que o estádio aumenta, a probabilidade de sobrevivência diminui em praticamente todos os tempos de

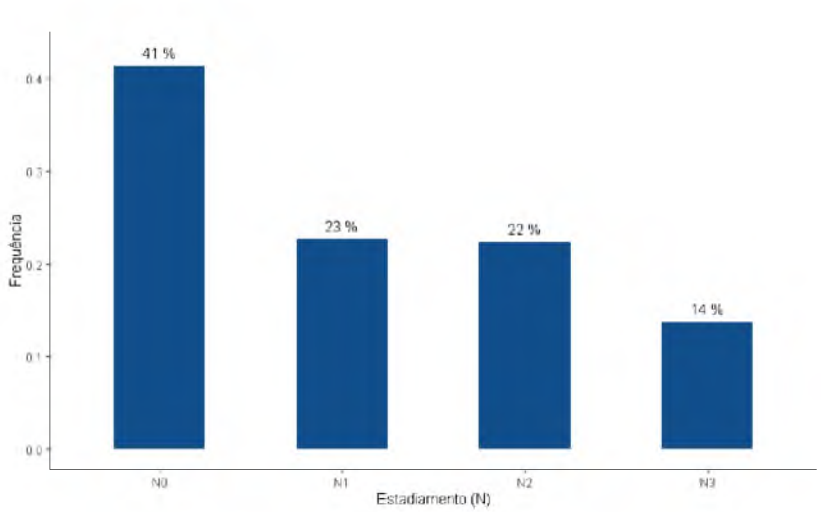

Figura 19: Gráfico de barras para a variável estadiamento (N)

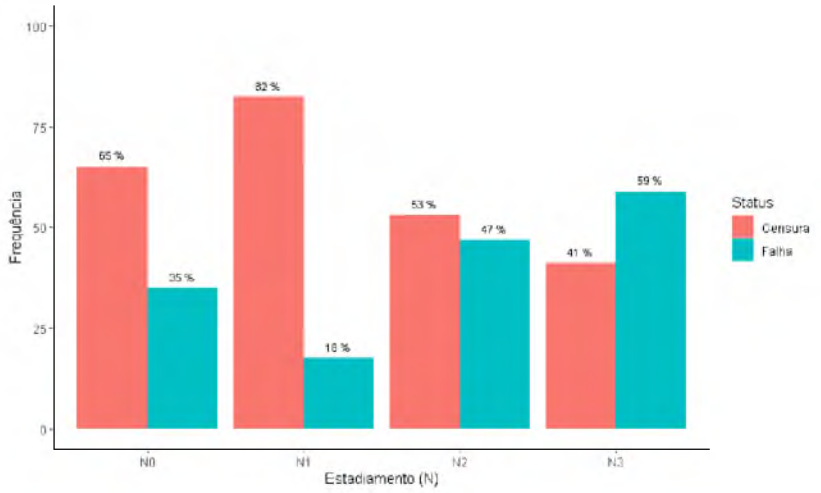

Figura 20: Gráfico de barras para a variável status agrupada por estadiamento (N)

observação (Figura 21).

Foi feito um teste de *logRank* para as diferenças entre as curvas de sobrevivência (Tabela 8).

Tabela 8: Teste de *logRank* para a variável estadiamento (N)

Variável	Estatística do teste	g.l.	p-valor
Estadiamento (N)	57,6	3	2×10^{-12}

A um nível de significância de 5%, o p-valor obtido de 3×10^{-13} nos dá evidências estatísticas suficentes para a rejeição da hipótese nula.

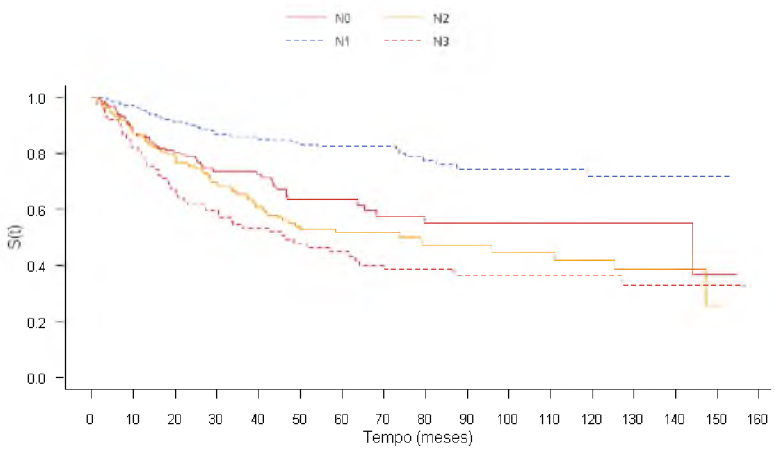

Figura 21: Gráfico das curvas de sobrevivência por estadiamento (N)

5.3.9 Presença de câncer prévio que não o de cabeça e pescoço

A variável que corresponde à presença de câncer prévio que não o de cabeça e pescoço é organizada em sim, não e não soube responder. A nossa base de dados é composta majoritariamente por indivíduos que negaram a presença de câncer prévio, como apresentado na Figura 22.

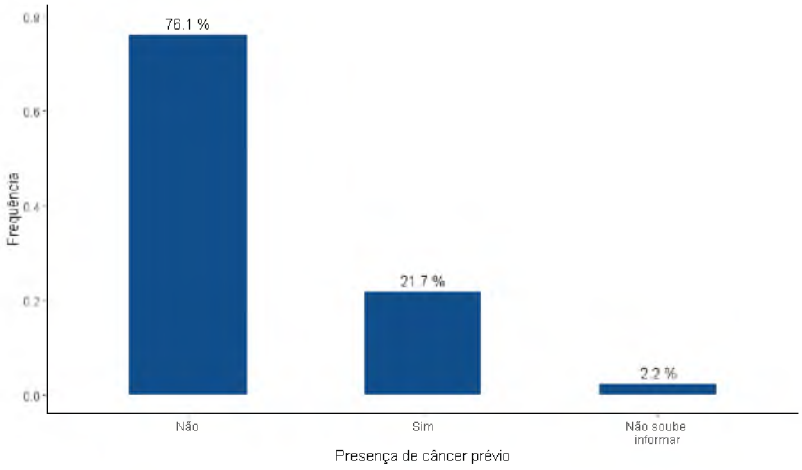

Figura 22: Gráfico de barras para a variável presença de câncer prévio

A análise da proporção de falha e censura segundo a presença de câncer prévio que não o de cabeça e pescoço nos mostra que não há diferença significativa entre as proporções dos grupos Sim e Não, mas que há diferença entre o grupo que não soube informar e as demais categorias (Figura 23).

O gráfico das curvas de sobrevivência nos mostra um comportamente bastante similar entre as curvas até os 60 meses de estudo, período em que a curva que corresponde aos indivíduos que não souberam informar começa a decair mais rapidamente que

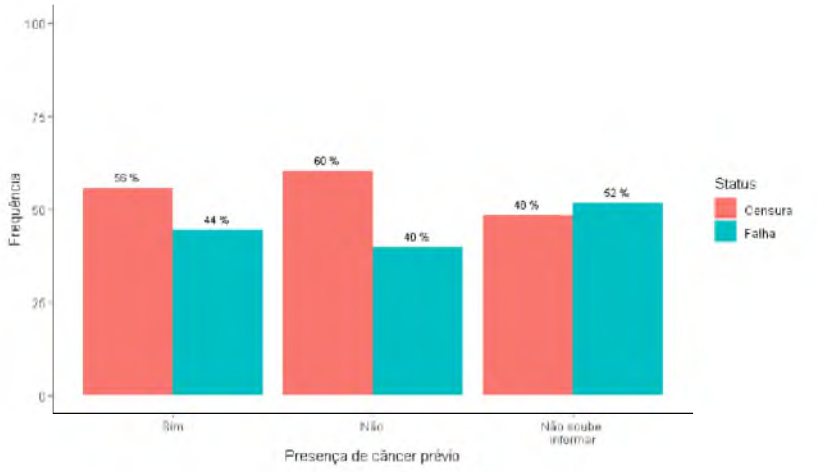

Figura 23: Gráfico de barras para a variável status agrupada por presença de câncer prévio

as demais (excluindo a curva de valores não disponíveis). Adicionalmente, a curva de indivíduos que informaram a presença de câncer prévio apresenta uma queda brusca na probabilidade de sobrevivência após os 140 meses de estudo, como apresentado na Figura 24.

