Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Pedro Henrique Pavan Gonçalves

FATORES DE RISCO PARA ÓBITO POR HANTAVIROSE NO PARANÁ, 1992-2016, ABORDAGEM VIA UM MODELO DE FRAÇÃO DE CURA.

Curitiba
2022

Pedro Henrique Pavan Gonçalves

FATORES DE RISCO PARA ÓBITO POR HANTAVIROSE NO PARANÁ, 1992-2016, ABORDAGEM VIA UM MODELO DE FRAÇÃO DE CURA.

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador(a): Silvia Emiko Shimakura

Curitiba
2022

Dedico esse trabalho a todos que estiveram ao meu lado durante todo esse tempo na Universidade, me dando forças e auxiliando sempre que necessário.

Agradecimentos

A minha familia, que sempre me apoiou em todas as decisões e sempre cultivou a importância dos estudos.

Aos meus companheiros Paulo, William, Felipe, Nilton e Lucka, por toda a parceria durante todos esses anos. Sem a parceria deles, nada seria possível.

A minha namorada Fernanda Ribeiro Scharman, por todo o apoio e paciência nesse periodo.

A minha orientadora, Profª Silvia Shimakura, por toda a disponibilidade e auxilio.

Ao Bruno Vinicius Nassar e Iara Claudia de Jesus Chagas, por serem as pessoais responsáveis por acreditarem no meu potencial e darem uma oportunidade no mercado de trabalho mesmo antes de um diploma.

Por fim agradecer a Deus.

Quando você resolve um problema, você verá mais dez.
Masaaki Imai

Resumo

Com a evolução da medicina surge também a necessidade de métodos capazes de estudar fatores de interesse. Com a hantavirose, zoonose viral aguda que não conta com uma cura ou vacina, essa necessidade também tem se mostrado necessária. Visando entender os fatores de risco que levam ao óbito os infectados pela hantavirose, o presente trabalho propõe a utilização de um modelo de Fração de Cura. Através de dados fornecidos pelo Sistema de Informação de Agravos de Notificação (SINAN), foi possível realizar a análise contendo todos os casos de hantavirose confirmados no estado do Paraná e que apresentaram sintomas de janeiro de 1992 até junho de 2016. Os resultados obtidos foram satisfatórios, onde pudemos analisar e avaliar fatores associados ao óbito dos infectados. O estudo é promissor, podendo ser aplicado em outras áreas oferecendo diversas possibilidades, se aplicados a dados mais completos.

Palavras-chave: Hantavirose. Análise de Sobrevivência. Fração de Cura. Kaplan Meier.

Sumário

1 Introdução

A hantavirose, zoonose viral aguda, cuja infecção em humanos no Brasil se apresenta na forma da Síndrome Cardiopulmonar por Hantavírus, apresentou seus primeiros casos registrados em 1993, e desde então tem sido notificada em todas as regiões do país.

Zoonoses são doenças infecciosas que saltam de animais não humano para humanos. Os patógenos zoonóticos podem ser bacterianos, virais ou parasitários, ou podem envolver agentes não convencionais, se espalhando assim para humanos por contato direto ou por meio de alimentos, água ou meio ambiente. Eles representam um grande problema de saúde pública em todo o mundo devido à nossa estreita relação com os animais, seja na agricultura, como companheiros ou no ambiente natural (OMS, 2020).

A transmissão da hantavirose é feita por roedores. O mais comum é que o contágio ocorra diretamente pela inalação de partículas de urina, fezes e saliva de roedores silvestres, não pelo contato com outros humanos infectados. Por isso, os casos da doença costumam ser isolados. Diferentemente dos seres humanos, roedores, como ratos e ratazanas, podem carregar o hantavírus por toda a vida sem adoecer (BBC, 2021).

Quando olhamos os dados relacionados a essa zoonose, de 2007 a 2015, foram notificados 13.181 casos de hantavirose no Brasil, dos quais 8% ($N = 1,060$) foram confirmados e 3,1% ($N = 410$) evoluíram para óbito. Observou-se uma média de 1.465 casos suspeitos notificados por ano, sendo 2008 (N=1.148) e 2013 (N=1.804) os períodos de menor e maior número de notificações, respectivamente (OLIVEIRA; DUARTE, 2018).

Segundo a OMS, não há nenhum tratamento, cura ou vacina para a infecção. A alta taxa de mortalidade e dificuldade com o tratamento são fatores que juntamente com a não descoberta de um tratamento tem aumentado a preocupação com a doença que desafia as autoridades de saúde pública ao redor do mundo. As alternativas terapêuticas para os indivíduos infectados limitam-se à introdução de medidas de suporte na fase aguda em ambiente hospitalar, preferivelmente em UTIs.

Dadas as circunstâncias, este presente trabalho de conclusão de curso tem como finalidade estender o trabalho realizado pela Daniele Akemi Aritra (ARITA, 2019), utilizando os mesmos dados e buscando analisar a sobrevida desses pacientes diagnosticados com hantavirose.

Dado o fato de que nem todos indivíduos do estudo experimentaram o evento de interesse (isto é, não foram a óbito devido à contaminação da hantavirose) até o termino do estudo, é proposta a utilização do modelo de sobrevivência com fração de cura apresentado por Corbiere e Joly (2007), para analisar a sobrevida desses pacientes, onde a variável resposta é o tempo decorrido entre a data do primeiro sintoma do paciente e a data em que o mesmo foi levado a óbito.

2 Objetivos

2.1 Objetivo Geral

O presente estudo tem como objetivo estudar fatores associados ao tempo de cura ou óbito, de pacientes infectados Hantavirose no Estado do Paraná no período de 1992 a 2016.

2.2 Objetivos Específicos

a. Revisar a literatura no que diz respeito às abordagens propostas para Análise de Sobrevivência;
b. Revisar os pacotes disponíveis no software R Team (2021) para estimação e diagnóstico de uma modelagem usando Análise de Sobrevivência fazendo uso dos pacotes survival Terry M. Therneau e Patricia M. Grambsch (2000) e smcure Cai et al. (2022);
c. Realizar uma análise descritiva dos dados descritos na Seção 3.1.1 para entendimento mais detalhado e consistente das informações;
d. Com base nos dados obtidos do banco de monitoramento da Secretaria de Estado da Saúde do Paraná (SESA/PR), realizar uma modelagem usando Análise de Sobrevivência em busca de definir os fatores associados ao óbito ou cura de um paciente infectado por Hantavirose.
e. Apresentar o modelo, discutir os resultados obtidos e tirar conclusões.

