UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Departamento de Sistemas de Computação

Análise de sobrevida aplicada a dados clínicos de câncer de mama

Rafael Pastre

São Carlos - SP

Análise de sobrevida aplicada a dados clínicos de câncer de mama

Rafael Pastre

Orientadora: Mariana Curi

Monografia referente ao projeto de conclusão de curso dentro do escopo da disciplina SSC0670 – Projeto de Formatura I do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC-USP para obtenção do título de Engenheiro de Computação.

Área de Concentração: Estatística

USP – São Carlos
23 de novembro de 2021

Agradecimentos

Agradeço a minha orientadora Prof. Dra. Mariana Curi, ao orientador do Hospital A.C. Camargo Prof. Dr. Israel Tojal da Silva, e ao aluno de mestrado Khennedy Bacule dos Santos, por todos os ensinamentos aprendidos e por todo o apoio na confecção deste trabalho.

Agradeço a Deus, que me deu forças para não desistir dos meus objetivos.

Agradeço a todos da minha família que de uma forma ou outra sempre estiveram do meu lado acreditando no meu potencial e dando sempre a força que eu precisava.

Aos meus pais e amigos que estiveram sempre ao meu lado, me apoiando, motivando, vendo meu esforço para chegar até essa etapa de minha vida.

Resumo

O câncer de mama é uma doença altamente relevante no Brasil e no Mundo, cujas causas podem ser variadas, assim como suas características clínicas, genéticas, e as relações da doença com o organismo. Por este motivo, o presente trabalho, feito em parceria com o hospital de câncer A.C. Camargo se dedica a realizar uma análise de sobrevida nos dados de câncer de mama da base *The Cancer Genome Atlas - Breast Invasive Carcinoma* (TCGA-BRCA) disponibilizadas pelo hospital. A análise realizada se consiste no tratamento inicial da base de dados a fim de proporcionar consistência aos dados, seguido de uma breve análise descritiva, da estimação da curva de sobrevida geral e por categorias através do estimador de Kaplan-Meier, e finalmente, pela aplicação do Modelo de Riscos Proporcionais de Cox, com uma e várias variáveis, bem como a tentativa de ajuste deste modelo e interpretação dos riscos relativos. Ao fim desse estudo, pudemos perceber que, apesar de realizadas todas as análises, para o estudo completo destes dados seria necessário a utilização de modelos de Cox mais robustos, visto que a condição de riscos proporcionais não foi garantida, o que fica proposto para futuros trabalhos.

Índice

CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização e Motivação

O câncer de mama é uma das doenças mais incidentes em mulheres na faixa etária de 40 a 59 anos, com múltiplos fatores de risco associados: fatores genéticos, ambientais e comportamentais, caracterizando-se pela proliferação desordenada e em constante crescimento das células deste órgão.

Este tipo de câncer representa um grave problema de saúde pública no Brasil e no mundo, dada a sua alta incidência, morbidade/mortalidade, como também pelo alto custo no tratamento, seguimento e reabilitação. Estimativas apontam que em 2020, serão cerca 15 milhões de novos casos podendo atingir 12 milhões de mortes.

Quando se refere especificamente ao câncer de mama, tem-se que ele corresponde a 22% dos tipos de câncer detectado anualmente e que possui uma maior incidência em mulheres do que em homens, sendo ele considerado o tipo de câncer mais frequente no mundo e o primeiro que leva ao óbito indivíduos do sexo feminino.

No Brasil, o Instituto Nacional de Câncer (INCA) relata que no ano de 2014 ocorreram 57.124 mil novos casos de câncer de mama, uma média de 57 casos a cada 100.000 mulheres. Quanto ao número de óbitos neste mesmo ano foi de 13.345 ocorrências.

Em um paciente que possuí câncer, o crescimento acelerado das células tende a ser incontrolável e agressivo ao tecido ou órgão atingido, fazendo com que haja um acúmulo de células cancerígenas (tumores) que são considerados neoplasias malignas. Em contraponto, temos os tumores benignos que são massas que se assemelham ao tecido original e se proliferam mais vagarosamente (INCA, 2015a).

O câncer tem causa variada, podendo ser elas externas ou internas ao organismo e ambas estão interligadas. As causas externas têm relação com o meio ambiente e com os costumes ou hábitos adotados pelos indivíduos, já as causas internas estão quase sempre relacionadas ao quesito genético. Todavia 80 a 90% dos cânceres estão ligados ao fator ambiental, que atua alterando a estrutura do DNA das células. Com isso, pode-se concluir

que o surgimento do câncer vai depender da duração e da intensidade da exposição do organismo aos agentes causadores.

Dessa forma, a alta relevância do tratamento do câncer no Brasil e no Mundo, assim como a alta variabilidade de causas, clínicas ou genéticas, e as relações da doença com o com meio em que está inserido, este estudo é motivado a analisar como estas diferentes variáveis afetam a sobrevida de um paciente com câncer de mama.

1.2. Objetivos

Este trabalho tem como principal objetivo o aprendizado e desenvolvimento de uma Análise de Sobrevivência, que será aplicada em dados de câncer de mama, através da confecção de curvas de Kaplan-Meier e do ajuste de Modelos de Cox, além de tratar e explorar os dados clínicos da base TGCA-BRCA. O seu desenvolvimento foi feito em parceria com o hospital de câncer A.C. Camargo, que disponibilizou os dados para a realização da análise, em um grupo de estudos formado pela orientadora prof. Dr. Mariana Curi, pelo orientador do A.C. Camargo (Fundação Antônio Prudente) prof. Dr. Israel Tojal da Silva, pelo aluno de mestrado Khennedy Bacule dos Santos e pelo aluno de graduação Rafael Pastre.

1.3. Organização do Trabalho

Este trabalho se consiste da realização de uma Análise de Sobrevivência aplicada em dados de câncer de mama. Para isso, o trabalho está estruturado de forma que no Capítulo 2 serão apresentados os conceitos fundamentais e técnicas que serão utilizadas, assim como serão apresentadas detalhes e considerações destas. No capítulo 3 será explicado detalhadamente todo o processo de realização da análise, além de serem apresentados os resultados e dificuldades encontradas no processo. No capitulo 4 será feita a conclusão do trabalho, através da revisão das contribuições realizadas, do relacionamento do trabalho com o curso, e da proposição de etapas para trabalhos futuros.

CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

2.1. Considerações Iniciais

Neste capitulo será feita a apresentação dos conceitos necessários para o entendimento e realização de uma Análise de Sobrevida. Inicialmente, serão abordados os conceitos de falha, censura, tempo de falha, e tempo de censura, assim como será definido o evento de interesse e o tempo de referência para o início do estudo. Em seguida, serão apresentadas a função de sobrevida e a função risco, e, finalmente, serão apresentadas as técnicas de Kaplan-Meier e do Modelo de Cox de Riscos Proporcionais, assim como as considerações e nuances destas técnicas.

2.2. Conceitos e Técnicas Relevantes

2.2.1. Conceitos fundamentais

A Análise de Sobrevida é amplamente aplicada na área médica, pois ela permite que o tempo de ocorrência do evento de interesse seja considerado na análise, ou seja, a variação da medida de interesse ao longo do tempo é considerada no estudo (COLOSIMO, 2006; ROSNER, 2011; SHREFFLER, 2021). Sendo assim, antes de iniciar a análise, é necessário definir os conceitos fundamentais que serão necessários para desenvolvê-la.