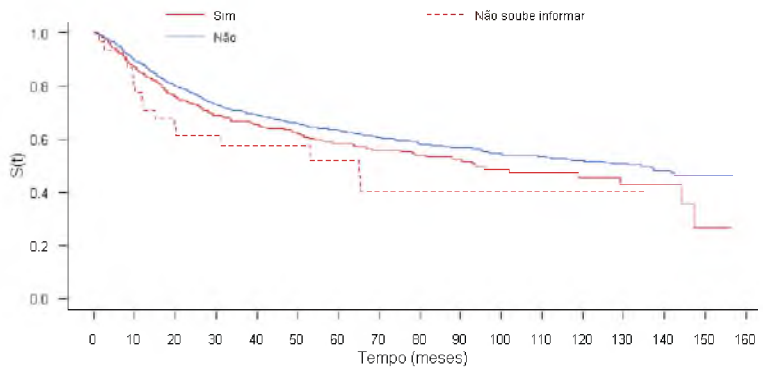

Figura 24: Gráfico das curvas de sobrevivência por presença de câncer prévio

Realizamos o teste de *logRank* para a diferença entre as curvas de sobrevivência, e os resultados são apresentados na Tabela 9.

Tabela 9: Teste de *logRank* para a variável presença de câncer prévio que não o de cabeça e pescoço

Variável	Estatística do teste	g.l.	p-valor
Câncer prévio que não o de cabeça e pescoço	6,0	2	0,049

A um nível de significância de 5%, o p-valor obtido de 0,049 nos dá evidências estatísticas suficientes para a rejeição da hipótese nula de que não há diferença entre as curvas de sobrevivência.

5.3.10 Escore geral de comorbidades ACE-27

O escore geral do índice de avaliação de comorbidades em adultos (ACE, da sigla em inglês) é representado no nosso banco de dados por uma variável categórica divida em quatro grupos: nenhum, leve, moderado e severo. A Figura 25 nos traz a distribuição das categorias em nossa base de dados.

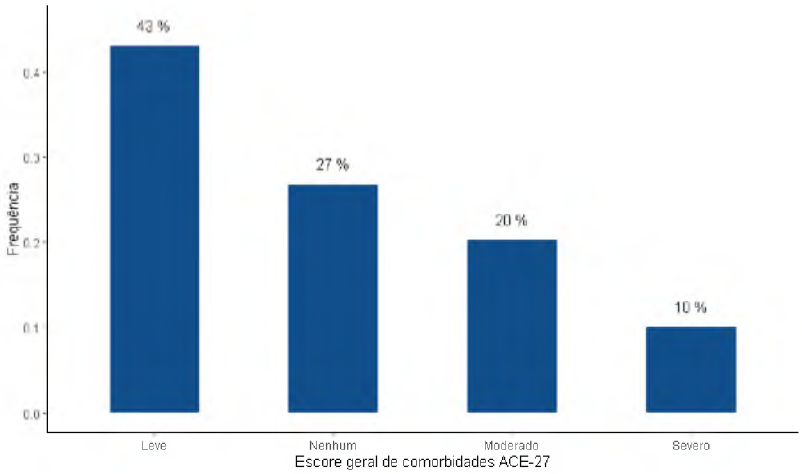

Figura 25: Gráfico de barras para a variável Escore geral ACE-27

A análise da proporção de falha e censura nos mostra que, à medida que escore de comorbidades aumenta, também aumenta a proporção de falhas entre os indivíduos (Figura 26), o que é intuitivo e esperado.

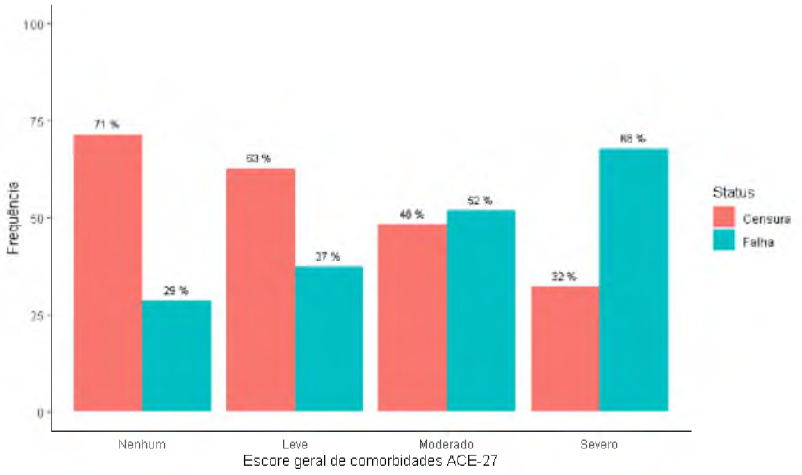

Figura 26: Gráfico de barras para a variável status agrupada por escore geral de comorbidades ACE-27

Adicionalmente, o gráfico das curvas de sobrevivência traz comportamentos claramente diferentes para as curvas de cada uma das categorias, e podemos constatar que a probabilidade de sobrevivência diminui à medida que o escore de comorbidades aumenta.

Para obter mais embasamento e precisão para nossas análises prévias, realizamos

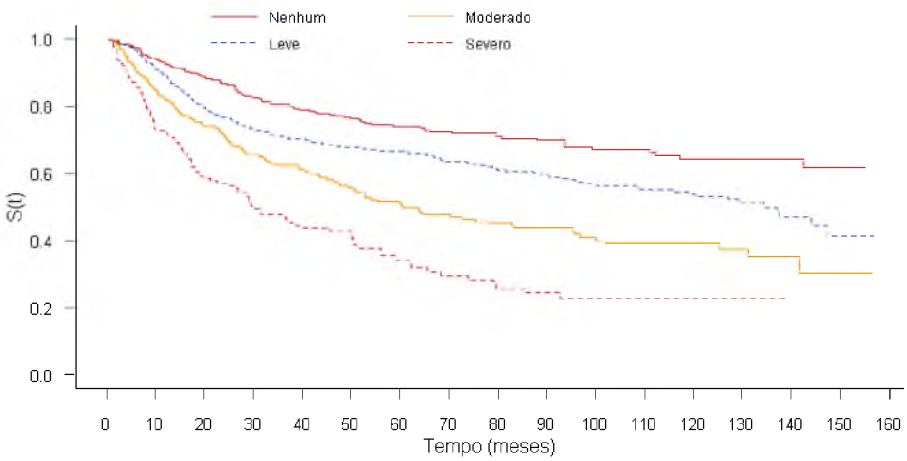

Figura 27: Gráfico das curvas de sobrevivência por escore geral de comorbidades

o teste de *logRank*. Os resultados obtidos são apresentados na Tabela 10.

Tabela 10: Teste de *logRank* para a variável escore geral de comorbidades ACE-27

Variável	Estatística do teste	g.l.	p-valor
Escore geral ACE-27	109,0	3	2×10^{-16}

Como esperado, o p-valor próximo de zero nos dá evidências estatísticas significativas para a rejeição da hipótese nula de que não há diferença entre as curvas, a um nível de significância de 5%.

5.3.11 Índice de massa corporal (IMC)

Esta é uma variável quantitativa contínua obtida pela divisão do peso em kilos do paciente pelo quadrado de sua altura em metros. As medidas-resumo da distribuição de frequências das índices de massa corporal dos pacientes são apresentadas na Tabela 11. Na Figura 28, temos o respectivo histograma.