3 Revisão de Literatura

A hantavirose é uma doença grave e aguda, com alta taxa de letalidade, cujo nome tem origem no rio "Hantan" na Coréia, onde vários soldados americanos adoeceram durante a guerra dos anos 50. A doença manifesta-se sob a forma renal, com febre hemorrágica, e sob a forma pulmonar (Ambientebrasil, 2021).

A Hantavirose é uma das zoonoses que vem preocupando as autoridades sanitárias de todo o mundo. Sua ocorrência se deve principalmente a distúrbios ecológicos, destacando-se desmatamentos, alterações em ecossistemas associados ao comportamento econômico, social e cultural do homem.

A virose surge como um importante problema de saúde pública tanto em zonas rurais como em zonas urbanas (Ambientebrasil, 2021).

Atualmente, a hantavirose distribui-se globalmente incluindo a Europa, a Ásia, África e Américas. Na América do Sul, os países mais afetados são Brasil, Argentina, Chile e Paraguai (KRUGER et al., 2015).

Com o agravamento do número de casos dessa zoonose ao redor do mundo se faz necessário utilizar métodos capazes de entender os fatores de risco para o óbito dessa doença.

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas décadas do século passado. A razão deste crescimento é o desenvolvimento e aprimoramento de técnicas estatísticas combinado com computadores cada vez mais velozes. Uma evidência quantitativa deste sucesso é o número de aplicações de análise de sobrevivência em medicina (COLOSIMO et al., 2006).

O termo análise de sobrevivência refere-se basicamente a situações médicas envolvendo dados censurados (observação parcial da resposta). Entretanto, condições similares ocorrem em outras áreas que usam as mesmas técnicas de análise de dados (COLOSIMO et al., 2006).

A grande vantagem da Análise de Sobrevivência é permitir utilizar a informação de todos os participantes até o momento que experimentam o evento final, ou são censurados. Assim, essa é a técnica ideal para analisar respostas binárias (ter ou não ter um evento), em estudos longitudinais. Portando esse tipo de análise, compara a rapidez com que os participantes desenvolvem determinado evento, ao contrário, de comparar as porcentagens de observações que o desenvolvem ao fim do período (BOTELHO; SILVA; CRUZ, 2009).

Com a evolução da medicina, doenças antes letais tem se tornado curáveis. Tal fato aumentou a necessidade de desenvolver modelos estatísticos capazes de analisar quando o tratamento é capaz de curar a doença ou retardar a progressão caso seja não curável. O modelo de fração de cura, primeiramente introduzido por Boag, Berkson e Gage (1952), é um dos mais populares modelos para estimar a taxa de cura do tratamento e a taxa de sobrevivência de pacientes não curado ao mesmo tempo (CAI et al., 2012).

Sabendo da ampla utilização dos métodos de Análise de Sobrevivência e do baixo nível de conhecimento quanto aos fatores de riscos da hantavirose, faz-se necessário estudos mais detalhados que consigam entender os fatores de risco e tratamento adequado para a doença.

4 Material e Métodos

4.1 Material

4.1.1 Conjunto de dados

Os dados analisados neste trabalho são provenientes do Sistema de Informação de Agravos de Notificação (SINAN), onde são registradas as informações contidas nas fichas de investigação.

A população do estudo compreendeu todos os casos de hantavirose confirmados no estado do Paraná e que apresentaram início dos sintomas dentro do período do estudo (janeiro de 1992 a junho de 2016).

Neste trabalho, a variável resposta de interesse tempo (em dias), foi calculada através do tempo entre as datas data de óbito e a data do 1º sintoma do indivíduo conforme a função a seguir.

$$tempo = \text{data óbito} - \text{data primeiro sintoma}$$

Para os indivíduos que não apresentaram data de óbito, o tempo (em dias) foi calculada através do tempo entre as datas data de encerramento e a data do 1º sintoma do indivíduo conforme a função a seguir.

$$tempo = \text{data encerramento} - \text{data primeiro sintoma}$$

Feito isso, os dados foram divididos entre cura e óbito, sendo cura o "indivíduo notificado por serviço de saúde do Estado do Paraná, no período de estudo e que tenha sido confirmado para hantavirose com evolução para cura" e óbito, o "indivíduo notificado por serviço de saúde do Estado do Paraná, no período de estudo e que tenha sido confirmado para hantavirose com evolução para óbito". Pacientes que não apresentaram data de óbito nem data de cura foram considerados como censura.

4.1.2 Recursos Computacionais

O software escolhido para a condução do estudo é o software livre R Core Team (2021), que será utilizado como ferramenta para a análise exploratória, bem como para ajustar os modelos. Os pacotes *survival** Therneau (2022) e *smcure** Cai et al. (2022) serão utilizados no ajuste dos modelos de Análise de Sobrevivência.

4.2 Métodos

4.2.1 Análise de Sobrevivência

Em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse, tempo esse denominado tempo de falha.

O tempo de falha é a ocorrência de um determinado evento, que pode ou não ser pré-estabelecido no início da pesquisa. Por exemplo, uma falha pode ser a morte de um ser em estudo ou uma recaída, mas também pode ser considerada como a melhora do quadro clínico do paciente. É muito importante que o tempo de início do estudo seja precisamente definido

4.2.1.1 Função de Sobrevivência

Uma das principais funções probabilísticas usadas para descrever um estudo de sobrevivência é a função de sobrevivência, que em termos probabilísticos é escrita como:

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

A função de sobrevivência é definida com a probabilidade da falha ocorrer até o tempo t, ou seja, a probabilidade de um indivíduo sobreviver ao tempo t.

Em consequência, a função de distribuição acumulada é definida como a probabilidade da falha ocorrer antes do tempo t, isto é, $F(t) = 1 - S(t)$.

4.2.1.2 Função de Taxa de Falha

Dado um intervalo de tempo $[t_1, t_2)$, a probabilidade da falhar ocorrer no intervalo pode ser expressa em termos da função de sobrevivência como:

$$S(t_1) - S(t_2).$$

Dado que a falha não ocorreu antes do intervalo $t1$, definimos a taxa de falha durante o período $[t1, t2)$ como a probabilidade de ocorrer a falha durante esse intervalo, dividida pelo comprimento do intervalos. Sendo assim, podemos definir a função que expressa a taxa de falha no intervalo $[t1, t2)$ por:

$$\frac{S(t_1) - S(t_2)}{(t_1 - t_2)S(t_1)}$$

De forma geral, redefinindo o intervalo como $[t, t + \Delta_t)$, a expressão anterior pode assumir a seguinte forma:

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

4.2.1.3 Censura

A principal característica da análise de sobrevivência é a utilização de todas as informações disponíveis, ou seja, tanto os indivíduos de um experimento que apresentaram o tempo de falha quanto os que não experimentaram o evento. Essa informação é chamada de Censura, que se refere aos que não experimentaram o evento de interesse

Mesmo com dados censurados contendo informações incompletas ou parciais sobre um paciente, todos os dados registrados devem ser considerados, visto que o tempo até a ocorrência do evento (falha), para todos os pacientes, é superior ao tempo registrado até o último acompanhamento. Ressaltando que os dados, mesmo que censurados, fornecem informações importantes sobre o tempo de vida do paciente. A não utilização de tais dados pode fornecer resultados viciados.