O primeiro conceito é o **tempo de falha**, que é o tempo a partir do início do estudo e até a constatação do evento de interesse, que, neste caso, é a morte do paciente causada devido ao câncer de mama (COLOSIMO, 2006). Vale acrescentar ainda, que, neste trabalho, o tempo de início do estudo ($t = 0$) se dá no momento do diagnóstico da doença.

Outro conceito importante é a **censura**, que se consiste da falta de informação completa sobre o tempo de falha. Existem diferentes tipos de censura, entretanto este estudo se restringirá a censura à direita, que se caracteriza pela não observação do evento de interesse, que pode ser causada, por exemplo, pela morte do paciente por outros motivos além do evento de interesse, pelo abandono do estudo pelo paciente, e pelo término do estudo (COLOSIMO, 2006; ROSNER, 2011).

Em caso de censura, como não observamos o tempo de falha, na prática o que pode ser observado é o tempo em que se constatou pela ultima vez que o evento de interesse não

ocorreu, e, este tempo é chamado de **tempo de censura**. No caso deste estudo, este tempo também corresponde ao último tempo de acompanhamento do paciente (COLOSIMO, 2006).

Definidos estes conceitos, devemos então apresentar a **função de sobrevivência**, que é a principal função utilizada nos estudos de sobrevivência, e é definida através de uma função probabilística que indica a probabilidade do indivíduo sobreviver pelo menos até um tempo *t*, ou seja, indica a probabilidade do evento de interesse ocorrer após *t*. Dessa forma, considerando que o tempo de falha é modelado pela variável aleatória *T*, com $T \geq 0$, definimos a função de sobrevivência (COLOSIMO, 2006; ROSNER, 2011).

$$S(t) = P(T \geq t)$$

Vale notar também que existem diversos modelos para a análise de sobrevida, que podem considerar taxas de reincidência, de cura, e múltiplos eventos de interesse, entretanto, por simplicidade, o estudo feito nesse trabalho não levará em conta nenhum destes fatores, portanto, no modelo aqui considerado temos também que no limite quanto o tempo vai para infinito, temos $S(t) = 0$.

Após definida a função de sobrevivência, podemos representar a probabilidade de falha, ou risco, em um intervalo $[t_1, t_2)$ como sendo $P(t_1 \leq T < t_2) = S(t_1) - S(t_2)$. Dessa forma, podemos definir ainda a **taxa de risco** em $[t_1, t_2)$, que se consiste na probabilidade de um evento ocorrer neste intervalo, dado que o indivíduo sobreviveu até t_1 (COLOSIMO, 2006; ROSNER, 2011). A taxa de risco é dada pela expressão:

$$\frac{P(t_1 \leq T < t_2 | T \geq t_1)}{(t_2 - t_1)} = \frac{S(t_1) - S(t_2)}{(t_2 - t_1).S(t_1)}$$

Definimos então a **função de risco** $\lambda(t)$, que se consiste da taxa de risco instantânea em um tempo *t* dado que o indivíduo sobreviveu até aquele momento, através da seguinte expressão (COLOSIMO, 2006).

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

Dessa definição, segue também a seguinte relação entre a função risco e a função de sobrevivência:

$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}(\ln S(t))$$

2.2.2. Estimador de Kaplan-Meier

Para realizar uma observação prática da função de sobrevivência é necessário utilizar um estimador para esta função, por exemplo, o estimador de Nelson-Aalen, o estimador da tabela de vida, ou o estimador de Kaplan-Meier. Este último é o mais utilizado em estudos clínicos, portanto ele foi escolhido para ser utilizado neste trabalho (COLOSIMO, 2006).

O **estimador de Kaplan-Meier** para a função de sobrevivência é um estimador não paramétrico, e sua construção é feita, primeiramente, através da divisão do tempo em intervalos, através dos valores $t_1 < t_2 < ... < t_k$ de todos os instantes em que houve ocorrência de falha nos dados observados. Em seguida, utilizaremos a ideia de que para um indivíduo sobreviver após um intervalo, é necessário que ele tenha sobrevivido a todos os anteriores, e então, dado que ele sobreviveu aos anteriores, que ele sobreviva ao período mais recente, ou seja

$$S(t) = P(T \geq t) = P(T \geq t | T \geq t_k).P(T \geq t_k)$$

$$S(t) = P(T \geq t | T \geq t_k)P(T \geq t_k | T \geq t_{k-1}) \dots P(T \geq t_1 | T \geq 0)P(T \geq 0)$$

Dessa forma, sendo d_j o número de eventos (falhas) que ocorreram em $[t_{j-1},t_j)$, e n_j o número de pacientes sobreviventes, ou seja, que não sofreram falha nem censura, até t_{j-1}, podemos estimar a função de sobrevivência através do estimador de Kaplan-Meier

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right)$$

Vale ainda notar que, neste estimador, temos que as censuras são consideradas através de $n_{j+1} = n_j - d_j - c_j$, sendo c_j a quantidade de censuras em $[t_{j-1},t_j)$.

2.2.3. Modelo de Cox de Riscos Proporcionais

Outro aspecto importante da análise de sobrevida é a modelagem e ajuste da função risco. Para isso, diversos modelos podem ser avaliados, e, neste trabalho, utilizaremos o **Modelo de Cox de Riscos Proporcionais**, que é o modelo mais utilizado na análise de

sobrevida, e se constituí de um modelo semi-paramétrico, onde a função de risco é modelada por:

$$\lambda(t|x) = \lambda_0(t)exp(\beta_1 x_1 + \cdots + \beta_n x_n)$$

onde x representa as características do paciente ou do tratamento, e β os coeficientes de cada respectiva característica (ROSNER, 2011). O modelo de Cox possuí um importante aspecto que é a suposição de riscos proporcionais, ou seja, para o modelo ser aplicável, é necessário que que a razão da função risco entre quaisquer dois grupos se mantenha constante independentemente do tempo. Essa suposição pode ser verificada através da seguinte expressão

$$\frac{\lambda(t|x = x_1)}{\lambda(t|x = x_2)} = \frac{\lambda_0(t)exp(x_1'\beta)}{\lambda_0(t)exp(x_2'\beta)} = exp(x_1'\beta - x_2'\beta) = cte$$