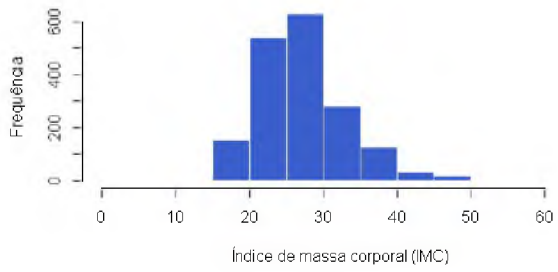

Figura 28: Histograma para a variável IMC

Tabela 11: Medidas-resumo para a variável índice de massa corporal (IMC)

Mínimo	1° quantil	Mediana	Média	3° quantil	Máximo
13,0	23,1	26,4	27,2	30,2	64,6

Com base no histograma e nas medidas-resumo, observamos que aproximadamente 50% dos indivíduos possui IMC entre 26,4 e 64,6 (valores entre 25 e 29,9 indicam sobrepeso). Além disso, aproximadamente 25% dos pacientes tinham IMC acima de 30,0 no momento do diagnóstico, o que representa obesidade ou obesidade severa.

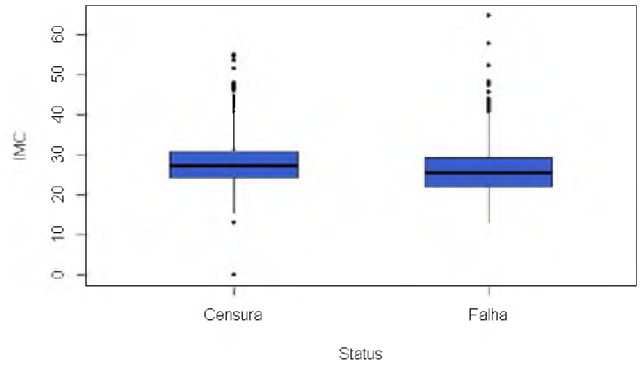

Figura 29: Gráfico *boxplot* para a variável IMC vs status

Tabela 12: Medidas-resumo agrupadas por status de censura para a variável IMC

Status	Mínimo	1° quantil	Mediana	Média	3° quantil	Máximo
Censura	13,1	24,2	27,1	27,9	30,6	55,0
Falha	13,0	21,9	25,4	26,2	29,3	64,6

Por último, a análise do gráfico *boxplot* e da tabela de medidas-resumo segundo o status (apresentados na Figura 29 e na Tabela 12, respectivamente) nos traz resultados bastante similares para os grupos de falha e censura. Nota-se também que a distribuição dos valores de IMC é bastante homogênea nos dois grupos quando não consideramos o valores atípicos, uma vez que as duas caixas são bastante achatadas.

5.3.12 Tabagismo

Esta variável categórica é divida em três grupos: nunca fumou, atual fumante e ex-fumante (não fuma há doze meses ou mais). Na nossa base de dados, aproximadamente 44% dos indivíduos é fumante, 31% são ex-fumantes e 23% dos indivíduos nunca fumou (Figura 30).

O gráfico da proporção de falha e censura segundo tabagismo nos mostra que,

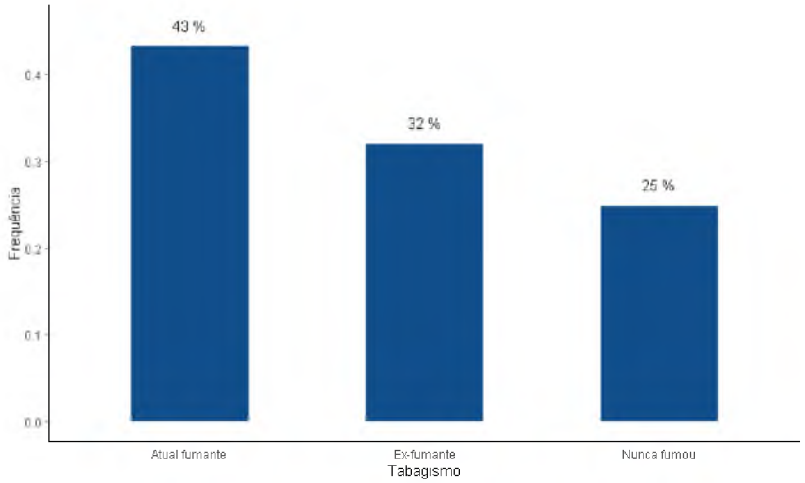

Figura 30: Gráfico de barras para a variável tabagismo

quando comparamos com indivíduos fumantes, a proporção de falhas é bastante inferior em indivíduos que nunca fumaram (Figura 31).

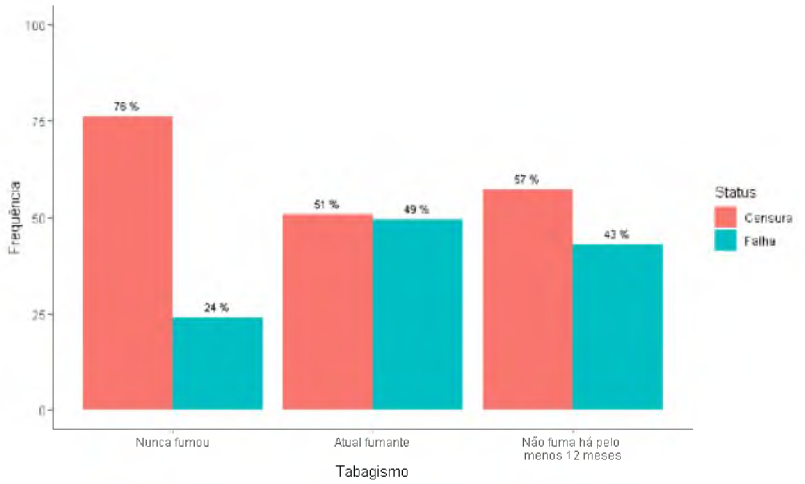

Figura 31: Gráfico de barras para a variável status agrupada por tabagismo

Analisando o gráfico das curvas de sobrevivência, observamos que a curva correspondente a indivíduos que nunca fumaram se comporta diferente das demais, uma vez que para ela a probabilidade de sobrevivência é superior em todos os tempos (Figura 32).

O teste de logRank é realizado para mensurar a diferença entre as curvas de sobrevivência segundo a variável tabagismo, e o resultados são apresentados na Tabela 13.

Tabela 13: Teste de *logRank* para a variável tabagismo

Variável	Estatística do teste	g.l.	p-valor
Tabagismo	48,50	2	3×10^{-11}

A um nível de significância de 5%, o p-valor próximo de zero obtido no teste nos

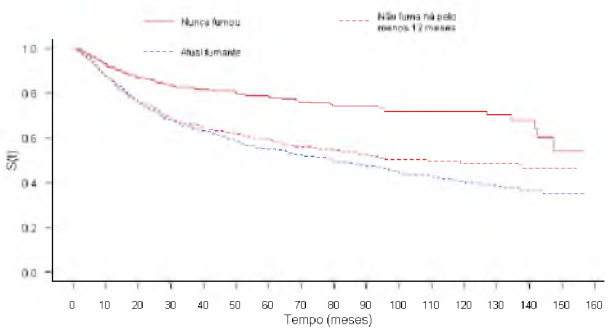

Figura 32: Gráfico das curvas de sobrevivência por tabagismo

traz evidências estatísticas suficientes para a rejeição da hipótese nula.

5.3.13 Consumo de bebida alcóolica

De forma similar à variável tabagismo, esta variável é dividida em três categorias: nunca bebeu, faz uso de bebida alcóolica atualmente e não bebe há doze meses ou mais. Como apresentado na Figura 33, a maior parcela de indivíduos faz uso de bebida alcóolica (aproximadamente 63%), ao passo que apenas 9,5% dos indivíduos nunca bebeu.