Há diferentes tipos de censura e esta depende exclusivamente da história do estudo e de mecanismos aleatórios, externos ao estudo em questão, podendo ser de um dos tipos a seguir:

1. Censura tipo 1 - Alguns indivíduos não apresentaram o evento até o final do experimento

2. Censura tipo 2 - Estudo é finalizado após a ocorrência de um número pré-estabelecido de falhas

3. Censura aleatória - O acompanhamento de alguns indivíduos foi interrompido por alguma razão e alguns indivíduos não experimentaram o evento até o final do periodo de acompanhamento.

Sendo t o tempo registrado e δ a variável indicadora de falha ou censura, a variável resposta é composta pelo par (t, δ). A variável indicadora de falha ou censura pode ser escrita como:

$$\delta = \begin{cases} 1, \text{ se t é um tempo de falha} \\ 0, \text{ se t é um tempo censurado} \end{cases}$$

4.2.2 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier, também chamado estimador produto-limite, foi proposto por Kaplan e Meier em 1958 e é sem dúvida o mais utilizado em estudos clínicos.

Na sua construção, o estimado de Kaplan Meier considera tanto intervalos de tempo quanto forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra(COLOSIMO et al., 2006).

Considere $t_1 < t_2 < \cdots < t_k$, os k tempos distintos e ordenados de falha, d_j o número de falhas em t_j, $j = 1, \cdots, k$, e n_j o número de indivíduos sob risco em t_j. O estimador de Kaplan Meier é, então, definido por:

$$\hat{S}(t) = P(T > t) = \prod_{j:t_j<t} (\frac{n_j - d_j}{n_j}) = \prod_{j:tj<t} (1 - \frac{d_j}{n_j})$$

As estimativas via estimador de Kaplan Meier são usualmente representadas graficamente e mostram o comportamento da curva de sobrevivência. Pressupõe uma única causa de falha de interesse, que ocorre com probabilidade igual a 1 se o tempo de seguimento for suficientemente longo, e trata como censura os tempos observados para aqueles indivíduos que não apresentam o evento de interesse até o final do tempo de seguimento.

Para considerarmos as hipóteses de diferenças significativas entre as curvas de sobrevivência, isto é, testar $S_1(t) = S_2(t)$ůůů $= Si(t)$ utilizaremos o teste Logrank que identifica a significância desta diferença por meio da atribuição de pesos.

4.2.3 Modelo de Fração de Cura

O modelo de riscos proporcionais e tempo de falha acelerado são os mais populares em análise de sobrevivência.

O modelo de riscos proporcionais proposto por Cox (1972) (PH), é dado por:

$$h(t|x) = g(\alpha' x) h_0(t)$$

Onde $g(\cdot)$ é uma função positiva, que assume o valor 1 quando seu argumento é igual a zero, $h_0(\cdot)$ representa a função de risco básica para uma unidade quando $x = 0$ e α' vetor de coeficientes a serem estimados.

O Modelo de Tempo de Falha acelerado (AFT) dado por Kalbáeish e Prentice (1980) é dado por:

$$h(t|x) = g(\alpha' x) h_0(g(\alpha' x)t)$$

Uma suposição comum não declarada por trás desses dois modelos é que todos os pacientes irão eventualmente experimentar o evento de interesse, dado que o tempo de acompanhamento é longo o suficiente.

O modelo de fração de cura é um tipo especial de modelo de sobrevivência e assume que a população do estudo é uma mistura de indivíduos suscetíveis que podem experimentar o evento de interesse, e indivíduos curados/não suscetíveis que nunca irão experimentar o evento. Para tal dado, modelos de sobrevivência mais usuais não seriam apropriados pois não contam com a possibilidade de cura.

Seja T o tempo de falha de interesse, $1 - \pi(z)$ a probabilidade de um paciente ser curado dependendo de z, e $S(t|x)$ sendo a probabilidade de sobrevivência de pacientes não curados dependendo de x. O modelo de fração de cura pode ser expresso como:

$$S_{pop}(t|x, z) = \pi(z)S(t|x) + 1 - \pi(z),$$

onde $\pi(z)$ é referido como "incidência" e $S(t|x)$ é referido como "latência". Se o modelo de riscos proporcionais for usado para modelar a parte latente, o modelo de fração de cura é chamado de modelo de fração de cura de riscos proporcionais. Em vez disso, se o modelo de tempo de falha acelerado for aplicado para modelar a parte latente, ele é chamado de modelo de fração de cura de tempo de falha acelerado.

Uma vantagem do modelo de fração de cura é a modelagem de indivíduos curados e não curados ser feita separadamente.

Geralmente, a função de ligação logito é usada para modelar o efeito z, mas também permite a utilização de outras funções de ligação.

4.2.3.1 Método de Estimação

Conforme mencionado na seção 4.1.2, o pacote com a implementação utilizado será o *smcure.** Portanto é importante ressaltar algumas particularidades e explicar a parte computacional do método.

Para a estimação dos parâmetros de interesse é utilizado o algoritmo EM (algoritmo de maximização de expectativa), que é um método iterativo para estimar parâmetros em modelos estatísticos, quando o modelo depende de variáveis latentes, ou seja, não observadas, que no presente estudo é dado pela cura do indivíduo.

Sendo $\mathbf{O} = (t_i, \delta, z_i, x_i)$ denotando os dados observados para o i-ésimo indivíduo onde $i = 1, \cdots, n$, t_i é o tempo de sobrevivência observado, δ o indicador de censura, e z_i, x_i são as possíveis covariáveis de incidência e latência respectivamente. Vale ressaltar que as mesmas covariáveis podem ser utilizadas para os componentes de incidência e latência, embora usemos notações diferentes.