Apesar desta forte suposição de riscos proporcionais, o modelo de Cox apresenta uma vantagem quanto à possibilidade de interpretação dos coeficientes do modelo, visto que é possível relacionar os termos exponenciais com a razão das taxas de riscos. Dessa forma, definindo o **risco relativo** para uma covariável X_k, como sendo o termo e^{β_k}, segue que, para covariáveis categóricas binárias, o risco relativo é a razão em que a taxa de risco varia entre os dois grupos, visto que, quando alteramos apenas a covariável X_k temos

$$\frac{\lambda(t|X_k = 1)}{\lambda(t|X_k = 0)} = e^{\beta_k}$$

Além disso, para variáveis categóricas em geral, ou para variáveis contínuas. Temos que o risco relativo se relaciona com a razão entre as funções risco, da seguinte forma

$$\frac{\lambda(t|X_k = a)}{\lambda(t|X_k = b)} = e^{\beta_k(a-b)}$$

Podemos perceber que a expressão acima também impõe uma restrição do modelo para variáveis não binárias. Para variáveis categóricas, podemos contornar esta restrição através da codificação da variável categórica X_k em n-1 novas variáveis binárias, onde n é a quantidade de categorias de X_k, de forma que é escolhida uma categoria de referência, e então cada uma destas novas variáveis representa as outras categorias que serão comparadas com a categoria de referência. Por exemplo, seja X uma variável categórica

que descreve a cor, com as categorias 0 = “Azul”, 1 = “Verde” e 2 = “Vermelho”. Podemos então evitar a restrição acima, escolhendo a categoria Azul como referência, e realizando a transformação de X nas variáveis Y_1 e Y_2 onde, para Y_1: 0 = “Não Verde” e 1 = “Verde”, e para Y_2: 0 = “Não Vermelho” e 1 = “Vermelho”. Dessa forma, podemos comparar o risco relativo entre as categorias Azul e Verde através de Y_1, e entre as categorias Azul e Vermelho através de Y_2, sem nenhuma restrição, e, por isso, esta transformação é frequentemente utilizada no modelo de Cox para a comparação dos riscos relativos em variáveis categóricas.

2.3. Considerações Finais

Este capítulo introduziu os conceitos fundamentais que serão necessários no decorrer da análise, além de apresentar as técnicas que serão utilizadas, assim como discutiu aspectos relevantes quanto às considerações e restrições que foram feitas, e também ilustrou como é possível interpretar os resultados das técnicas aplicadas.

CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

3.1. Considerações Iniciais

O presente capítulo apresentará os detalhes da realização do projeto, bem como os objetivos, as tecnologias e metodologias utilizadas, além de também apresentar os recursos desenvolvidos no decorrer da analise de sobrevivência, assim como as etapas realizadas para exploração da base, tratamento dos dados, os resultados obtidos, e os próximos passos que serão tomados no projeto.

3.2. Descrição do Problema e/ou Projeto

O presente trabalho foi desenvolvido utilizando a linguagem R, entretanto trabalhos futuros serão desenvolvidos utilizando a linguagem Python a fim de possibilitar o aprendizado da aplicação das técnicas em ambas as linguagens.

O principal objetivo do trabalho é a elaboração e o aprendizado de uma Análise de Sobrevida, considerando os dados de câncer de mama da base TCGA-BRCA. A análise se consiste de uma breve análise exploratória a fim de descrever e tratar os dados, seguida da estimação da curva de sobrevivência através do estimador de Kaplan-Meier, e finalmente, da utilização da Regressão de Cox para o ajuste de um modelo capaz de descrever o comportamento observado da sobrevida dos pacientes.

3.3. Descrição das Atividades Realizadas

3.3.1. Análise exploratória

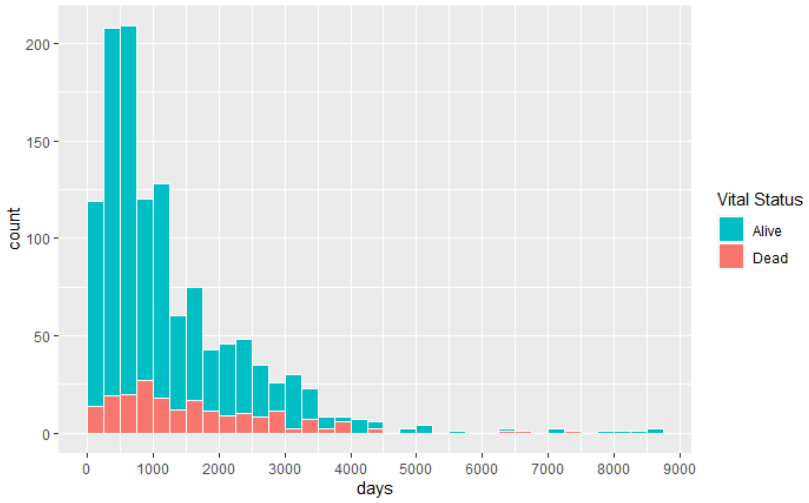
Para realizar uma Análise de Sobrevida, através da confecção da curva de Kaplan-Meier, e da aplicação da regressão de Cox, é fundamental para a estas técnicas que sejam conhecidos o tempo de falha ou censura, o estado vital do paciente nesta observação, que indica a ocorrência ou não do evento de interesse, e também é necessário escolher as covariáveis categóricas ou contínuas utilizadas para comparar a amostra.

Entretanto, observando a base de dados, podemos verificar que os dados do estado vital do paciente e do tempo de falha ou censura não estão prontos para uso. As

informações do estado vital estão contidas em duas colunas da base *vital_status.clinical* e *vital_status.mol*. Sendo assim, após verificar que estas colunas diferem entre si apenas em casos onde uma delas apresenta falta de informação, estas duas colunas foram fundidas em uma nova coluna chamada *vital_status*.

Já para o tempo de falha ou censura, foi observado que, inicialmente, os dados para os pacientes que estavam vivos na última vez em que foram observados (tempo de censura) estavam divididos entre as colunas *days_to_last_follow_up* e *days_to_last_followup*, enquanto os dados para os pacientes cuja ultima observação constatava a ocorrência de óbito (tempo de falha) estavam divididos entre as colunas *days_to_death.clinical* e *days_to_death.mol*. Dessa forma, para construir uma coluna contendo o tempo de falha ou censura, foi realizado um procedimento que se consiste em escolher, em caso de sobrevivência, o maior tempo entre as colunas *days_to_last_follow_up* e *days_to_last_followup*, e, em caso de óbito, o menor tempo entre as colunas *days_to_death.clinical* e *days_to_death.mol*. Após isso, a fim de explorar a distribuição do tempo de acompanhamento, foi construído o gráfico apresentado na Figura 1.

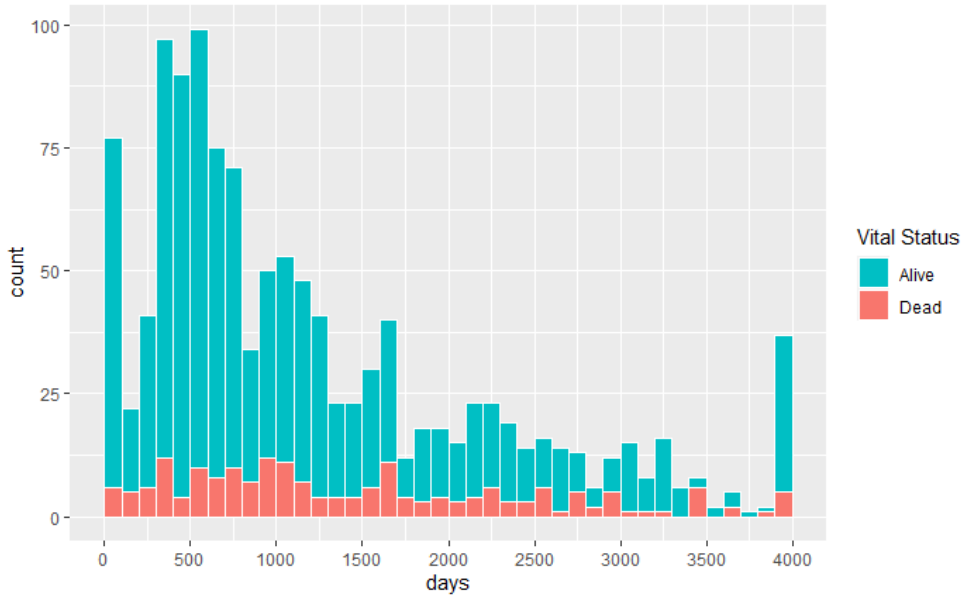
Figura 1 – Tempo de observação dos pacientes