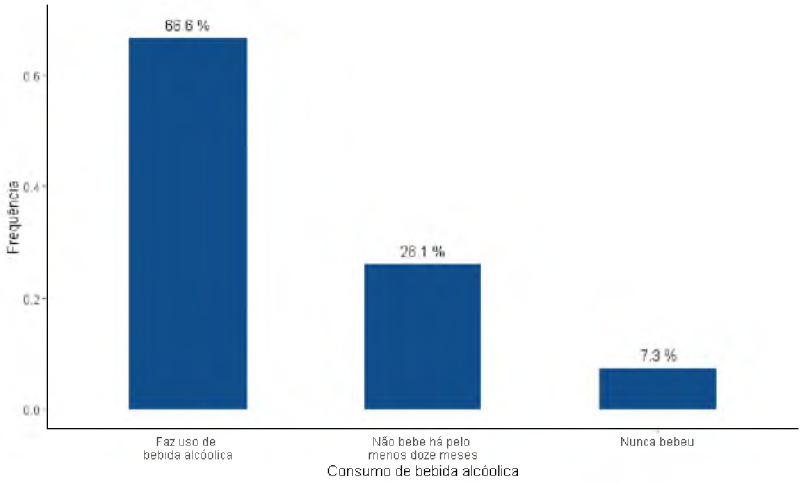

Figura 33: Gráfico de barras para a variável consumo de bebida alcóolica

Entretanto, a análise da proporção de falha e censura nos traz um resultado contraintuitivo: a proporção de falhas é menor no grupo que faz uso de bebida alcóolica do que no grupo que nunca bebeu (Figura 34). Adicionalmente, o grupo que não faz uso de bebida alcóolica há pelo menos doze meses apresentou a maior proporção.

O gráfico das curvas de sobrevivência revela um comportamento bastante similar entre as curvas do grupo que nunca bebeu e do que faz uso de bebida alcóolica até mais ou menos o 140° mês, que é quando a probabilidade de sobrevivência do primeiro grupo

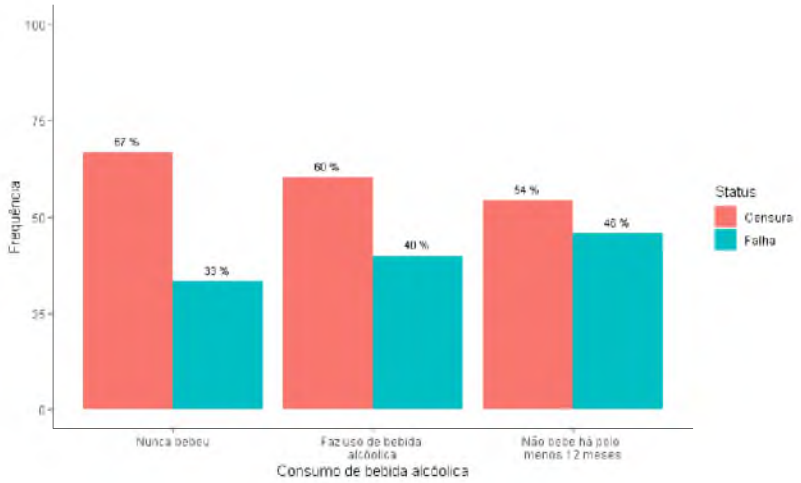

Figura 34: Gráfico de barras para a variável status agrupada por consumo de bebida alcóolica

diminui em relação ao segundo. A curva de indivíduos que não bebe há pelo menos doze meses apresenta um decaimento inferior às outras em praticamente todos os tempos.

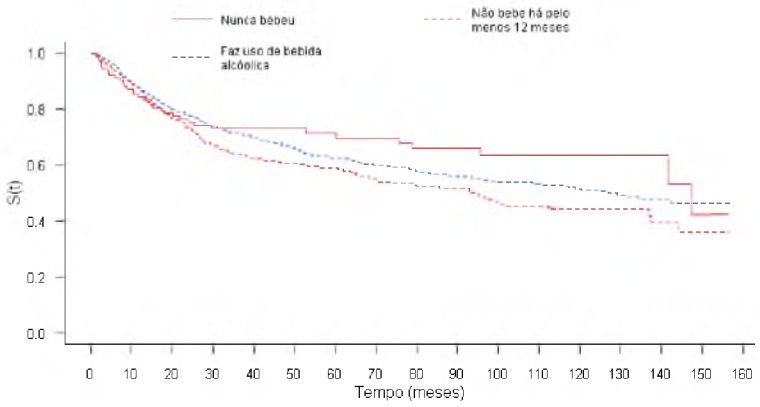

Figura 35: Gráfico das curvas de sobrevivência por consumo de bebida alcóolica

Realizamos o teste de *logRank* para mensurar a diferença entre as curvas. Os resultados estão expostos na Tabela 14.

Tabela 14: Teste de *logRank* para a variável consumo de bebida alcóolica

Variável	Estatística do teste	g.l.	p-valor
Consumo de bebida alcóolica	6,3	2	0,04

A um nível de significância de 5%, o p-valor de 0,01 nos dá evidências estatísticas o suficiente para rejeitarmos a hipótese nula de que não há diferença entre as curvas.

5.3.14 Local do Câncer

A variável local do câncer é dividida em dez categorias: glote, laringe, cavidade oral, orofaringe, hipofaringe, ossos cranianos (cavidade nasal, sinus, crânio), inicialmente desconhecido, nasofaringe, glândulas salivares e outros. A distribuição de cada uma das categorias em nossa base de dados é apresentada na Figura 36.

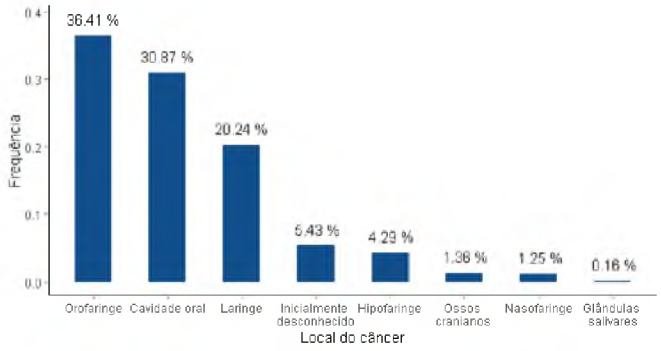

Figura 36: Gráfico de barras para a variável local do câncer

O gráfico da proporção de falha e censura segundo o local do câncer nos traz alguns resultados importantes: indivíduos onde o câncer manifestou-se na hipofaringe apresentaram a maior proporção de falhas, ao passo que pacientes cujo local do câncer foi as glândulas salivares ou inicialmente desconhecido tiveram a menor proporção de falhas.

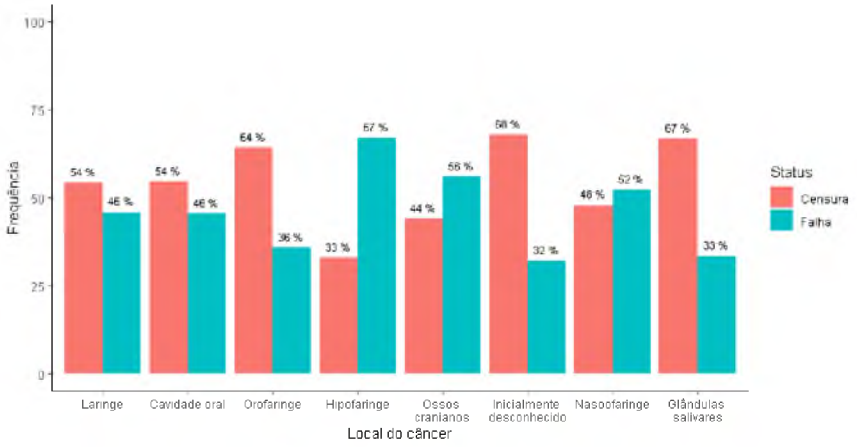

Figura 37: Gráfico de barras para a variável status agrupada por local do câncer

Analisando o gráfico das curvas de sobrevivência, notamos que as curvas para a orofaringe e local inicialmente desconhecido possuem maior probabilidade de sobrevivência para praticamente todos os tempos, ao passo que a curva que corresponde à hipofaringe possui menor probabilidade de sobrevivência para todos os tempos a partir de mais ou menos o 20° mês (Figura 38).