Sendo $\Theta = (b, \beta, S_0(t))$ os parâmetros desconhecidos. Deixando Y como o indicador de que um indivíduo vai eventualmente ($Y = 1$) ou nunca ($Y = 0$) experimentar o evento, com probabilidade $1 - \pi(z)$. Dado $y = (y_1, y_2, \cdots, y_n)$ e $\mathbf{O}$, a função de Verossimilhança pode ser expressa como:

$$\prod_{i=1}^{n}[1-\pi(z_i)]^{1-y_i}\pi(z_i)^{y_i}h(t_i|Y=1,X_i)^{\delta_i y_i}S(t_i|Y=1,X_i)^{Y_i}$$

Devido à complexidade da equação de estimativa no algoritmo EM, os erros padrão dos parâmetros estimados não estão diretamente disponíveis. Para obter a variância de $\hat{\beta}$ e $\hat{b}$, o método desenha aleatoriamente amostras bootstrap com substituição.

Para um entendimento mais detalhado do método reecomenda-se a leitura do artigo escrito por Chao Cai, Yubo Zou, Yingwei Peng and Jiajia Zhanga em 2012 que se encontra presente nas referência do trabalho.

5 Resultados e Discussão

5.1 Material final do Estudo

A princípio foi realizada uma análise descritiva dos dados visando um maior entendimento da base.

A base de dados original era composta por 280 observações e 69 variáveis, que são elas os 69 campos presentes na ficha de investigação preenchida pelos pacientes.

Devido a dificuldades com a tratativa de alguns dados presentes na base e presença de dados faltantes para várias variáveis, com algumas atingindo um percentual de 96% de dados faltantes, decidiu-se restringir as análises com as seguintes variáveis:

1. Tempo: Tempo decorrido do primeiro sintoma do paciente até o óbito ou perda do acompanhamento.
2. Idade: Idade do paciente.
3. Sexo: Sexo do paciente (homem ou mulher).
4. Tontura: Apresentou tontura (sim ou não).
5. Cefaleia: Apresentou cefaleia (sim ou não).
6. Sangramento Respiratório: Apresentou sangramento respiratório (sim ou não).
7. Dispneia: Apresentou falta de ar (sim ou não).
8. Hipotensão: Apresentou problemas com pressão baixa (sim ou não).
9. Mialgia: Apresentou dores musculares (sim ou não).
10. Regional de Saúde: Regional de saúde na qual o paciente foi atendido, podendo ser União da Vitória, Guarapuava, Irati e Outros.
11. Sinais Hemorrágicos: Apresentou sinais hemorrágicos (sim ou não).
12. Internação: Paciente foi internado no período em que esteve com a doença (sim ou não).
13. Diarreia: Apresentou diarreia (sim ou não).
14. Respirador Mecânico: Precisou de respirador mecânico (sim ou não).

Para fins de visualização e melhor entendimento da variável, entendeu-se que seria adequado categorizar a variável *idade* para a análise exploratória, entre indivíduos de 1 a 20 anos de idade, 20 a 29, 30 a 39, 40 a 49, 50 a 59 e com idade superior a 60 anos.

Conforme já foi discutido anteriormente na seção 4.2.1.3, em análise de sobrevivência, as censuras são compostas por indivíduos do estudo que ainda não experimentaram o evento final. Portanto, além das 14 variáveis escolhidas, ainda foi adicionada a variável de censura, em que o indivíduo que apresentou censura é *0* e *1* para o indivíduo não censurado.

5.2 Análise Descritiva

Inicialmente, foi realizada uma análise descritiva dos dados apresentados na Seção 5.1 a fim de analisar a quantidade de indivíduos em cada categoria das variáveis.

Visando entender o comportamento das variáveis assim como a distribuição entre a censura, as visualizações para as variáveis foram feitas distinguindo entre censura sim e não.

Na Figura 1, conseguimos analisar a distribuição e comportamento das variáveis idade e sexo, além de suas respectivas censuras.

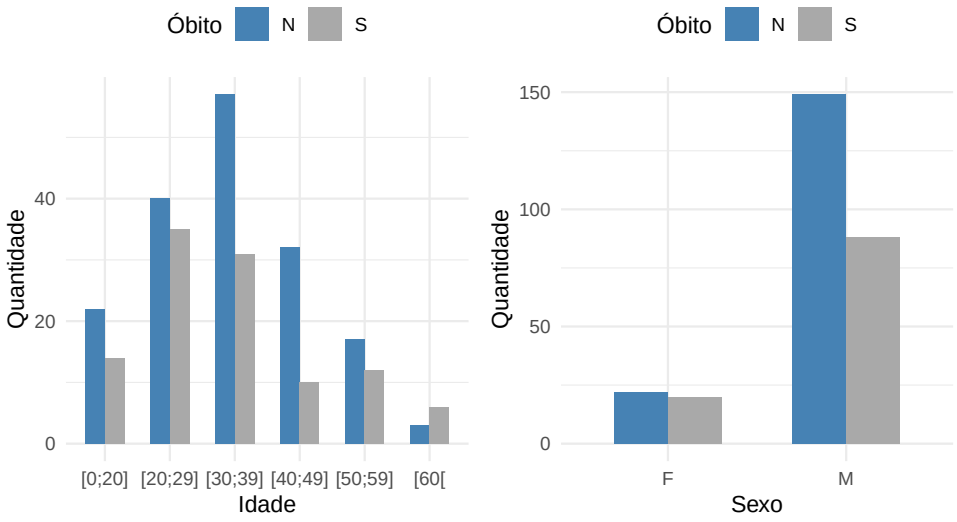

Figura 1 – Distribuição das variáveis Idade e Sexo

Dos 280 indivíduos presentes no estudo, 85,0% (238/280) eram do sexo masculino e os outros 15% (42/280) do sexo feminino. Em relação a idade, a média dos indivíduos é de 33 anos que também equivale a mediana dos dados. O indivíduo com a menor idade possui 2 anos, enquanto o indivíduo com a maior idade possui 80.

Entre as categorias da variável idade, 32% (88/280) estão presentes na categoria de indivíduos de 30 a 39 anos, e apenas 3% (9/280) dos indivíduos do estudo presentes na categoria de 60 ou mais anos de idade.

Quanto a variável resposta tempo, conforme a Figura 2, cerca de 88% (250/280) dos indivíduos apresentam um tempo igual ou inferior a 22 dias. Um ponto interessante de se observar é o fato de que para um tempo superior a 22 dias, apenas 2 dos 30 indivíduos apresentam censura, ou seja, experimentaram o evento de interesse (óbito). Possível notar uma alta taxa de óbito nos primeiros dias de infeção do paciente, que decai com o passar dos dias.