A partir da Figura 1, é possível perceber que, no período após o 4000º dia, a quantidade de observações é bastante reduzida, o que poderia causar o aumento da incerteza da análise, e a perda de significância, caso ela fosse realizada nesse período. Por esta razão, foi determinado que o estudo seria encerrado após os primeiros 4000 dias, e os dados observados após este período, ambos de pacientes sobreviventes ou não, foram

considerados como observações de pacientes sobreviventes no 4000º dia, visto que, a constatação de óbito em data superior ao 4000º dia implica que o paciente estava vivo no dia 4000. Após realizar este corte no tempo de estudo, a distribuição do tempo de observação foi novamente verificada e ilustrada no gráfico da Figura 2.

Figura 2 – Tempo de observação após corte do estudo em 4000 dias



A seguir, é necessário escolher quais características dos pacientes serão utilizadas na análise. Considerando que a base TCGA-BRCA de dados clínicos possui originalmente 154 variáveis descritivas, foram desconsideradas, inicialmente, as variáveis que apenas apresentavam valores nulos, ou que apresentavam apenas valores iguais. Em seguida, com ajuda do Prof. Dr. Israel Tojal da Silva, e considerando que os resultados da Análise de Sobrevida serão interpretados ao final da análise, foram escolhidas analisar as variáveis categóricas que descrevem: O Subtipo do tumor, o subtipo imune, o estadiamento, a patologia, recorrência da doença, recorrência do tratamento, raça, etnia, e gênero. E como variáveis numéricas, foram escolhidas: Idade, Fração de Leucócitos, Fração de *Stromals*, Fração de células alteradas.

Após realizados os ajustes iniciais na base, foi realizada uma breve descrição das variáveis categóricas escolhidas a fim de verificar a distribuição dos pacientes nestas categorias. Os resultados desta análise foram então sintetizados na Tabela 1.

Tabela 1 – Distribuição da amostra entre as variáveis categóricas

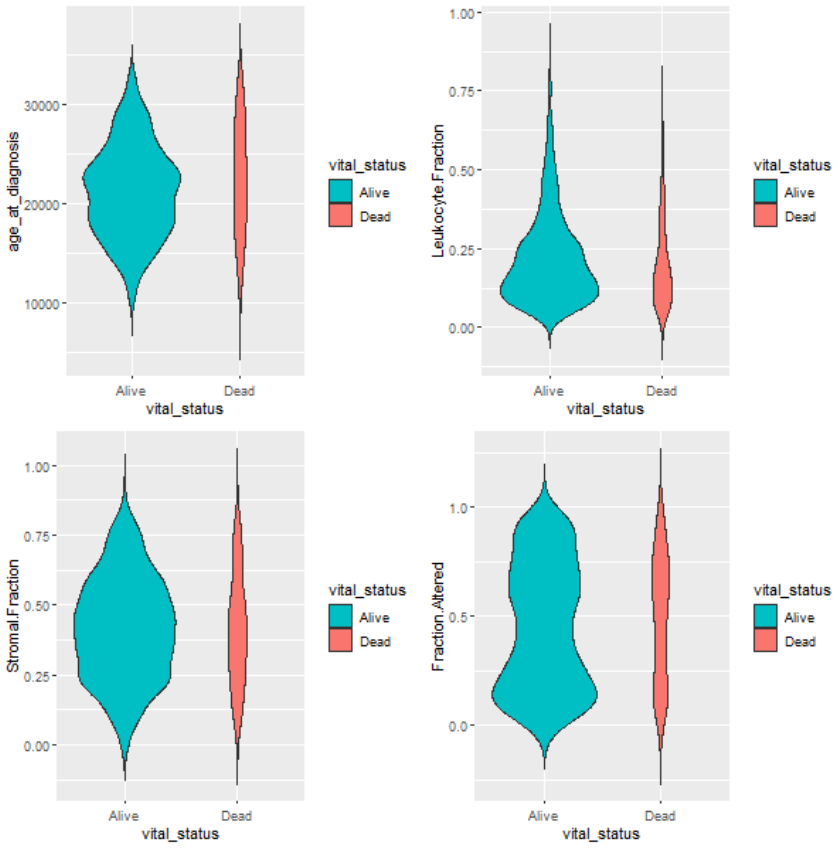
Variável	Categoria	Quantidade	Sobreviventes	Óbitos
Subtipo	Basal	191 (15,72%)	164 (85,86%)	27 (14,14%)
	Her2	82 (6,75%)	66 (80,49%)	16 (19,51%)
	LumA	580 (47,74%)	512 (88,28%)	68 (11,72%)
	LumB	219 (18,02%)	185 (84,47%)	34 (15,53%)
	Normal	143 (11,77%)	95 (66,43%)	48 (33,57%)
Subtipo Imune	C1	405 (33,72%)	344 (84,94%)	61 (15,06%)
	C2	434 (36,14%)	364 (83,87%)	70 (16,13%)
	C3	212 (17,65%)	179 (84,43%)	33 (15,57%)
	C4	104 (8,66%)	82 (78,85%)	22 (21,15%)
	C6	46 (3,83%)	40 (86,96%)	6 (13,04%)
Estadiamento	Stage_I	203 (17,25%)	178 (87,68%)	25 (12,32%)
	Stage_II	681 (57,86%)	594 (87,22%)	87 (12,78%)
	Stage_III	272 (23,11%)	219 (80,51%)	53 (19,49%)
	Stage_IV	21 (1,78%)	5 (23,81%)	16 (76,19%)
Patologia	IDC	593 (60,82%)	488 (82,29%)	105 (17,71%)
	ILC	142 (14,56%)	122 (85,92%)	20 (14,08%)
	Mixed	104 (10,67%)	89 (85,58%)	15 (14,42%)
	Other	136 (13,95%)	106 (77,94%)	30 (22,06%)
Recorrência da doença	no	1142 (93,99%)	958 (83,89%)	184 (16,11%)
	not reported	1 (0,08%)	1 (100%)	0 (0%)
	yes	72 (5,93%)	63 (87,5%)	9 (12,5%)
Recorrência do tratamento	No	1199 (98,68%)	1009 (84,15%)	190 (15,85%)
	Not Reported	2 (0,16%)	2 (100%)	0 (0%)
	Yes	14 (1,15%)	11 (78,57%)	3 (21,43%)
Raça	american indian or alaska native	1 (0,08%)	1 (100%)	0 (0%)
	asian	62 (5,1%)	59 (95,16%)	3 (4,84%)
	black or african american	188 (15,47%)	157 (83,51%)	31 (16,49%)
	not reported	96 (7,9%)	89 (92,71%)	7 (7,29%)
	white	868 (71,44%)	716 (82,49%)	152 (17,51%)
Etnia	hispanic or latino	39 (3,21%)	38 (97,44%)	1 (2,56%)
	not hispanic or latino	976 (80,33%)	795 (81,45%)	181 (18,55%)
	not reported	200 (16,46%)	189 (94,5%)	11 (5,5%)
Gênero	female	1202 (98,93%)	1010 (84,03%)	192 (15,97%)
	male	13 (1,07%)	12 (92,31%)	1 (7,69%)

Ao observar os dados da Tabela 1, é possível verificar que existem categorias cuja quantidade de observações é baixa, o que pode prejudicar a significância dos métodos

aplicados. Além disso, visto que o principal objetivo deste trabalho é a aplicação das técnicas de Análise de Sobrevida, optou-se por realizar um agrupamento nas variáveis categóricas a fim de minimizar esse problema. Dessa forma, foram agrupadas as categorias "Sim" e "Não reportado" das variáveis "Recorrência da doença" e "Recorrência do tratamento", foram agrupadas as categorias "Hispânico ou latino" e "Não reportado" da variável "Etnia", e para a variável "Raça", foram consideradas as categorias "Branco" e "Outros".