Novamente, o teste de *logRank* é realizado para fornecer uma conclusão mais

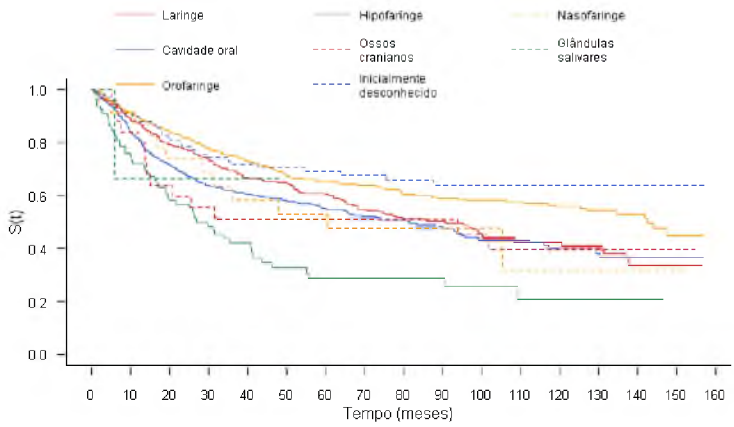

Figura 38: Gráfico das curvas de sobrevivência por local do câncer

precisa a respeito das diferenças entre as curvas (Tabela 15).

Tabela 15: Teste de *logRank* para a variável local do câncer

Variável	Estatística do teste	g.l.	p-valor
Local do câncer	63,13	7	4×10^{-11}

Analisando os resultados apresentados na Tabela 9, concluímos que, a um nível de significância de 5%, temos evidências estatísticas suficientes para a rejeição da hipótese nula de que não há diferença entre as curvas.

5.3.15 Tipo do tratamento recebido

Para o tipo de tratamento recebido, a variável correspondente foi organizada em quinze categorias: apenas cirurgia, cirurgia e radioterapia adjuvante, cirurgia e químio-radioterapia adjuvante, cirurgia e quimioterapia adjuvante, apenas radioterapia, radioterapia e cirurgia adjuvante, radioterapia e quimioterapia adjuvante, apenas químio-radioterapia, químio-radioaterapia e cirurgia adjuvante, apenas quimioterapia, nenhum tratamento curativo recebido antes do óbito, nenhuma modalidade de tratamento recebida após um ano de acompanhamento, nenhuma modalidade de tratamento recebida durante o primeiro ano de acompanhamento, paliativo ou desconhecido. As distribuições para cada tipo de tratamento são apresentadas na Figura 39.

O gráfico das proporções de falha e censura segundo o tipo de tratamento traz informações relevantes: indivíduos em que o tratamento utilizado foi apenas a quimioterapia, tratamento paliativo ou desconhecido, as proporções de falha foram significativamente superiores aos demais tipos de tratamento. Analogamente, a menor proporção de falha observada foi a do grupo que teve apenas a cirurgia como tipo de tratamento (Figura 40).

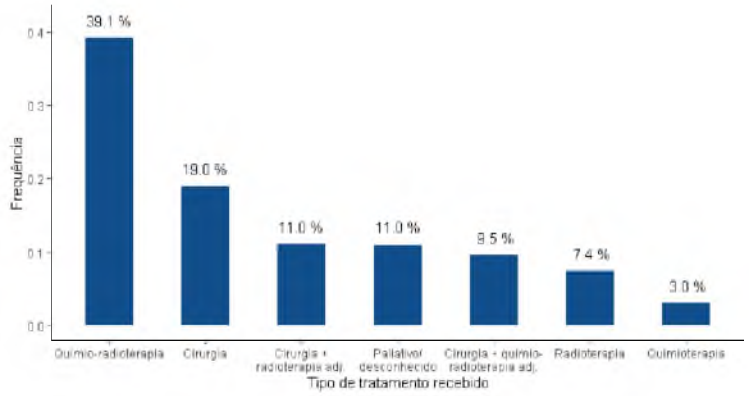

Figura 39: Gráfico de barras para a variável tipo de tratamento

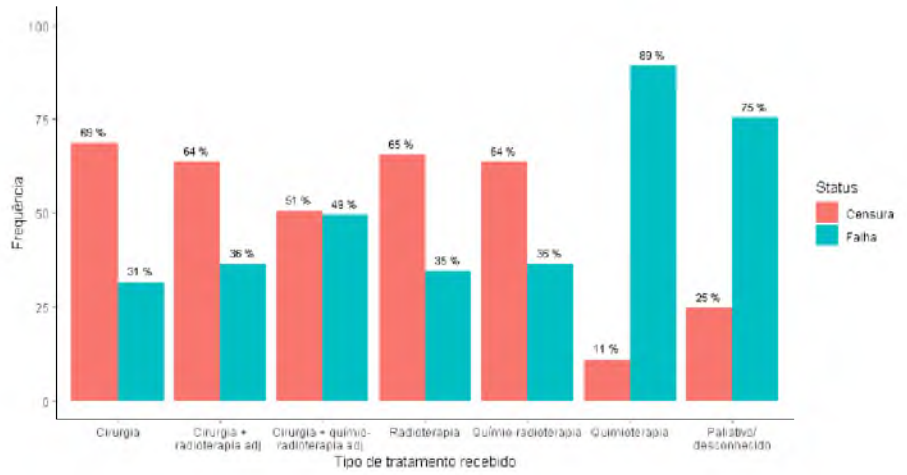

Figura 40: Gráfico de barras para a variável status agrupada por tipo de tratamento

O gráfico das curvas de sobrevivência corrobora as observações feitas a partir das proporções de falha e censura, uma vez que as probabilidades de sobrevivência para as curvas que correspondem à quimioterapia e a tratamentos paliativos ou desconhecidos são consideravelmente menores para praticamente todos os tempos (Figura 41). As demais curvas apresentaram comportamento semelhante, com exceção da curva para a cirurgia aliada à químio-radioterapia adjuvante, que teve probabilidades de sobrevivência levemente inferiores para todos os tempos a partir de mais ou menos o 20° mês.

Mais uma vez, utilizamos o teste de *logRank* para corroborar as hipóteses levantadas até aqui (Tabela 16).

Tabela 16: Teste de *logRank* para a variável tipo do tratamento recebido

Variável	Estatística do teste	g.l.	p-valor
Tipo do tratamento recebido	423,04	6	2e-16

Como esperado, o p-valor obtido muito se aproximou de zero. Portanto, a um nível de significância de 5%, temos evidências estatísticas suficientes para a rejeição da hipótese nula.

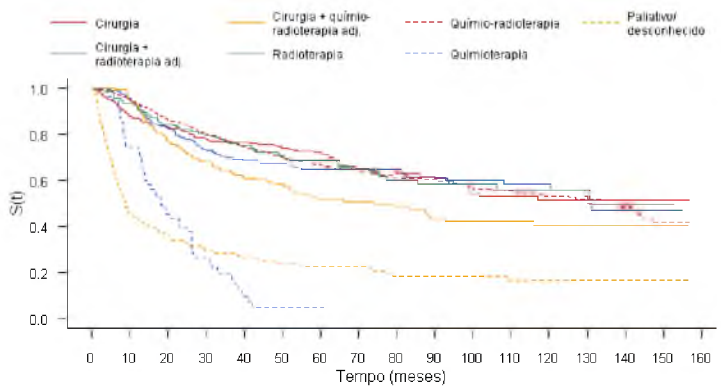

Figura 41: Gráfico das curvas de sobrevivência por tipo de tratamento

5.3.16 Status de infecção por HPV determinado por PCR ou índice patológico clínico p16

Esta variável categórica foi organizada em três grupos: positivo, negativo e inválido/indisponível. As distribuições para cada categoria são apresentadas na Figura 42.

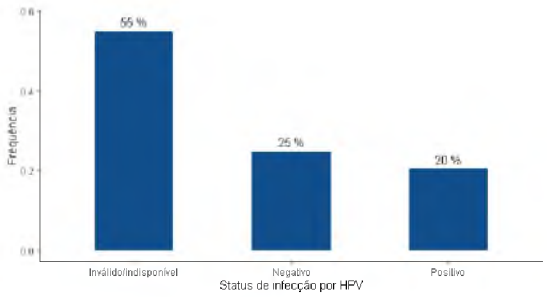

Figura 42: Gráfico de barras para a variável Status de infecção por HPV

O gráfico das proporções de falha e censura segundo o status de infecção por HPV nos permite observar que, contraintuitivamente, a proporção de falhas é significativamente inferior no grupo que testou positivo para o vírus (Figura 43).