Dentre as 22 regionais de saúde do Paraná, no banco analisado o maior número de casos foi registrado em União da Vitória com 33% (93/280), seguido de Guarapuava com 23% (65/280) e Irati com 16% (46/280). Os demais foram agrupados na categoria "Outros".

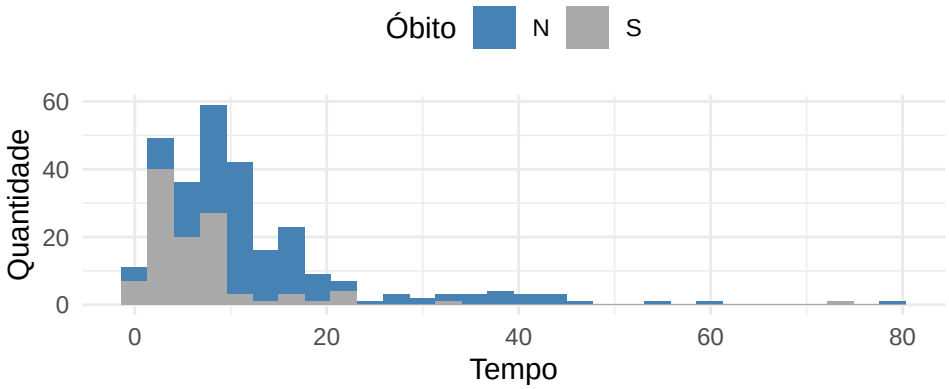

Figura 2 – Distribuição da variável tempo

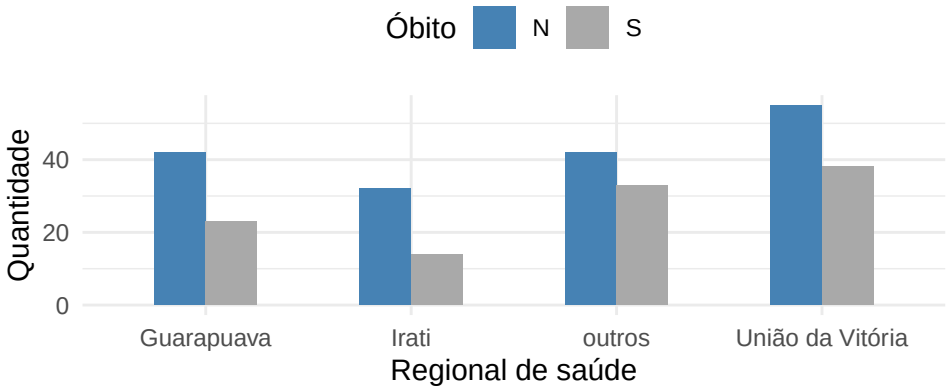

Figura 3 – Distribuição dos casos segundo a regional de saúde

Quanto as variáveis relacionadas a sintomas, na Tabela 1 são apresentadas as frequências absolutas e relativas (percentuais) em cada uma das categorias, bem como o número de óbitos e censuras.

Tabela 1 – Frequências absolutas, relativas, quantidade de censuras e quantidade de falhas para as variáveis dicotômicas apenas para a categoria sim.

Variável	Freq.Absoluta	Freq.Relativa	Censura	Falha
Tontura	162	57,50%	49	113
Cefaleia	242	86,43%	87	155
Sangramento Respiratório	123	43,93%	78	45
Dispineia	199	71,07%	92	107
Hipotensão	148	52,86%	94	54
Mialgia	229	81,43%	82	147
Sinais Hemorrágicos	31	11,07%	17	14
Diarreia	72	25,36%	27	45
Respirador Mecânico	118	32,50%	73	45

O sintoma que esteve mais presente entre os infectados foi a cefaleia, presente em 242 dos 280 indivíduos (86%), seguido da mialgia em 232 (83%).

Os sintomas que se mostraram menos presentes foram os sintomas sinais hemorrágicos com 31 aparições seguido da diarreia em 71 dos 280 indivíduos.

Os sintomas que se mostraram mais letais entre os estudados com 94 (33,5%) e 92 (32,8) respectivamente foram hipotensão, que significa uma baixa na pressão arterial do paciente infectado, e dispneia representada pela falta de ar ou dificuldade de respirar.

Na sequência, foi realizada uma análise uni variada (uma a uma) das variáveis presentes no estudo. Para isso, foi utilizado o estimador de Kaplan-Meier que nos permite visualizar as curvas de sobrevida associadas às categorias de cada variável, auxiliando assim a verificar a existência de associação delas com o tempo até o óbito.

Buscando oferecer uma melhor visualização dos resultados, os gráficos foram divididos em duas figuras. A Figura 4 contém 8 das 14 imagens, e o restante presente na Figura 5. O valor p resultante do teste logrank calculado está presente no canto inferior esquerdo de cada uma das imagens.

Na Figura 4 as variáveis Idade, Sexo, Tontura, Cefaleia, Sangramento Respiratório, Dispneia e Hipotensão se mostraram significativas através do teste *logrank*, com valores p inferiores a 0,10. Embora significativas, as variáveis Sexo e Cefaleia apresentaram um intervalo de confiança bem elevado. Um ponto interessante é o fato da variável Tontura, diferente das demais, se mostrar significativa com a categoria "NÃO" apresentando uma maior probabilidade de sobrevivência, ou sja, pacientes que apresentam esse sintoma apresentam uma probabilidade de sobrevivência superioro aos indíviduos infectados que não apresentam esse sintoma.

Na Figura 5 as variáveis Respirador Mecânico e Mialgia se mostraram significativas através do teste *logrank*, com valor p inferior a 0,10, e as demais todas não significativas.

Feita a análise exploratória dos dados e tendo um entendimento amplo do funcionamento e distribuição de cada uma das variáveis presentes no estudo, o próximo passo foi a modelagem desses dados utilizando do modelo de Fração de Cura modelando a parte latente com o Modelo de Riscos Proporcionais (PH). A intenção era utilizar os dois modelos apresentados na seção 4.2.3 (Modelo de Riscos Proporcionais e Tempo de Falha Acelerado) mas devido a problemas com o pacote e os dados utilizados foi utilizado apenas um deles para modelar a parte latente.