Além disso, também foi feita uma breve descrição das variáveis numéricas, a fim de verificar suas distribuições. O gráfico confeccionado para isso pode ser encontrado na Figura 3, e após observá-lo verificou-se que não seria necessário realizar nenhum tipo de tratamento nestas variáveis.

Figura 3 – Distribuições das variáveis numéricas escolhidas

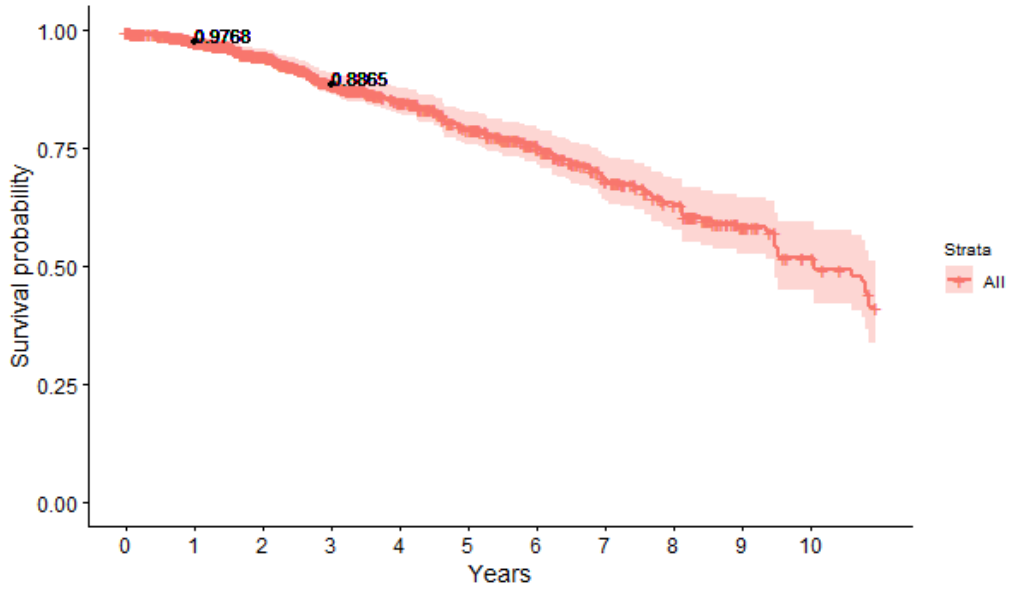


3.3.2. Análise Univariada

Para a realização da Análise de Sobrevida em R, a biblioteca *survival* foi utilizada para estruturar os dados. Dessa forma, foi necessário codificar o estado vital do paciente na forma de uma variável numérica booleana que indica a ocorrência ou não ocorrência do evento de interesse, ou seja, "0" indica que o evento não ocorreu, portanto o paciente está vivo até a última observação, enquanto "1" indica que o evento ocorreu, portanto o paciente veio a óbito. Este tratamento é necessário para todo o desenvolvimento da análise que será feito a seguir.

A primeira técnica da Análise de Sobrevida aplicada foi a confecção das curvas de sobrevivência através do estimador de Kaplan-Meier. Inicialmente foi feita a curva para todos os dados da base, que pode ser encontrada na Figura 4, na qual é possível verificar o comportamento geral da doença e obter as probabilidades de sobrevivência em tempos específicos, como a probabilidade de 97,68% de sobrevivência após 1 ano, e 88,65% após 3 anos, que são tempos relevantes para os estudos clínicos.

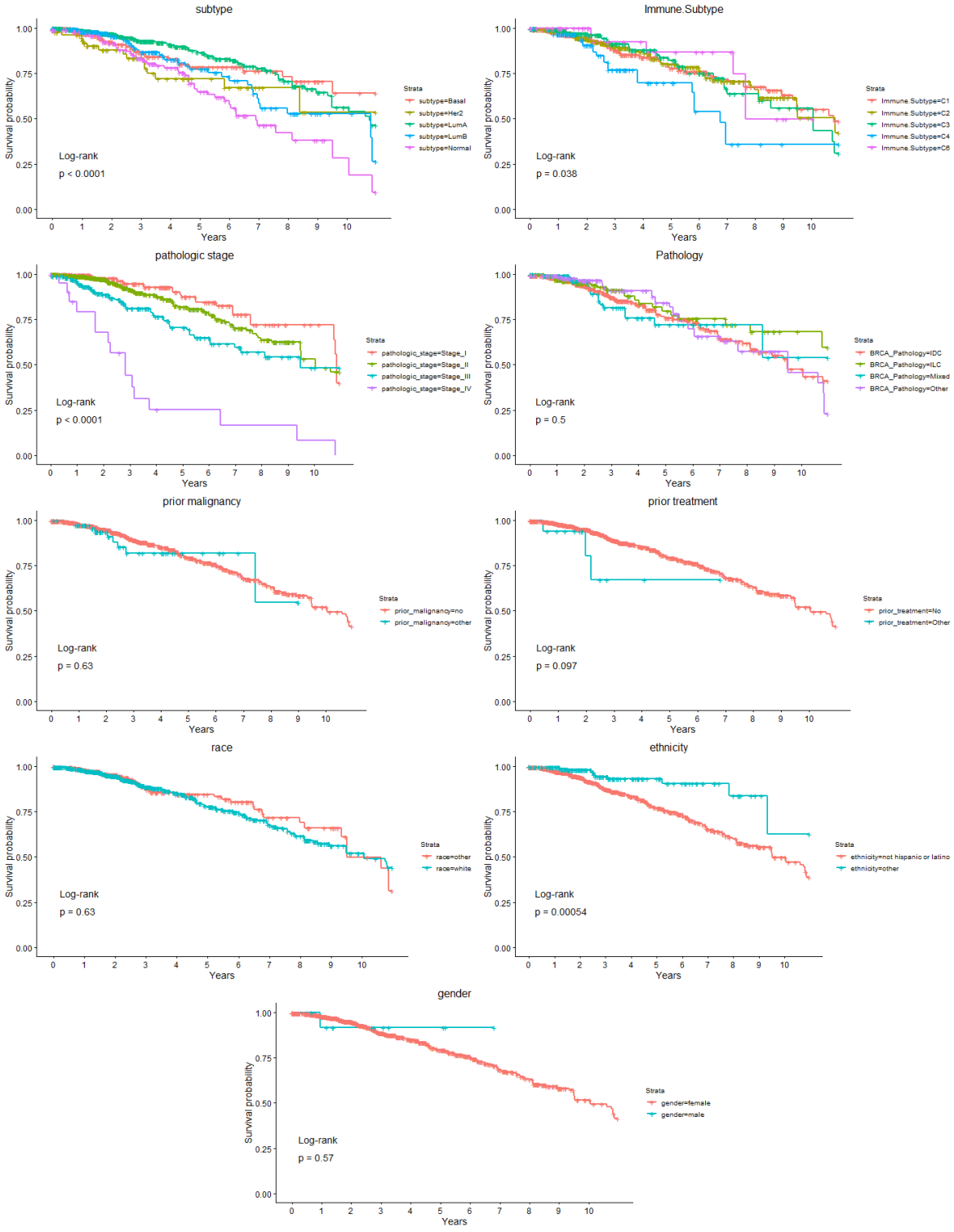
Figura 4 – Curva de Kaplan-Meier para todos os pacientes