Adicionalmente, o gráfico das curvas de sobrevivência indica que, para todos os tempos observados, a probabilidade de sobrevivência é superior para o grupo que testou positivo para o HPV (Figura 44).

Por último, realizamos o teste de *logRank* para mensurar as diferenças entre as curvas. Os resultados obtidos são apresentados na Tabela 17.

Tabela 17: Teste de *logRank* para a variável status de infecção por HPV

Variável	Estatística do teste	g.l.	p-valor
Status de infecção por HPV	62,46	2	3×10^{-14}

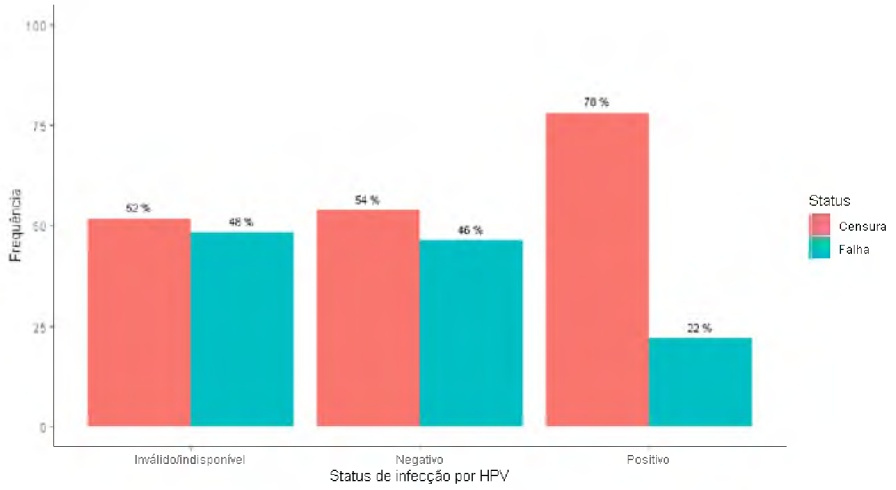

Figura 43: Gráfico de barras para a variável status agrupada por status de infecção por HPV

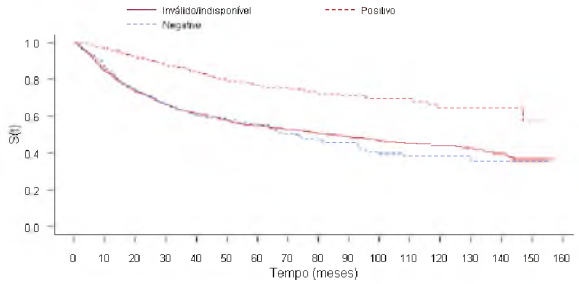

Figura 44: Gráfico das curvas de sobrevivência por status de infecção por HPV

O p-valor obtido muito se aproxima de zero, o que nos dá evidências estatísticas suficientes para a rejeição da hipótese nula, a um nível de significância de 5%.

5.4 Modelo

Inicialmente, utilizamos o pacote penPHcure, no R, para o ajuste do modelo de riscos proporcionais de Cox com fração de cura contendo todas as variăveis explicativas selecionadas no nosso estudo. A modelagem da parte dos incidentes, que diz respeito ă probabilidade de cura, e da parte da latência, que corresponde ao comportamento do tempo de falha, é feita separadamente por meio do uso de uma função de ligação *logit* e o ajuste de um modelo de riscos proporcionais de Cox, respectivamente. As estimativas dos parâmetros e os intervalos de confiança a um nível de confiança de 95% para as duas partes são apresentados abaixo:

Tabela 18: Estimativas para os coeficientes do modelo referente à probabilidade de cura (incidentes) e intervalos de confiança

Parâmetro	Estimativa	2,5%	97,5%
Intercepto	-4,57	-11,01	-1,04
IMC	-0,11	-1,05	0,27
Idade	0,59	-0,04	0,98
Gênero_1	-0,23	-0,94	1,06
Local do câncer_2	0,58	-1,27	2,03
Local do câncer_3	0,49	-1,98	2,66
Local do câncer_4	0,18	-5,82	1,77
Local do câncer_6	-0,86	-2,54	1,23
$\text{Local do câncer}_{\text{RARO}}$	0,29	-3,53	2,79
Raça_6	0,56	-5,48	1,67
$\text{Maior grau de instrução}_2$	-0,58	-2,24	1,72
$\text{Maior grau de instrução}_3$	-0,08	-1,61	2,44
$\text{Maior grau de instrução}_4$	-0,14	-2,63	2,00
$\text{Maior grau de instrução}_5$	-1,17	-3,40	1,46
$\text{Maior grau de instrução}_6$	0,10	-4,16	6,88
Estado civil_2	0,29	-23,25	3,16
Estado civil_3	0,51	-2,19	2,08
Estado civil_4	0,17	-1,45	1,16
Estado civil_5	-0,05	-1,13	2,36
Estado civil_6	-1,15	-5,45	3,10
$\text{Estadiamento (T)}_2$	1,49	-0,55	4,78
$\text{Estadiamento (T)}_3$	1,44	-0,38	5,45
$\text{Estadiamento (T)}_4$	2,91	1,89	7,07
$\text{Estadiamento (T)}_5$	1,90	-0,57	7,63
$\text{Estadiamento (N)}_1$	0,48	-1,90	2,46
$\text{Estadiamento (N)}_2$	0,78	-5,83	3,02
$\text{Estadiamento (N)}_3$	1,81	-2,26	5,15
$\text{Estadiamento (N)}_4$	0,73	-2,84	3,19
Escore ACE-27_1	0,67	-0,75	1,62
Escore ACE-27_2	1,31	-1,48	2,07
Escore ACE-27_3	2,74	-29,78	3,70
Tratamento_1	0,70	-0,43	2,98
Tratamento_2	1,56	-1,36	3,08
Tratamento_3	0,63	-2,54	1,67
Tratamento_4	1,27	-0,16	2,83
Tratamento_5	33,71	30,75	63,85

Parâmetro	Estimativa	2,5%	97,5%
Tratamento_6	1,71	-0,54	3,64
HPV_1	-0,32	-2,72	0,63
HPV_2	0,30	-1,95	0,94
Câncer prévio_2	-0,01	-0,98	1,88
Câncer prévio_3	0,26	-5,63	2,51
Tabagismo_1	0,46	-1,24	1,85
Tabagismo_2	-0,07	-1,51	1,15
$\text{Consumo de bebida alcóolica}_1$	0,54	-1,87	2,14
$\text{Consumo de bebida alcóolica}_2$	1,01	-1,35	2,72

Tabela 19: Estimativas para os coeficientes do modelo referente à latência (sobrevivência) e intervalos de confiança

Parâmetro	Estimativa	2,5%	97,5%
IMC	-0,16	-0,29	0,04
Idade	0,16	-0,13	0,24
Gênero_1	-0,07	-0,45	0,23
Local do câncer_2	0,41	-0,02	0,78
Local do câncer_3	-0,22	-0,91	0,21
Local do câncer_4	0,36	-0,23	0,93
Local do câncer_6	0,28	-0,31	1,19
$\text{Local do câncer}_{\text{RARO}}$	0,55	-0,36	1,13
Raça_6	0,13	-0,29	0,58
$\text{Maior grau de instrução}_2$	0,27	-0,07	0,86
$\text{Maior grau de instrução}_3$	0,02	-0,50	0,42
$\text{Maior grau de instrução}_4$	0,03	-0,54	0,65
$\text{Maior grau de instrução}_5$	0,44	$1{,}21\times10^{3}$	1,32
$\text{Maior grau de instrução}_6$	0,00	-0,51	0,76
Estado civil_2	0,28	-0,80	1,18
Estado civil_3	-0,12	-0,53	0,24
Estado civil_4	0,02	-0,31	0,32
Estado civil_5	0,56	-0,05	0,96
Estado civil_6	0,09	-1,63	1,20
$\text{Estadiamento (T)}_2$	-0,96	-3,06	-0,71
$\text{Estadiamento (T)}_3$	-0,57	-2,83	-0,30
$\text{Estadiamento (T)}_4$	-0,86	-3,36	-0,64
$\text{Estadiamento (T)}_5$	-0,84	-3,38	-0,40
$\text{Estadiamento (N)}_1$	-0,67	-1,69	0,01