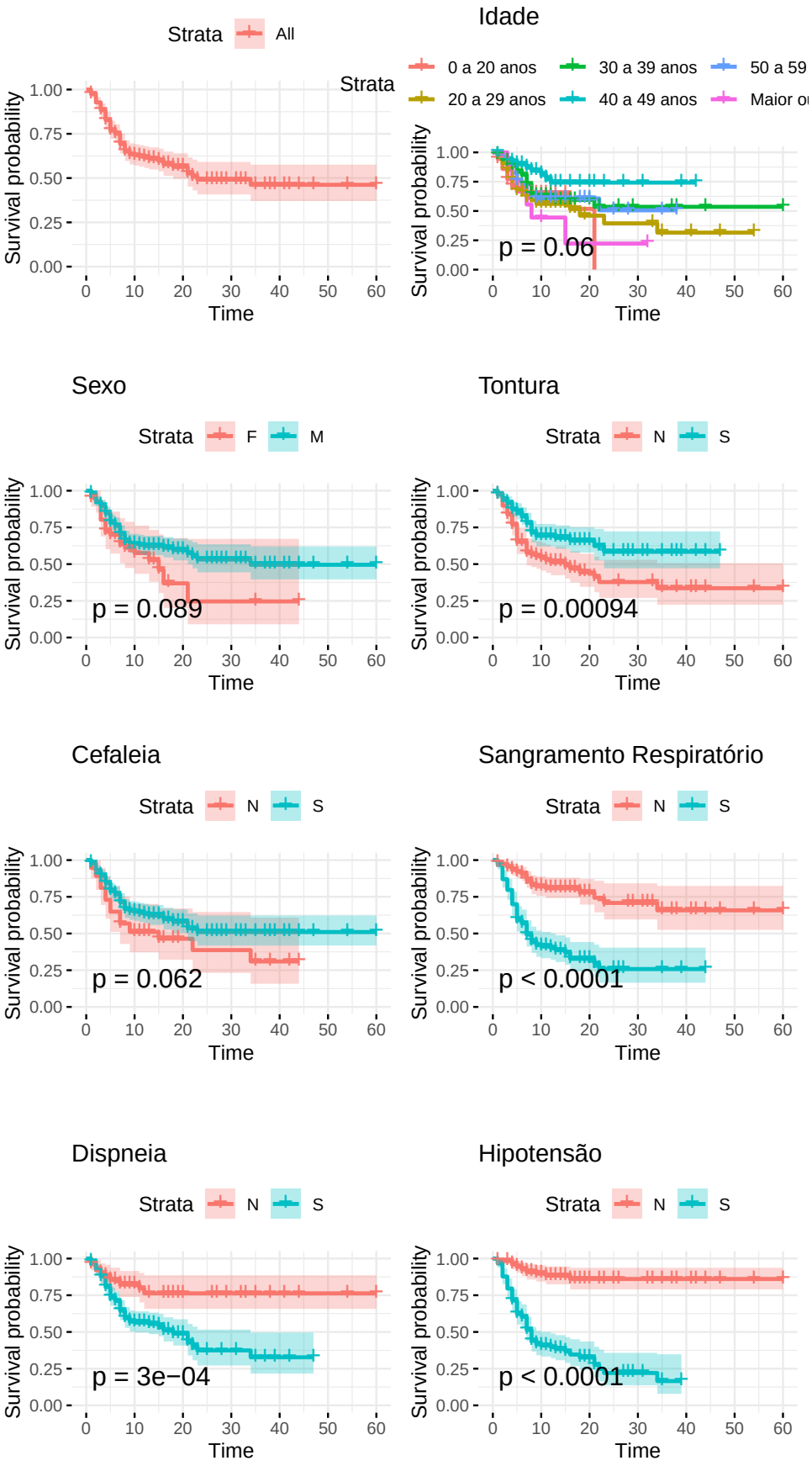

Figura 4 – Curvas de Sobrevivência univariada para o estimador de Kaplan Meier

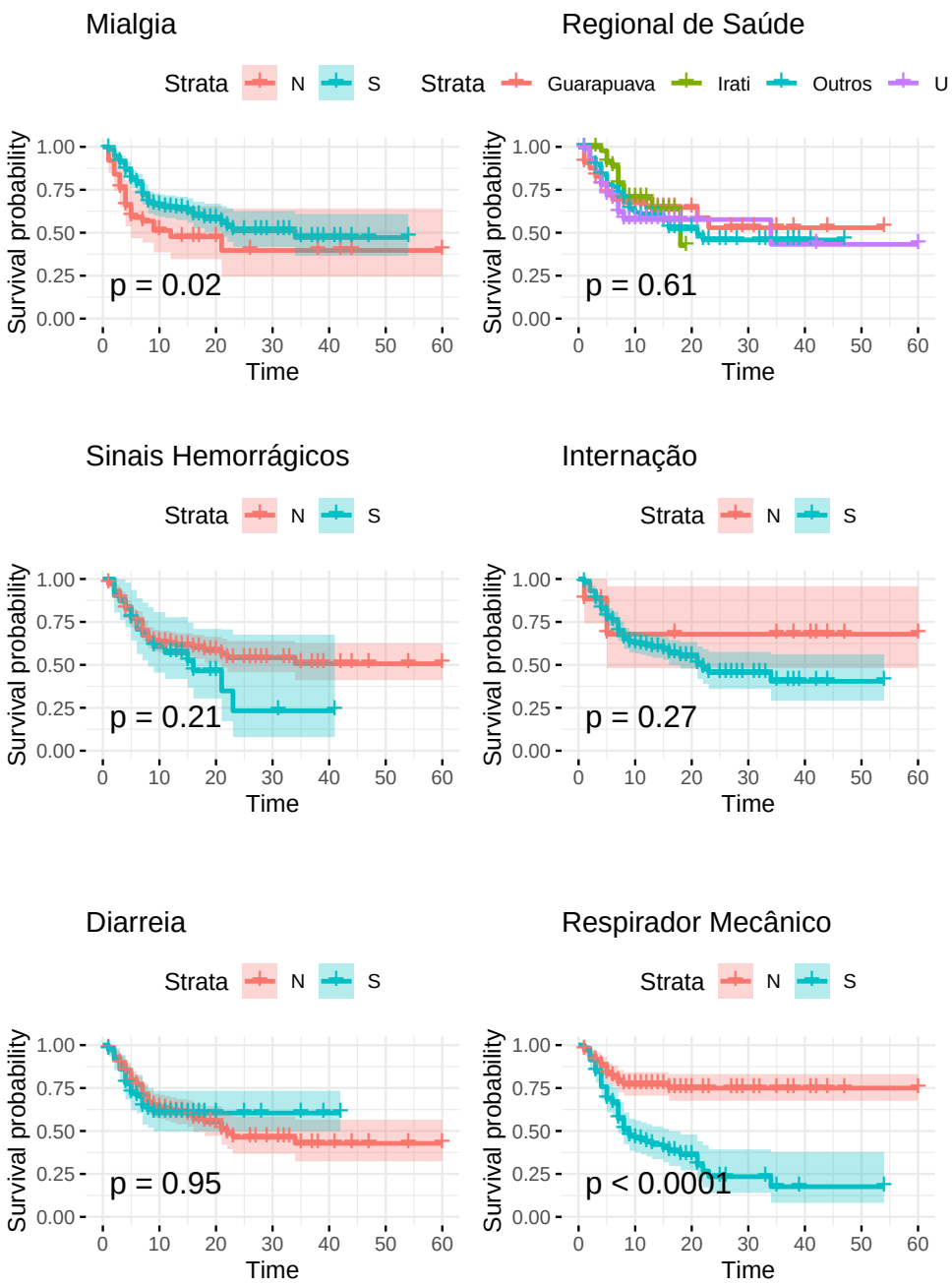

Figura 5 – Curvas de Sobrevivência univariada para o estimador de Kaplan Meier

5.3 Modelo de Fração de Cura

Os dados foram aplicados no modelo de fração de cura através do pacote *smcure*, considerando as covariáveis: Idade, sexo, tontura, cefaleia, sangramento respiratório, dispneia, hipotensão, mialgia, regional de saúde, sinais hemorrágicos, internação, diarreia e respirador mecânico.

As variáveis dicotômicas foram introduzidas aos modelos por meio de variáveis dummy, onde a primeira categoria de cada uma delas ficou como referência. O pacote utilizado para a modelagem apresenta uma particularidade para variáveis categóricas com mais de duas categorias, fazendo necessária a criação de $k-1$ variáveis dummy, onde k é o número de categorias da variável, portanto, para a variável *Regional de Saúde* 3 variáveis dummy foram adicionadas aos dados.

Na primeira etapa, foi ajustado um modelo separado para cada variável selecionada, avaliando as suas respectivas significâncias ao nível de 0,10.

Na Tabela 2, podemos observar as estimativas para cada um dos modelos, tanto para a probabilidade de cura quanto para o tempo de falha respectivamente, bem como os respectivos erros padrões e valores de p. Para a variável regional de saúde, a categoria "Guarapuava" ficou como categoria de referência.

Os erros padrões estimados foram obtidos baseados em 200 amostras de bootstrap. A escolha desse tamanho de amostra foi baseada em outros materiais e referências utilizadas pelo autor deste trabalho durante o processo de estudo do trabalho.

Muito importante entender que embora estejamos utilizando um modelo de Fração de Cura no estudo o interesse é estudar o óbito, portanto estaremos modelando a probabilidade de óbito ao invés da probabilidade de cura.

Ajustado um modelo para cada uma das variáveis selecionadas na seção 5.1, para a probabilidade de óbito 12 variáveis se mostraram significativas ao nível de 10% de confiança, fora elas Sexo, Tontura, Sangramento Respiratório, Dispneia, Hipotensão, Mialgia, Regional de Saúde, Sinais Hemorrágicos, Internação, Diarreia e Respirador Mecânico.