Após feita a curva geral de Kaplan-Meier, foi realizado, para as covariáveis categóricas, as curvas de Kaplan-Meier para cada subgrupo, ilustrado na Figura 5, assim como as probabilidades de sobrevivência após 1 e 3 anos, ilustradas na Tabela 2, e finalmente foi realizada a regressão de Cox, considerando os modelos univariados de cada

respectiva variável, cujo resultado está ilustrado na Figura 6, para variáveis categóricas, e na Figura 7 para variáveis contínuas.

Figura 5 – Curvas de Kaplan-Meier e teste Log-Rank por categoria



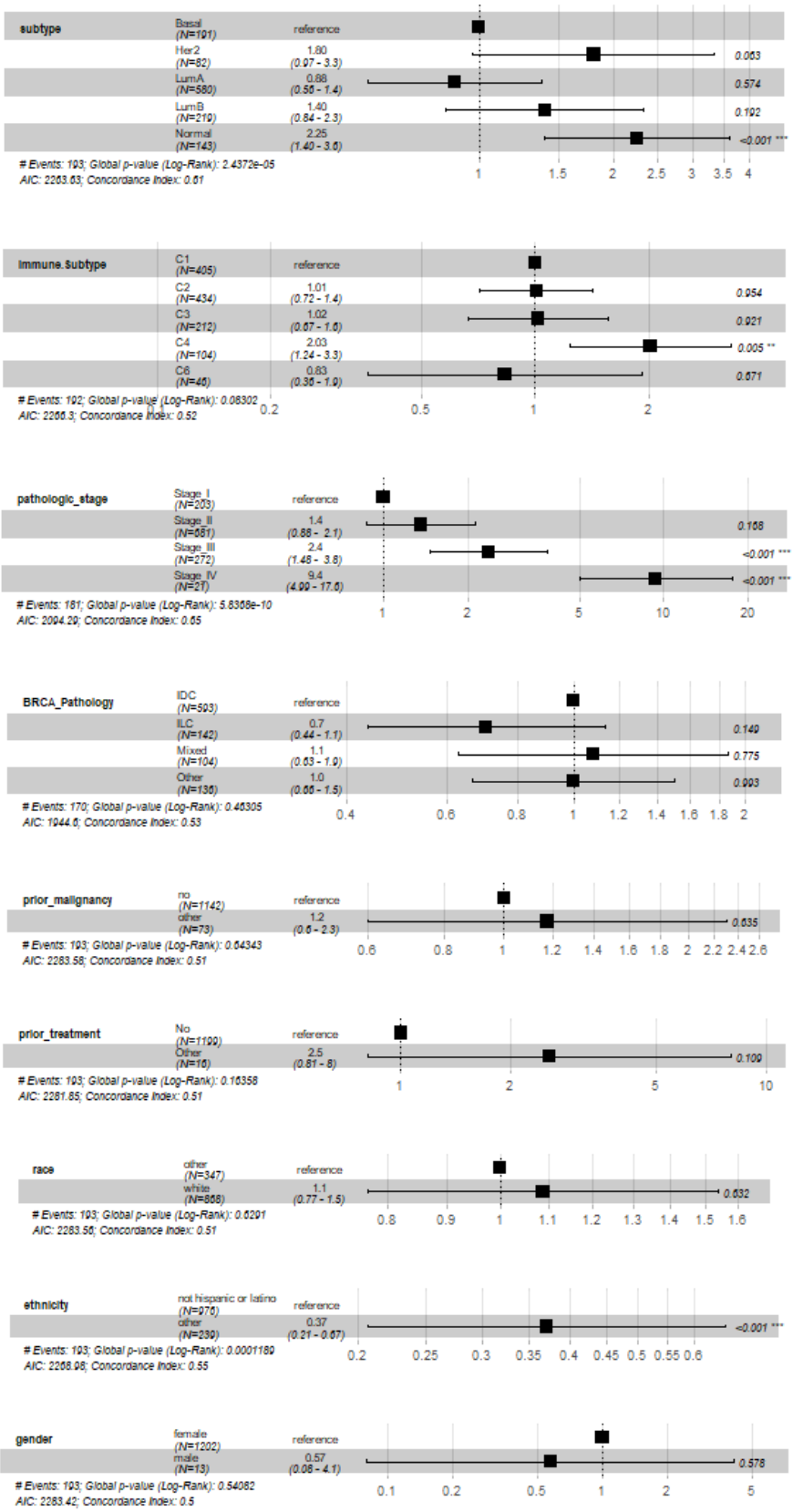
Na Figura 5 podemos verificar de forma visual as diferenças na sobrevida entre os grupos, por exemplo, nas curvas agrupadas por subtipo, subtipo imune, estadiamento, e etinia, podemos observar diferenças significativas entre os grupos.

Tabela 2 – Probabilidade de sobrevivência após 1 e 3 anos

Variável	Categoria	Quantidade	Sobreviventes
Subtipo	Basal	97,03%	85,44%
	Her2	93,24%	83,26%
	LumA	98,71%	92,80%
	LumB	97,93%	86,59%
	Normal	96,36%	83,38%
Subtipo Imune	C1	96,74%	87,92%
	C2	97,76%	89,78%
	C3	99,04%	91,32%
	C4	96,64%	76,81%
	C6	100%	92,31%
Estadiamento	Stage_I	99,47%	94,44%
	Stage_II	98,73%	91,43%
	Stage_III	95,52%	80,97%
	Stage_IV	79,06%	43,92%
Patologia	IDC	97,29%	87,10%
	ILC	96,99%	91,26%
	Mixed	99,01%	81,29%
	Other	99,26%	91%
Recorrência da doença	no	97,71%	89,02%
	other	97,18%	81,91%
Recorrência do tratamento	No	97,74%	88,86%
	Other	93,75%	66,96%
Raça	other	98,06%	88,08%
	white	97,53%	88,74%
Etnia	not hispanic or latino	97,25%	87,40%
	other	99,57%	94,26%
Gênero	female	97,75%	88,58%
	male	91,67%	91,67%

Na Tabela 2, podemos verificar e comparar mais precisamente as diferenças entre grupos no início do estudo, entre 1 e 3 anos, o que corrobora com as diferenças observadas na Figura 5.