Parâmetro	Estimativa	2,5%	97,5%
Estadiamento $(N)_2$	-0,41	-1,18	0,15
Estadiamento $(N)_3$	-0,17	-1,32	0,17
Estadiamento $(N)_4$	-0,66	-1,65	-0,10
Escore ACE-27_1	-0,22	-0,75	0,10
Escore ACE-27_2	0,04	-0,49	0,34
Escore ACE-27_3	-0,06	-0,76	0,23
Tratamento_1	-0,31	-1,28	-0,04
Tratamento_2	0,01	-0,73	0,33
Tratamento_3	0,07	-0,58	0,67
Tratamento_4	0,09	-0,69	0,36
Tratamento_5	1,16	0,16	1,41
Tratamento_6	1,64	0,89	1,96
HPV_1	-0,29	-0,56	0,38
HPV_2	0,06	-0,20	0,40
Câncer prévio_2	0,10	-0,22	0,47
Câncer prévio_3	0,11	-0,47	0,58
Tabagismo_1	0,48	0,03	0,99
Tabagismo_2	0,54	0,25	1,17
Consumo de bebida alcóolica_1	-0,33	-0,97	0,21
Consumo de bebida alcóolica_2	-0,49	-1,26	0,02

Para o modelo dos incidentes, referente à probabilidade de cura, as variáveis estatisticamente significativas são aquelas para os quais os intervalos de confiança não contèm o zero. Primeiramente, a variável estadiamento (T) mostrou associação significativa apenas para o estadiamento T4, indicando que indivíduos nesse grupo apresentam menor probabilidade de cura comparado à categoria de referência (estadiamento T1), uma vez que o sinal da estimativa foi positivo. As demais categorias de T não apresentaram diferenças estatisticamente significativas em relação à referência. Além disso, para a variável tipo de tratamento recebido, observou-se uma associação extremamente significativa e negativa com a cura para o grupo que recebeu apenas quimioterapia (tratamento 5) em comparação com a categoria de referência, que corresponde ao grupo que recebeu apenas cirurgia (tratamento 0). Para a segunda parte do modelo, ou seja, o modelo da latência, apesar de parecer contraintuitivo, o estadiamento T apresentou associação significativa com a latência para todas as categorias T2 a T5. As estimativas negativas das categorias de estadiamento (T), indicam que esses indivíduos apresentam maior tempo de sobrevida em relação à categoria de referência (T1), mesmo sem cura. Esse achado pode sugerir que pacientes com doença mais avançada, mas não curados, ainda assim têm sobrevida prolongada, possivelmente em função de tratamentos mais intensivos. Com relação à

variável tipo de tratamento, observou-se que a categoria referente ao tratamento paliativo ou desconhecio (tratamento 45) e a modalidade quimioterapia isolada (tratamento 30) estão associadas a menor tempo de sobrevida, enquanto a modalidade de tratamento 11 (cirurgia associada à radioterapia) apresentou efeito protetor, com aumento da latência. Ainda, o tabagismo foi associado a menor tempo até o evento entre os não curados, com significância estatística para ambas as categorias de fumantes em comparação aos não fumantes. Por fim, apenas uma categoria da variável escolaridade (categoria 5), que corresponde a indivíduos com pós-graduação ou mais, apresentou associação significativa com menor latência, embora o intervalo de confiança marginalmente inclua o zero, indicando uma possível relação entre menor sobrevida e esse nível de escolaridade.

Em seguida, empregamos o método de penalização do desvio absoluto suavemente recortado (*SCAD*) para a seleção de variáveis com o intuito de identificar covariáveis que expliquem as partes da incidência e latência. Após o ajuste de todos os modelos possíveis, o modelo com o menor valor do critério de informação bayesiano (*BIC*, da sigla em inglês) é retornado, juntamente com as covariáveis selecionadas para as duas partes do modelo. Os resultados são apresentados abaixo:

Tabela 20: Estimativas finais para os coeficientes do modelo escolhido a partir do BIC

Probabilidade de cura (incidência)			
Parâmetro	Estimativa	IC 2,5%	IC 97,5%
Intercepto	-3,01	-3,96	-1,76
IMC	-0,26	-0,47	-0,02
Idade	0,62	0,34	0,78
Estadiamento T_2	0,74	-0,18	1,33
Estadiamento T_3	1,12	0,25	1,99
Estadiamento T_4	2,25	1,01	3,04
Estadiamento T_5	1,19	0,31	1,95
Estadiamento N_3	1,49	0,26	2,46
Escore ACE-27_2	1,24	0,64	1,66
Escore ACE-27_3	2,18	0,61	2,99
Tratamento_2	1,48	0,51	2,22
Tratamento_4	1,02	0,29	1,65
Tratamento_5	4,92	-24,75	7,11
Tratamento_6	1,71	0,88	2,52
HPV_1	-0,55	-1,15	0,04
Tabagismo_1	1,10	0,56	1,49
Tabagismo_2	0,55	0,03	0,93
Comportamento do tempo de falha (latência)			

Parâmetro	Estimativa	IC 2,5%	IC 97,5%
Local do câncer$_2$	0,56	0,22	0,83
Tratamento$_5$	1,03	0,60	1,42
Tratamento$_6$	1,53	1,00	1,92

O critério de informação bayesiano é minimizado para $\lambda_1 = 0,04$. Para o modelo retornado, as covariáveis selecionadas para a parte da incidência são IMC, idade, estadiamento (T), estadiamento (N) (apenas o estádio N3), escore de comorbidades ACE-27 (apenas categorias leve e moderado), tipo de tratamento (somente as categorias cirurgia aliada à químio-radioterapia, químio-radioterapia, quimioterapia isolada e tratamento desconhecido/paliativo), status positivo de infecção por HPV, e tabagismo (categorias atual ou ex-fumante). No modelo dos incidentes, os sinais negativos dos coeficientes para as variáveis IMC e status positivo de infecção por HPV indicam que indivíduos com índices de massa corporal menores e/ou que testaram positivo para HPV possuem maiores probabilidades de cura. Já para o modelo da latência, pacientes suscetíveis cujo o local do câncer foi a cavidade oral (local 2) e/ou que receberam como tipo de tratamento a quimioterapia isolada ou tratamento paliativo/desconhecido apresentaram maiores probabilidades de falha.

Adicionalmente, podemos selecionar o modelo que retorne o menor valor para o critério de informação de Akaike (*AIC*, da sigla em inglês). Os resultados são apresentados na Tabela 19.

Tabela 21: Estimativas finais para os coeficientes do modelo escolhido a partir do AIC, com intervalos de confiança de 95%

Probabilidade de cura (incidência)			
Parâmetro	Estimativa	IC 2,5%	IC 97,5%
Intercepto	-3,42	-6,23	-1,85
Idade	0,61	0,21	0,88
Local do câncer$_2$	0,46	-0,71	1,24
Local do câncer$_3$	0,32	-1,26	1,43
Estadiamento T$_2$	0,81	-0,74	1,69
Estadiamento T$_3$	1,04	-1,06	1,98
Estadiamento T$_4$	2,11	0,39	3,03
Estadiamento T$_5$	1,22	-0,65	2,29
Estadiamento N$_1$	0,10	-1,93	1,70
Estadiamento N$_3$	1,49	-0,36	3,88
Escore ACE-27$_1$	0,73	-0,24	1,30