Para o tempo de falha, as variáveis que se mostraram significativas ao nível de 10% de confiança foram Sexo, Tontura, Sangramento Respiratório, Regional de Saúde para os níveis Irati e Outros, Diarreia e Respirador Mecânico. Olhando para a variável sangramento respiratório como exemplo, olhando para o coeficiente concluímos que pacientes que apresentaram esse sintoma tem um aumento no tempo médio de falha.

Na segunda etapa, foi ajustado um modelo com todas as variáveis, novamente avaliando as que são significativas ao nível de significância de 0,10. As que apresentaram um nível de significância abaixo do estabelecido foram retiradas do modelo. Conforme mencionado na Seção 4.2.3.1, importante lembrar que uma mesma covariável pode ser utilizada tanto para modelar o óbito quanto para modelar o tempo de falha, da mesma forma que uma covariável pode ser significativa para explicar o óbito, mas não significativa para explicar o tempo de falha e vice-versa.

A partir da Tabela 3 observam-se as estimativas dadas pelos modelos apenas com as variáveis significativas para o explicar o óbito e o tempo de falha respectivamente.

Tabela 2 – Estimativas, coeficientes, erros padrões e p-valores para cada um dos modelos com uma variável.

		Probabilidade de óbito			Tempo de falha		-
Parâmetro	Categoria	Coeficiente	Erro Padrão	Valor-p	Coeficiente	Erro Padrão	Valor-p
Idade	-	-0,005	0,156	0,768	-0,122	0,009	0,19
Sexo	Masculino	-1,915	0,612	0,001	-0,173	0,612	0,001
Tontura	Sim	-1,13	0,439	0,001	-0,438	0,239	0,067
Cefaleia	Sim	-1,395	1,262	0,268	-0,253	0,322	0,432
Sangramento Respiratório	Sim	1,997	0,491	0,001	1,092	0,322	0,001
Dispneia	Sim	2,923	0,462	0,001	0,053	0,362	0,883
Hipotensão	Sim	6,673	0,576	0,001	0,402	0,298	0,177
Mialgia	Sim	-0,189	0,478	0,001	-0,498	0,338	0,14
Regional de Saúde	Irati	0,983	0,651	0,001	-0,954	0,505	0,058
-	Outros	0,242	0,523	0,001	-0,88	0,49	0,072
-	União da Vitória	2,297	0,626	0,001	-0,79	0,581	0,174
Sinais Hemorrágicos	Sim	4,469	2,566	0,001	-0,045	0,342	0,896
Internação	Sim	2,489	0,76	0,001	-1,531	0,67	0,224
Diarreia	Sim	-1,636	0,391	0,001	0,943	0,29	0,001
Respirador Mecânico	Sim	7,394	0,585	0,001	-0,741	0,27	0,006

Para chegar nesse modelo final as variáveis foram introduzidas uma a uma, avaliando seu efeito e significância para o modelo.

Para a probabilidade de óbito as variáveis que restaram foram Sexo, Tontura, Sangramento Respiratório, Sinais hemorrágicos e Cefaleia.

Tabela 3 – Estimativas das variáveis significativas para a probabilidade de óbito do modelo final.

