UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

ANÁLISE DO TEMPO DE INTERNAÇÃO EM HOSPITAIS PELO SUS NO ESTADO DE SÃO PAULO NO PERÍODO DE 2018 A 2022.

Iniciação Científica

Discente: Leticia Tiago Maia Santos

leticia.tiago@unesp.br

RA: 211250041

Orientador: Mário Hissamitsu Tarumoto

mario.tarumoto@unesp.br

Presidente Prudente - SP

Outubro/2024

Resumo

A evolução tecnológica permitiu o aumento exponencial na geração e fluxo de dados, dificultando a análise e tomada de decisões o que nos leva a fazer um estudo mais detalhado e organizado a fim de tomar decisões mais assertivas. Entre os vários conjuntos de dados públicos existentes, no Brasil, o DATASUS é uma importante fonte de informações que pode ser usada como dados principais na construção de modelos. Este projeto foca na construção de modelos de análise de sobrevivência usando dados do SIHSUS, especificamente de internações por Neoplasias no estado de São Paulo entre 2018 e 2022, com o objetivo de identificar diferenças regionais nos tempos de internação. A análise foi realizada utilizando o software R-Studio, empregando o método de Máxima Verossimilhança para estimar os parâmetros e de acordo com a distribuição dos dados foi decidido que seria utilizado modelo paramétrico. O ajuste do modelo foi verificado por meio dos resíduos de Cox-Snell e da estimativa de Kaplan-Meier. Embora tenhamos rodado o modelo com as covariáveis idade, motivo e gênero, o modelo de regressão de Weibull apresentou um desempenho superior em termos de ajuste ao incluir apenas o intercepto.

Palavras-chave: Análise de sobrevivência; DataSus; R; Modelo Paramétrico; Kaplan-Meier.

Introdução

A evolução tecnológica dos últimos anos está possibilitando a geração cada vez maior e rápida de dados. Diariamente, muitas empresas geram grandes quantidades de dados e o fluxo de informação aumenta de maneira exponencial, o que pode dificultar muito a análise e uma tomada de decisão consciente e eficaz. As informações provêm das inúmeras fontes de dados e é preciso fazer um estudo mais detalhado e organizado a fim de direcionar o rumo do empreendimento. (https://blog.fortestecnologia.com.br/como_garantir-a-tomada-de-decisao/).

Existem várias formas de obtenção de dados. Algumas das formas podem ser automatizadas, como por exemplo, as informações de pessoas que acessam determinados conteúdos de internet, de veículos que passam por determinados locais, de pessoas que transitam por algumas ruas. Outras podem ser por meio de inserção de dados pelo usuário, como por exemplo, opinião dos usuários em relação a uma prestação de serviços ou a qualidade de produtos, ou ainda a opinião de proprietários de veículos. Existe ainda aquelas em que alguns órgãos são responsáveis pelo preenchimento de informações para subsidiar órgãos governamentais para tomada de decisões ou para disponibilizar informações ao público para atender a lei de acesso as informações.

Entre os vários conjuntos de dados públicos existentes, no Brasil, o DATASUS é uma importante fonte de informações que pode ser usada como meta dados ou dados principais na construção de modelos. Este sistema conhecido como Departamento de Informática do Sistema Único de Saúde (DATASUS) foi criado no início dos anos 90, buscando promover ações direcionadas a coleta, processamento e disseminação de informação sobre saúde. Toda produção de dados públicos relacionados a essa área, são gerenciados pelos sistemas de informações pertencentes ao DATASUS. Este departamento pertence à Secretaria Executiva do Ministério da Saúde (MS), tem como principal objetivo estruturar sistemas de informação, integrar dados em saúde e auxiliar na gestão dos diversos níveis de atenção em saúde.

Entre os vários sistemas administrativos existentes no DATASUS, neste projeto será utilizado os dados dos arquivos disseminâveis para tabulação do

Sistema de Informações Hospitalares do SUS (SIHSUS) que podem ser encontrados no endereço: http://www2.datasus.gov.br/ DATASUS. Estas bases de dados, por se tratar de informações a nível nacional, abrangendo todos os municípios da Federação, são bases muito grandes, necessitando de um tratamento de dados iniciais para posterior análise.