Parâmetro	Estimativa	IC 2,5%	IC 97,5%
Escore ACE-27_2	1,20	-0,33	1,71
Escore ACE-27_3	2,45	-9,84	3,32
Tratamento_2	1,55	0,41	2,86
Tratamento_4	0,84	-0,02	2,40
Tratamento_5	33,44	32,40	63,82
Tratamento_6	1,58	0,61	3,44
Tabagismo_1	1,11	0,39	2,09
Comportamento do tempo de falha (latência)			
Parâmetro	Estimativa	IC 2,5%	IC 97,5%
IMC	-0,18	-0,29	-0,04
Idade	0,14	-0,11	0,28
Gênero_1	-0,11	-0,36	0,15
Local do câncer_2	0,44	0,07	0,82
Local do câncer_3	-0,21	-0,68	0,22
Local do câncer_4	0,19	-0,57	0,64
Local do câncer_{RARO}	0,54	-0,39	1,21
Raça_6	0,29	-0,03	0,65
Maior grau de instrução_2	0,10	-0,26	0,48
Estado civil_2	0,46	-0,31	1,32
Estado civil_3	-0,12	-0,48	0,19
Estado civil_5	0,49	0,03	0,86
Estadiamento N_3	0,08	-0,84	0,72
Estadiamento N_4	-0,54	-1,35	0,02
Escore ACE-27_1	-0,24	-0,66	0,18
Escore ACE-27_2	0,06	-0,31	0,50
Escore ACE-27_3	-0,05	-0,53	0,36
Tratamento_3	0,27	-0,44	0,94
Tratamento_4	0,24	-0,45	0,72
Tratamento_5	1,27	0,46	1,73
Tratamento_6	1,71	0,92	2,08
HPV_1	-0,38	-0,69	0,14
HPV_2	0,13	-0,09	0,40
Tabagismo_1	0.21	-0.42	0.57
Tabagismo_2	0.47	-0.04	0.92
Consumo de $\text{bebida alcóolica}_2$	-0,22	-0,63	0,23

O critério de informação de Akaike é minimizado para $\lambda_1 = 0,02$. Como esperado,

o modelo selecionado é menos penalizado e possui mais covariáveis explicativas.

Enquanto o BIC tende a selecionar modelos mais parcimoniosos, penalizando fortemente o número de parâmetros, o AIC prioriza a qualidade do ajuste, resultando em modelos com mais covariáveis. A escolha entre eles depende do objetivo do estudo: se a interpretabilidade e a simplicidade são prioritárias, o BIC é mais adequado; se o foco é a precisão preditiva, o AIC pode ser preferível. Como o objetivo principal deste trabalho é identificar os principais fatores associados ao tempo até a falha e à probabilidade de cura, priorizou-se o modelo selecionado pelo BIC, por ser mais parcimonioso e favorecer modelos com maior evidência de ajuste, penalizando a complexidade desnecessária.

Por fim, utilizamos o modelo escolhido para obter a proporção de curados na amostra por meio do cálculo da média do vetor das probabilidades de cura de cada um dos indivíduos. O valor obtido foi 56,77%, que é a proporção estimada de indivíduos não-suscetíveis.

6 Conclusão

Não é incomum, no âmbito da análise de sobrevivência, encontrarmos situações em que parte da população não é suscetível ao evento de interesse, e portanto o uso de modelos com fração de cura é mais adequado. Neste trabalho, utilizamos o pacote penPHcure, do R, para aplicar o modelo semiparamétrico com fração de cura de Sy e Taylor (2000) a um estudo de pacientes com câncer de cabeça e pescoço.

De início, a análise descritiva das covariáveis de interesse possibilitou a identificação de fatores que influenciam diretamente na probabilidade de cura dos pacientes em estudo.

No que diz respeito aos fatores sócio-econômicos, as variáveis raça, maior grau de instrução e estado civil todas apresentaram diferenças significativas nas curvas de sobrevivências de suas respectivas categorias, o que nos permitiu traçar um perfil de grupos que possam estar associados negativamente a uma menor probabilidade de sobrevivência. A saber, os grupos de pacientes não-brancos, com grau de instrução inferior a ensino médio completo e viúvos.

Adicionalmente, a análise de fatores clínicos permitiu a identificação de covariáveis cujos grupos apresentaram diferenças significativas em suas curvas de sobrevivência, especificamente as variáveis presença de câncer prévio que não o de cabeça e pescoço, escore geral de comorbidades ACE-27, tabagismo, consumo de bebida alcóolica, local do câncer, tipo do tratamento recebido e status de infecção por HPV. Analogamente à análise dos fatores sócio-econômicos, identificamos aqui os grupos com menor probabilidade de sobrevivência, de acordo com os gráficos de Kaplan-Meier obtidos: pacientes com histórico prévio de câncer que não o de cabeça e pescoço, indivíduos cujo escore de comorbidade ACE-27 foi severo, fumantes, pacientes cujo local do câncer foi a hipofaringe e, por último, indivíduos que receberam apenas a quimioterapia ou tratamento paliativo. Curiosamente, indivíduos cujo status de infecção por HPV foi positivo apresentaram maior probabilidade de sobrevivência.

Foi feito o ajuste de um modelo para cada uma das combinações de covariáveis disponíveis, juntamente com o cálculo dos critérios de informação bayesiano e de Akaike. Em seguida, os dois modelos que apresentaram os menores valores para cada um dos índices foram selecionados. Por fim, optou-se utilizar o modelo selecionado com base no BIC, uma vez que este está mais alinhado com os objetivos do trabalho.

Os resultados obtidos são muito relevantes no desenvolvimento e elaboração de protocolos de diagnóstico, além de contribuir para a identificação de tipos de tratamento que podem influenciar na probabilidade de cura de pacientes diagnosticados com o CCECP. Para trabalhos futuros, a inclusão de outras covariáveis de interesse é incenti-

vada, como a renda e marcadores de metilação, por exemplo.

Referências

BARSOUK, A. et al. Epidemiology, risk factors, and prevention of head and neck squamous cell carcinoma. *Med Sci (Basel)*, PubMed Central, v. 11, n. 2, p. 42, 2023.

BERETTA, A.; HEUCHENNE, C. Variable selection in proportional hazards cure model with time-varying covariates, application to us bank failuresfootnote. *Journal of Applied Statistics*, v. 46, n. 9, p. 1529–1549, 2019.

BERETTA, A.; HEUCHENNE, C. penphcure: Variable selection in proportional hazards cure model with time-varying covariates. *The R Journal*, v. 13, n. 1, p. 116–129, 2021.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, American Statistical Association, v. 47, n. 259, p. 501–515, 1952.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, Wiley for the Royal Statistical Society, v. 11, n. 1, p. 15–53, 1949.

CAI, C. et al. smcure: An r-package for estimating semiparametric mixture cure models. *Comput Methods Programs Biomed*, PubMed Central, v. 108, n. 3, p. 1255–1260, 2012.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Blucher, 2006.

CORBIÈRE, F.; JOLY, P. A sas macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 85, n. 2, p. 173–180, 2007.

COX, D. R.; OAKES, D. *Analysis of Survival Data*. [S.l.]: CRC Press, 1984.

DEMPSTER, A. et al. Maximum likelihood from incomplete data via the em algorithm. *Royal Statistical Society*, v. 39, n. 1, p. 1–22, 1977.

FAN, J.; LI, R. Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, v. 30, n. 1, p. 74–99, 2002.

FANG H.-B., L. G.; SUN, J. Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, Wiley, v. 32, p. 59–75, 2005.

GALATI, L. et al. Hpv and head and neck cancers: Towards early diagnosis and prevention. *Tumour Virus Res*, PubMed Central, v. 14, n. 200245, 2022.

KLEIN, J. P. et al. *Handbook of Survival Analysis*. [S.l.]: CRC Press, 2014.

SABATINI, M. E.; CHIOCCA, S. Human papillomavirus as a driver of head and neck cancers. *British Journal of Cancer*, Springer Nature, v. 122, p. 306–314, 2020.

SOBIN, L.; WITTEKIND, C. (Ed.). *TNM Classification of Malignant Tumours*. [S.l.]: John Wiley Sons, 2002.

SY, J.; TAYLOR, J. Estimation in a cox proportional hazards cure model. *Biometrics*, v. 56, n. 1, p. 227–236, 2000.

VIRANI, S. et al. Ndn and cd1a are novel prognostic methylation markers in patients with head and neck squamous carcinoma. *BMC Cancer*, BioMed Central, v. 15, n. 825, 2015.