Parâmetro	Categoria	Coeficiente
Sexo	Masculino	-1,358
Tontura	Sim	-0,886
Sangramento Respiratório	Sim	2,299
Sinais Hemorrágicos	Sim	1,187
Cefaleia	Sim	-0,499

As que apresentaram um coeficiente negativo, ou seja, diminuem a probabilidade de óbito são Sexo (masculino), tontura e cefaleia. As demais apresentaram um coeficiente positivo. Interessante notar que a variável tontura (que já havia se mostrado significativa no Kaplan-Meier com probabilidade de sobrevivência superior para indivíduos que apresentam o sintoma) o coeficiente contém um sinal negativo, influenciando assim na diminuição da probabilidade de óbito.

Conforme apresentado na Seção 4.2.3, a probabilidade de óbito pode ser calculada baseada nos resultados da parte de probabilidade de Cura do modelo. Para um indivíduo do Sexo masculino, que apresentou todos os sintomas da Tabela 3, a probabilidade de cura é dada por:

$$\pi(z) = 1 - \frac{exp(\mathbf{bz})}{1 + exp(\mathbf{bz})} = 0,082$$

Ou seja, paciente do sexo masculino que apresentaram os sintomas da Tabela 3 tem uma probabilidade de cura de 8.2%.

Para o tempo de falha restaram Sexo, Sinais hemorrágicos, Respirador Mecânico, Cefaleia e Hipotensão.

Tabela 4 – Estimativas das variáveis significativas para o tempo de falha do modelo final.

Parâmetro	Categoria	Coeficiente
Sexo	Masculino	-0,002
Sinais Hemorrágicos	Sim	-0,349
Respirador Mecânico	Sim	0,048
Cefaleia	Sim	-0,020
Hipotensão	Sim	1,595

Das variáveis presentes para explicar o tempo de falha Respirador Mecânico e Hipotensão apresentam um coeficiente positivo, ou seja, aumentam a probabilidade de o indivíduo apresentar a falha, enquanto as demais apresentaram um coeficiente negativo.

6 Considerações Finais

O presente estudo apresentou dificuldade quanto a seleção e tratativa dos dados. Diversas informações que poderiam ser interessantes para a análise tiveram de ser descartas devido a problemas com a coleta das informações. Alguns sintomas presentes na base por exemplo, contaram com mais de 95% de dados *missing*, impossibilitando assim a utilização da variável. Seria interessante buscar formas de introduzir variáveis com baixa quantidade de dados *missing* utilizando técnicas existentes.

Com as informações que tínhamos disponível para a realização do trabalho, o objetivo principal era analisar fatores de risco para o óbito por hantavirose no Paraná, nos anos de 1992 até 2016, utilizando uma abordagem via um modelo de Fração de Cura. Para isso, foram aplicadas técnicas estatísticas com o intuito de ajustar o melhor modelo possível para os dados e, a partir dele, extrair conclusões.

Na análise descritiva, por meio do estimado de Kaplan Meier, pudemos identificar informações relevantes sobre os dados, como as taxas de sobrevivência e o respectivo valor de p extraído do teste log-rank.

O modelo de Fração de Cura aplicado apresentou um resultado interessante para a análise de interesse. As variáveis significativas no presente estudo foram coincidentes com as variáveis significativas no modelo de Riscos Proporcionais proposto pela Daniele Arita (ARITA, 2019).

Como sugestão para estudos futuros, seria interessante avaliar a maneira como os dados foram coletados. Um acompanhamento mais severo dos infectados talvez fosse mais recomendado para um melhor entendimento dos fatores associados ao óbito da doença, além de diminuir a quantidade de informações em branco na base. Quanto a parte computacional, seria interessante buscar criar uma implementação do modelo ao invés de utilizar implementações já existentes no R, dada as dificuldades que tivemos com o pacote selecionado para a modelagem dos dados. Além disso, para os dados utilizados não foi possível utilizar todas as possibilidades do pacote, o que talvez ressalte a importância de alguns testes por simulação para estudos futuros.

É importante ressaltar a necessidade de uma conscientização da população quanto a hantavirose, principalmente na zonas rurais do país, conscientizando da população quanto a doença.

Referências

Ambientebrasil. 2021. Disponível em: <https://ambientes.ambientebrasil.com.br/agropecuario/doencas_agropecuarias/hantavirose.html>.

ARITA, D. A. Survival of persons with hantavirus infection diagnosed in parana state, brazil. *Cadernos de Saude Publica*, Fundacao Oswaldo Cruz, v. 35, 2019. ISSN 16784464.

BBC. 2021. Disponível em: <https://www.bbc.com/portuguese/brasil-57245848>.

BOTELHO, F.; SILVA, C.; CRUZ, F. Artigos de revisão epidemiologia explicada-análise de sobrevivência. 2009. Disponível em: <www.apurologia.pt>.

CAI, C. et al. smcure: An r-package for estimating semiparametric mixture cure models. *Computer methods and programs in biomedicine*, NIH Public Access, v. 108, p. 1255, 12 2012. ISSN 01692607. Disponível em: </pmc/articles/PMC3494798//pmc/articles/PMC3494798/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3494798/>.

CAI, C. et al. *smcure: Fit Semiparametric Mixture Cure Models*. [S.l.], 2022. R package version 2.1. Disponível em: <https://CRAN.R-project.org/package=smcure>.

COLOSIMO, E. A. et al. Análise de sobrevivência aplicada. *Revista Entreteses*, p. 77, 2006. Disponível em: <http://cursodegestaoelideranca.paginas.ufsc.br/files/2016/03/Apostila-OrientaÃğÃčo-ao-TCC.pdf>.

CORBIERE, F.; JOLY, P. A sas macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, Elsevier Ireland Ltd, v. 85, p. 173–180, 2007. ISSN 01692607. Disponível em: <https://www.researchgate.net/publication/6641628_A_SAS_macro_for_parametric_and_semiparametric_mixture_cure_models>.

KRUGER, D. H. et al. Hantaviruses–globally emerging pathogens. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology*, J Clin Virol, v. 64, p. 128–136, 3 2015. ISSN 1873-5967. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25453325/>.

OLIVEIRA, S. V. D.; DUARTE, E. C. Magnitude and distribution of deaths due to hantavirus in brazil, 2007-2015. *Epidemiol. Serv. Saude*, v. 27, p. 11, 2018.

OMS. 2020. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/zoonoses>.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

TEAM, R. D. C. *Download R-4.0.4 for Windows. The R-project for statistical computing.* 2021. Disponível em: <https://cran.r-project.org/bin/windows/base/>.

Terry M. Therneau; Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000. ISBN 0-387-98784-3.

THERNEAU, T. M. Survival analysis [r package survival version 3.4-0]. Comprehensive R Archive Network (CRAN), 8 2022. Disponível em: <https://CRAN.R-project.org/package=survival>.