Neste projeto, o interesse principal, será o de construção de modelos de análise de sobrevivência. Uma das variáveis resposta a ser utilizada, é o tempo de internação em hospitais. Neste caso, a censura ocorre quando o indivíduo morre antes da alta. Desta forma, este projeto tem como o intuito, estudar os principais modelos de regressão aplicáveis em dados de tempos de vida. Os dados a serem utilizados serão os casos de internação no estado de São Paulo, no período de 2018 a 2022. Para este estudo, o objetivo é fazer um recorte dos dados, ou seja, serão selecionados somente as Neoplasias (CID - C00 a D48). Tem como intuito ainda o de verificar se existe diferenças entre as regiões de governo, em termos de ocorrência, tempos de internação e tipos de CIDs, levando-se em consideração as várias variáveis existentes na base de dados.

REVISÃO BIBLIOGRÁFICA

Nesta seção são apresentados alguns conceitos de Análise de Sobrevivência importantes para a realização do estudo, de forma que seja possível atingir o objetivo deste trabalho.

2.1 Análise de Sobrevivência

A Análise de Sobrevivência é um ramo da Estatística que pode ser aplicado também nas áreas de Medicina, Finanças e Engenharia chamada nesta última de Análise de Confiabilidade.

Nesta área a variável resposta de interesse, geralmente, é o tempo até a ocorrência de um evento de interesse, este é denominado tempo de falha.

Tempo pode ser anos, meses, semanas ou dias desde o início do acompanhamento de um indivíduo até a ocorrência de um evento. Nesse trabalho é medido em dias.

Por evento de interesse entendemos morte, incidência de doença, recaída da remissão, recuperação (por exemplo, retorno ao trabalho) ou qualquer experiência designada de interesse que possa acontecer a um indivíduo. No presente trabalho o evento de interesse é a morte do paciente

A principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Isto é, não se tem, nesses casos, o tempo decorrido até o evento de interesse, mas tem-se o tempo até a ocorrência da censura e, apesar de incompleta, essa informação é muito útil para a análise.

2.2 Distribuições de Probabilidade

Ao realizar modelagem para dados de sobrevivência, é essencial considerar a distribuição de probabilidade que a variável resposta pode ter.

Na Estatística, o modelo normal é frequentemente utilizado. No entanto, devido à característica particular da variável resposta, que só pode assumir valores positivos, utilizam-se distribuições que respeitem essa condição, como as distribuições Exponencial, Weibull, Log-logística e Log-normal, que são os casos paramétricos.

2.2.1 Distribuição Exponencial

A distribuição exponencial é um dos modelos probabilísticos mais simples usados para descrever o tempo de sobrevivência. Apresenta um único parâmetro e é a única que apresenta uma taxa de falha constante. A distribuição exponencial apresenta uma propriedade característica chamada de falta de

memória da distribuição exponencial, devido a sua taxa de falha ser constante o que significa que tanto uma unidade velha quanto uma nova, que ainda não falharam, apresentam a mesma taxa de falha em um intervalo futuro.

. Seja T uma variável aleatória referente ao tempo de falha. Se T segue uma distribuição exponencial, sua função de densidade é dada por:

$$f(t) = \frac{1}{\alpha}\, exp\left\{-\frac{t}{\alpha}\right\}, \qquad t \geq 0,$$

sendo α>0 o tempo médio de vida.

Ainda, as funções de sobrevivência, $S(t)$,, e de taxa de falha, $\lambda(t)$, são dadas, respectivamente, por:

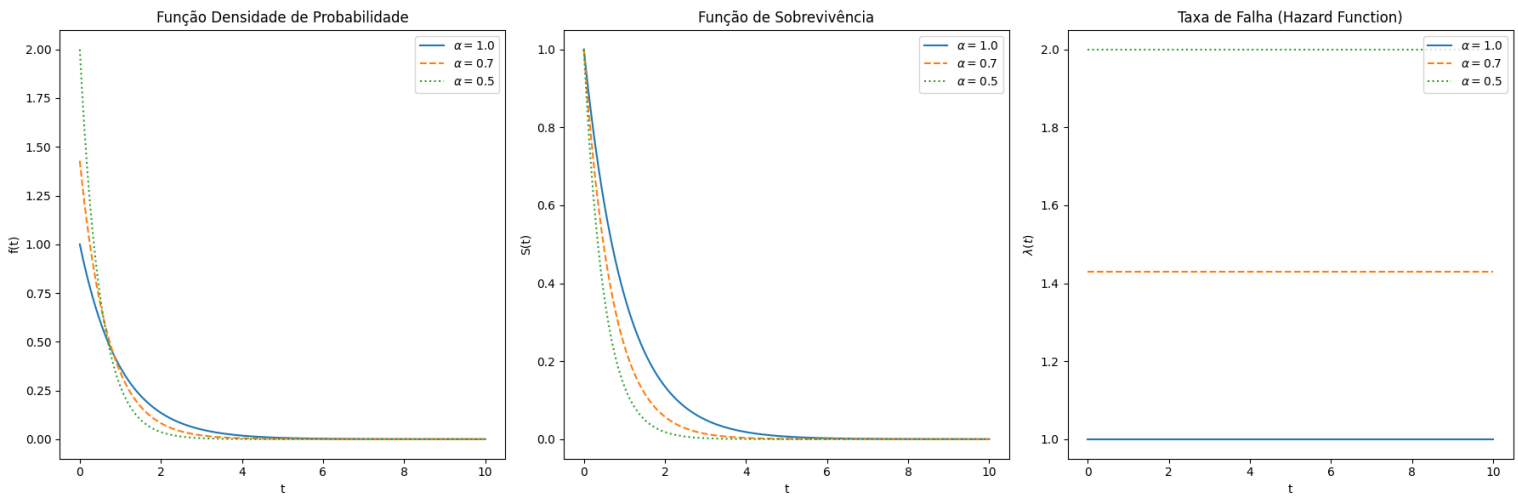
$$S(t) = \; exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$$

e

$$\lambda(t) = \frac{1}{\alpha}\; para\; t \geq 0$$

Na figura 4, temos, respectivamente, as funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição exponencial para $\alpha = 1{,}0\ (-), 0{,}7$ (--) e $0{,}5(\cdots)$.

Figura 1 – Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição exponencial.



Fonte: Autora (2024).

2.2.2 Distribuição Weibull

A distribuição de Weibull é frequentemente utilizada em estudos biomédicos e industriais. Isso se deve ao fato dela apresentar uma grande variedade de formas e uma função de taxa de falha monótona, isto é, ela é crescente, decrescente ou constante.

Seja T uma variável aleatória referente ao tempo de falha. Se T segue uma distribuição de Weibull, sua função de densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \qquad t \geq 0,$$

sendo γ o parâmetro de forma e α o parâmetro de escala, ambos positivos.

As funções de sobrevivência, $S(t)$ e de taxa de falha, $\lambda(t)$, são dadas, respectivamente, por

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$$

e

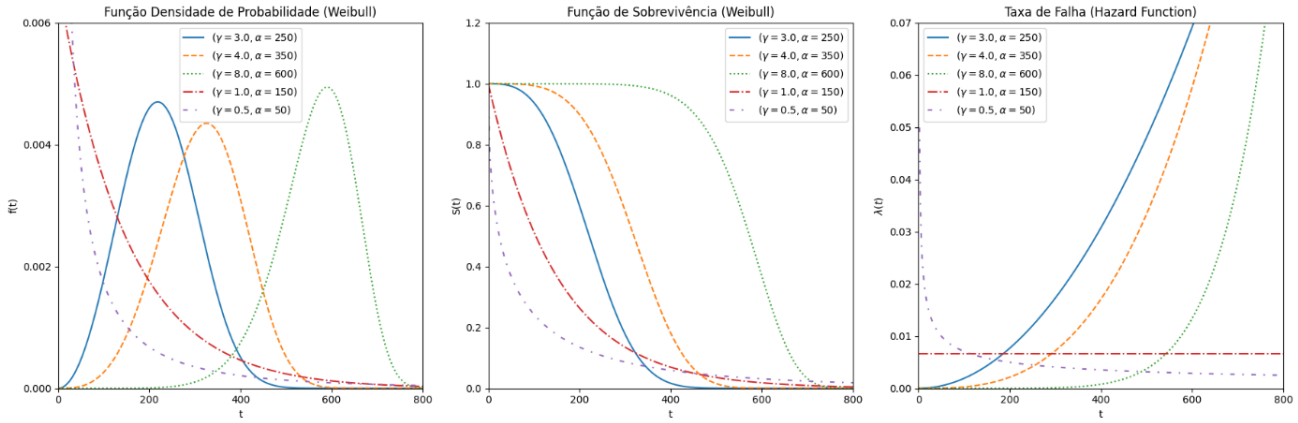
$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}$$

para $t \geq 0, \gamma > 0$ e $\alpha > 0$.