Figura 6 – Riscos relativos para modelos de Cox univariados de variáveis categóricas



A Figura 6 apresenta o resultado da regressão do modelo de Cox, considerando modelos univariados para cada característica. Dessa forma, podemos observar os valores dos coeficientes do modelo, e também podemos observar o nível de significância do teste de Log-Rank.

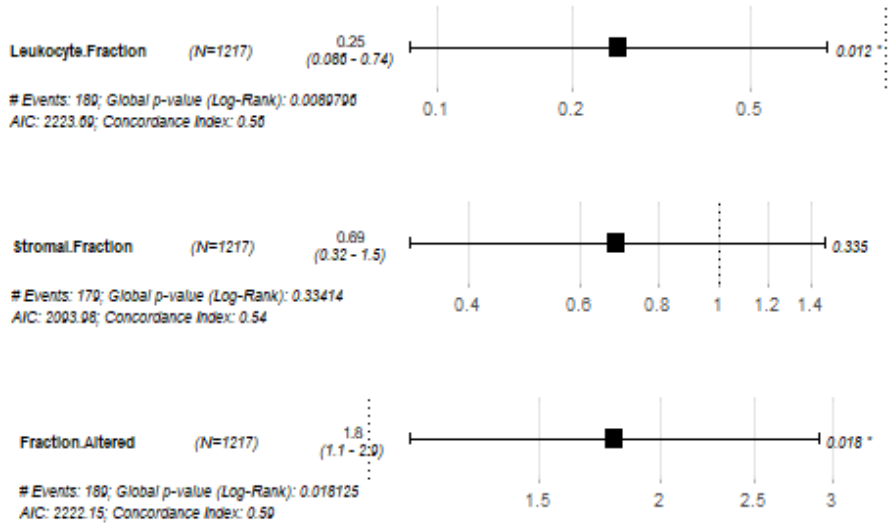
O teste de Log-Rank é utilizado para verificar se há ou não diferenças entre os grupos, visto que sua hipótese nula é que não há diferenças no resultado do modelo quando as variáveis se alteram, ou seja, considerando um nível de significância de 5%, podemos dizer que as categorias cujo p-valor do teste de Log-Rank é inferior a 0,05 são significantes para o modelo, pois a alteração destas variáveis altera significativamente a função risco.

Outro fato que corrobora com esta afirmação é que, ao analisarmos os intervalos de confiança para as covariáveis que possuem significância no modelo, observamos que os riscos relativos são sempre diferentes de 1, o que significa que a covariável interfere na função risco.

Dessa forma, podemos verificar através do valor global do teste de Log-Rank que existem diferenças significativas nos agrupamentos por subtipo, subtipo imune, estadiamento, e etinia. Além disso, podemos verificar, por exemplo, que no agrupamento por subtipo, há diferença significativa entre o grupo Basal (escolhido como referência) e o grupo Normal, enquanto que ou outros grupos não apresentam diferenças significativas com o grupo Basal.

O teste de Log-Rank também é utilizado para verificar a relevância de variáveis continuas para o modelo, visto que as análises feitas anteriormente são aplicáveis apenas em variáveis categóricas. A Figura 7 mostra o resultado dos modelos de Cox univariados para cada variável contínua que foi considerada, e também os resultados dos testes de Log-Rank para essas variáveis.

Figura 7 – Riscos relativos para modelos de Cox univariados para variáveis contínuas

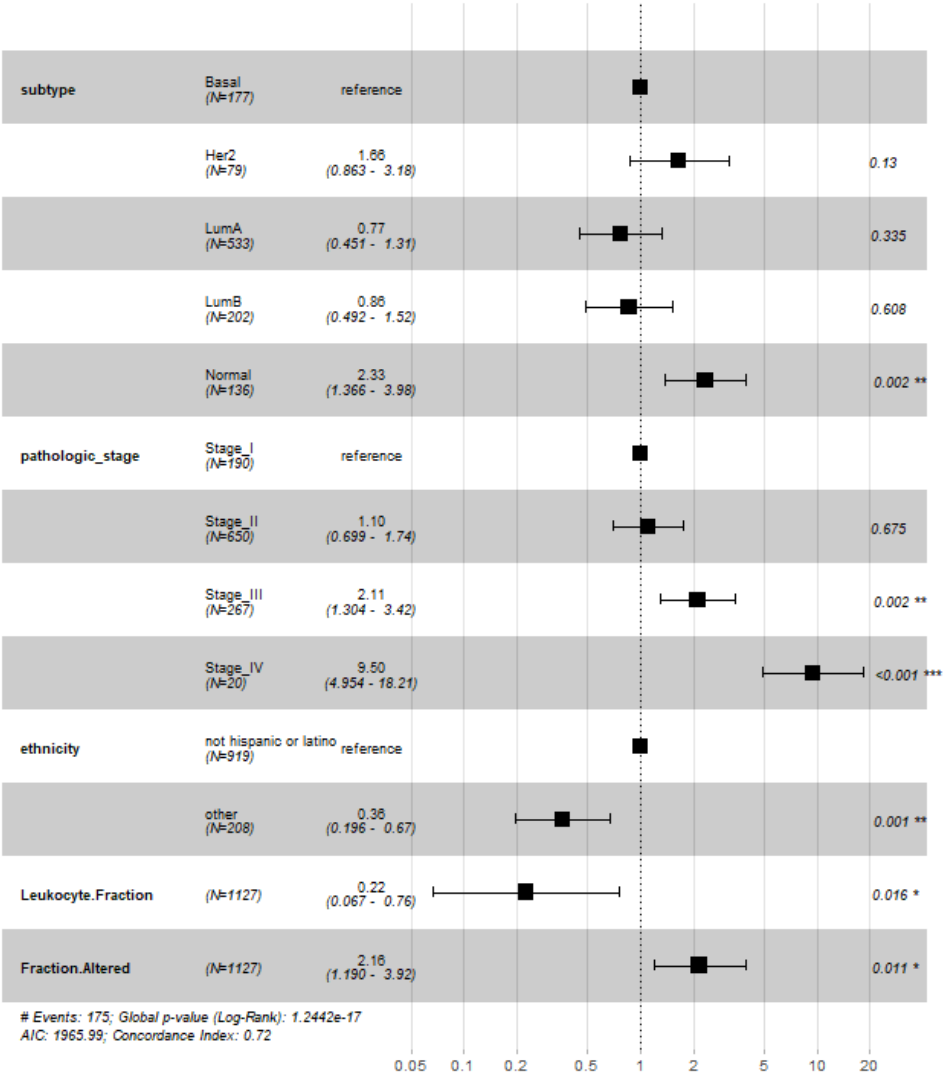


Dessa forma, podemos interpretar os resultados obtidos, de forma a concluir que, nos modelos univariados, as seguintes características são relevantes: o subtipo Normal, que possuí um risco relativo, ou seja, aumenta a taxa de risco, em 2,25, quando comparado com o subtipo Basal; o subtipo imune C4, que possuí risco relativo 2,03 comparado com o subtipo C1; o estadiamento 3 e 4, que possuem riscos relativos de 2,4 e 9,4, respetivamente, quando comparados com o estágio 1; a etnia, cujas pessoas que declararam ser hispânicos, latinos, ou não declarados apresentam risco relativo de 0,37 quando comparados aos pacientes que se declararam não ser hispânicos ou latinos, e, neste caso, podemos interpretar o risco relativo menor que 1 como sendo um fator de proteção apresentado por esta característica; a fração de leucócitos, que possuí risco relativo de 0,25; e a fração alterada, que possuí risco relativo de 1,8.