Quando o parâmetro de forma for igual a 1, tem-se como caso particular a distribuição Exponencial, com função de taxa de falha constante. Quando, essa função é crescente e para valores menores do, $\gamma > 1$, ela é decrescente.

Na figura 5, temos respectivamente, a forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull para alguns valores (mostrado na legenda da figura) dos parâmetros (γ, α).

Figura 2 – Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull.



Fonte: Autora (2024).

2.2.3 Modelo de Regressão Exponencial

O modelo de regressão exponencial é o modelo de regressão mais simples, com uma única covariável. Segundo Colosimo e Giolo (2006), a combinação de um componente determinístico e uma distribuição exponencial com média unitária para o erro $(f(\varepsilon) = exp\{-\varepsilon\})$ produz o modelo de regressão exponencial:

$$T = exp\{\beta_0 + \beta_1 x\}\varepsilon$$

Este modelo é linearizável se for considerado o logaritmo de T, escrito da forma:

$$Y = \log(T) = \beta_0 + \beta_1 X + \upsilon$$

onde $\upsilon = \log(\varepsilon)$. O erro υ segue uma distribuição do valor extremo padrão $(f(v) = exp\{v - exp\{v\}\})$. A função de sobrevivência para Y condicional a x é expressa por:

$$S(y|x) = exp\{-exp\{y - (\beta_0 + \beta_1 x)\}\}$$

Para T condicional a x, a função de sobrevivência correspondente é:

$$S(t|x) = exp\left\{-\left(\frac{t}{exp\{\beta_0 + \beta_1 x\}}\right)\right\}$$

2.2.4 Modelo de Regressão Weibull

O modelo de regressão Weibull, de acordo com David e Kleibaum (2012), é o modelo paramétrico mais utilizado. Assim como no modelo exponencial, será reparametrizado com coeficientes de regressão. O parâmetro adicional γ é chamado de parâmetro de forma e determina a forma da função de risco. A adição deste dá ao modelo de Weibull maior flexibilidade do que o modelo exponencial. Este modelo pode ser expresso como:

$$Y = \log(T) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \sigma v = \boldsymbol{x}'\boldsymbol{\beta} + \sigma v$$

em que $\boldsymbol{x}' = (1, x_1, \cdots, x_p)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$. Neste modelo, T deve ter uma distribuição de Weibull para que $\log(T)$ tenha uma distribuição do valor extremo com parâmetro de escala σ.

O modelo nessa forma também é conhecido como modelo de tempo de vida acelerado. Isto porque a função das covariáveis é acelerar ou desacelerar o tempo de vida. Ele pode ser expresso como:

$$T = exp\{\boldsymbol{x}'\boldsymbol{\beta}\}exp\{\sigma v\}$$

A função de sobrevivência para Y condicional a $\boldsymbol{x}$ é expressa por:

$$S(y|\boldsymbol{x}) = exp\left\{-exp\left\{\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma}\right\}\right\}$$

e para T condicional a $\boldsymbol{x}$:

$$S(t|\boldsymbol{x}) = exp\left\{-\left(\frac{t}{exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}\right)^{1/\sigma}\right\}$$

2.3 Análise de resíduos

A avaliação do ajuste do modelo é extremamente importante na análise de dados. O principal objetivo desse passo é examinar a adequação da distribuição escolhida para a variável resposta, verificar as suposições básicas do modelo, identificar pontos extremos, observar a relevância de possíveis fatores omitidos e analisar a forma funcional do modelo.

De acordo com Klein e Moeschberger (2003), essas técnicas devem ser aplicadas para rejeitar modelos claramente inadequados, em vez de “provar” que um modelo específico está correto.

Devido à presença de observações censuradas, uma maneira de se fazer a análise é utilizando os resíduos de Cox-Snell (1968), útil para examinar o ajuste global do modelo.