3.3.3. Análise Multivariada

O próximo passo da análise é a elaboração e ajuste de um modelo de Cox para múltiplas covariáveis. Para isso, selecionamos inicialmente as variáveis que apresentaram significância global inferior a 0.1 nos modelos univariados apresentados nas Figuras 6 e 7. Dessa forma, foi construído um modelo de Cox multivariado com as variáveis subtipo, subtipo imune, estadiamento, etinia, fração de leucócitos e fração de células alteradas.

Figura 8 – Riscos relativos para modelo de Cox multivariado



Para verificar a significância das variáveis de um modelo de Cox multivariado, podemos, inicialmente, verificar os valores dos riscos relativos e suas significâncias, que parecem ser relevantes, conforme apresentado na Figura 8. Entretanto, também é necessário verificar se a condição de riscos proporcionais do modelo de Cox está satisfeita, o que pode ser feito através do teste de resíduos de Schoenfeld, cujo resultado é apresentado na Tabela 3.

Tabela 3 – Probabilidade de sobrevivência após 1 e 3 anos

Variável	Qui Quadrado	Graus de Liberdade	P-Valor
Subtipo	13,4557	4	0,0093
Subtipo Imune	8,4421	4	0,0767
Estadiamento	9,9793	3	0,0187
Etnia	0,0757	1	0,7832
Fração de Leucócitos	0,4177	1	0,5181
Fração Alterada	5,2256	1	0,0223
Global	31,3361	14	0,0050

O teste de resíduos de Schoenfeld testa a hipótese nula de que não há dependência entre o tempo e a covariável, ou o modelo. Dessa forma, podemos perceber que as variáveis: Subtipo, Subtipo Imune, Estadiamento, Fração de Leucócitos e Fração alterada se mostram dependentes do tempo, uma vez que apresentam $p < 0,05$ e, portanto, violam a condição de riscos proporcionais. Dessa forma, fica evidente que para ajustar corretamente o modelo de Cox, será necessário utilizar variações mais robustas do modelo, que consideram a dependência temporal das covariáveis, o que foge do escopo deste trabalho.

3.4. Resultados Obtidos

A partir do desenvolvimento realizado na seção anterior, podemos verificar que a maior parte dos objetivos deste trabalho foi alcançada, visto que todas as análises propostas foram realizadas, sendo elas a análise exploratória e tratamento inicial da base, a descrição da função de sobrevida através das curvas de Kaplan-Meier, a utilização de testes de Log-Rank e da regressão de Cox para verificar a significância das covariaveis no modelo. Entretanto, o ajuste do modelo não foi satisfatório, visto que se constatou que, para tal, seria necessária a utilização de um modelo de Cox mais robusto, o que está fora do escopo do trabalho, o que também compromete a interpretação feita dos riscos relativos, que, apesar de poder ser feita, não é correta uma vez que o modelo não satisfaz a condição de riscos proporcionais.

3.5. Dificuldades e Limitações

Durante o desenvolvimento do trabalho, uma das dificuldades encontradas foi no tratamento e entendimento dos dados da base, visto em alguns casos eles estavam dispersos em várias colunas, e também pela dificuldade no entendimento de conceitos da área médica. Entretanto, ao mesmo tempo, este aspecto do trabalho também foi uma excelente oportunidade de aprendizado multidisciplinar, e que só pode ocorrer graças à parceria com o hospital A.C. Camargo.

Outra dificuldade encontrada foi quanto ao ajuste do modelo de Cox, visto que, após a técnica ser aplicada, se constatou que algumas das variáveis utilizadas não satisfaziam a condição de riscos proporcionais, portanto seria necessária a aplicação de modelos mais robustos, o que foge do escopo deste trabalho, mas que poderá ser realizado em trabalhos posteriores.

3.6. Considerações Finais

Nesta seção foi apresentado em detalhes todo o desenvolvimento realizado neste projeto para a realização da Análise de Sobrevida com a aplicação de Kaplan-Meier e dos modelos de Cox, assim como também foram apresentados os resultados e discussões relevantes. A seguir, no próximo capítulo serão feitas a conclusão e as considerações finais deste trabalho.

CAPÍTULO 4: CONCLUSÃO

4.1. Contribuições

A análise desenvolvida proporcionou ao aluno o aprendizado das técnicas utilizadas em estudos estatísticos aplicados à área médica, o que foi, sem dúvidas, uma oportunidade única de aprendizado multidisciplinar. Foram realizadas análises relevantes para o estudo da análise de sobrevivência para os dados de câncer de mama, através das curvas de Kaplan-Meier, e dos modelos de Cox. E, além disso, os códigos desenvolvidos podem ser utilizados como base para a realização de Análises de Sobrevivências aplicadas a outros conjuntos de dados, ou ainda para a continuação da análise com métodos mais robustos, ou para comparação com métodos de aprendizado de máquina.

4.2. Trabalhos Futuros

Para trabalhos futuros, um aspecto que pode ser explorado é a utilização de modelos de Cox que consideram a dependência temporal, ou a interação entre as covariáveis, a fim de tornar o modelo mais robusto de modo que a condição de riscos proporcionais seja satisfeita.

Outro aspecto que será explorado, tendo em vista o trabalho de mestrado que está sendo desenvolvido por outro integrante do grupo de pesquisa, é a relação do estudo feito nesse trabalho com a adição de dados genéticos e técnicas de aprendizado de máquina, a fim de verificar se esta combinação proporciona uma melhora na explicação do modelo.

REFERÊNCIAS

COLOSIMO, Enrico Antônio; GIOLO, Suely Ruiz. **Análise de sobrevivência aplicada**. ABE - Projeto Fisher, 2006.

INCA - Instituto Nacional do Câncer. **Estimativas 2014.** Ministério da Saúde. 2014. Dísponível em: < http://www.inca.gov.br >. Acesso: 01/11/2021.

INCA - Instituto Nacional do Câncer. **Falando sobre Doenças de Mama.** Ministério da Saúde. 2015a. Disponível em: <http://www.inca.gov.br>, Acesso em 01/11/2021.

INCA - Instituto Nacional de Câncer. **Câncer de mama: é preciso falar disso** / Instituto Nacional de Câncer José Alencar Gomes da Silva. – 3. ed. – Rio de Janeiro: Inca, 2015b.

INCA - Instituto Nacional de Câncer. **Estimativa 2016: incidência de câncer no Brasil.** Rio de Janeiro: INCA, Coordenação de Prevenção e Vigilância, 2015c.

ROSNER, Bernard. **Fundamentals of Biostatistics**. 7. ed. Cengage Learning, 2011.

SHREFFLER, Jacob; HUECKER, Martin R. Survival Analysis. StatPearls: Treasure Island (FL): StatPearls Publishing, 2021. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK560604/. Acesso em: 23 nov. 2021.

THORSSON, Vésteinn. *et al*. The Immune Landscape of Cancer. **Immunity**, v. 48(4), ed. 14, p. 812-830, 2018. DOI doi.org/10.1016/j.immuni.2018.03.023. Disponível em: www.ncbi.nlm.nih.gov/pmc/articles/PMC5982584/. Acesso em: 23 nov. 2021.