2.3.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell (1968) auxiliam no ajuste global do modelo e são quantidades determinadas por:

$$\widehat{e_i} = \widehat{H}(t_i|\boldsymbol{x}_i) \text{ ou } \widehat{e_i} = \widehat{H}(y_i|\boldsymbol{x}_i)$$

sendo que $\widehat{H}(\cdot)$ é a função de taxa de falha acumulada obtida do modelo ajustado, t_i os tempos de falha dos indivíduos e y_i a variável originada a partir da transformação $Y = \log(T)$. Para os modelos de regressão exponencial e Weibull, os resíduos de Cox-Snell são dados, respectivamente, por:

Exponencial: $\widehat{e_i} = \left[t_i exp\{-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\}\right]$

E Weibull: $\widehat{e_i} = \left[t_i exp\{-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\}\right]^{\widehat{\gamma}}$

Materiais e Métodos

Os dados analisados foram extraídos do SIHSUS para o estado de São Paulo com 12.322.623 de observações. As análises foram feitas utilizando o software R-Studio.

O primeiro passo para iniciar as análises foi filtrar esses casos em que a variável “DIAG_PRINC” referente ao diagnostico principal do paciente incluía apenas os casos de Neoplasias (CID – C00 a D48), ficando com 710.394 observações. Foi feita uma análise inicial e notou-se que havia diversas colunas com informações codificadas, foi preciso então enriquecer com os dados a partir da Plataforma de Ciência de Dados aplicada à Saúde (PCDAS) disponibilizada pela Fiocruz.

Em seguida foi feito um tratamento dos dados, algumas colunas como data de internação, data de saída e data de nascimento foram transformadas para o tipo de data (AAAAMMDD). A coluna SEXO, IDENT, COMPLEX, CEP, DIAS_PERM, CAR_INT, COBRANCA, MORTE foram enriquecidas de acordo com o que foi informado em PCDAS, com intuito de ter informação do que cada código significava. Na tabela 1 pode-se verificar a variável e sua respectiva descrição.

Tabela 1 – Descrição das variáveis presentes no conjunto de dados.

Variável	Descrição
SEXO	Sexo do paciente.
IDENT	Identificação do tipo da AIH.
COMPLEX	Complexidade do caso.
CEP	CEP do paciente.
DIAS_PERM	Dias de Permanência.
CAR_INT	Caráter da internação.
COBRANCA	Motivo de Saída/Permanência
MORTE	Indica Óbito ou não.
DIAG_PRINC	Código do diagnóstico principal (CID10).
DT_SAIDA	Data de saída, no formato aaaammdd.
DT_INTER	Data de internação no formato aaammdd.
NASC	Data de nascimento do paciente aaaammdd.
IDADE	Idade do paciente

Fonte: Autora (2024).

Para a análise exploratória dos dados, foram criadas algumas variáveis com o intuito de enriquecer nosso estudo. As variáveis criadas foram:

- **Faixa etária**: Criada a partir da variável **idade**, classificando os indivíduos em grupos etários.
- **Tempo de internação**: Calculado pela diferença entre a data de internação e a data de saída do hospital.
- **Censura**: Derivada da variável **morte**, onde:
    - 0: Indica "Com óbito", representando a censura do estudo.
    - 1: Indica "Sem óbito", representando a alta do paciente e a "falha" do estudo.

- **Região**: Criada a partir da variável **cep**, onde a região é identificada com base no primeiro dígito do código postal.
- **Motivo da internação**: Derivada da variável **car_int**, onde:
    - 0: Indica internação de urgência.
    - 1: Indica internação eletiva.
- **Gênero**: Criada a partir da variável **sexo**, onde:
    - 0: Indica gênero feminino.
    - 1: Indica gênero masculino.

Além desses tratamentos, foi tratado os casos em que as colunas possuíam nas. Outro filtro aplicado foi ignorar os casos em que a variável tempo ultrapassavam 100 dias, pois casos de neoplasias não há a necessidade de uma internação superior a 100 dias, na base possuía casos de 600 dias que foi considerado algum erro de digitação. Na base temos 636.858 casos sem óbito e 73.536 com óbito que é a nossa censura.

Resultados e Discussão

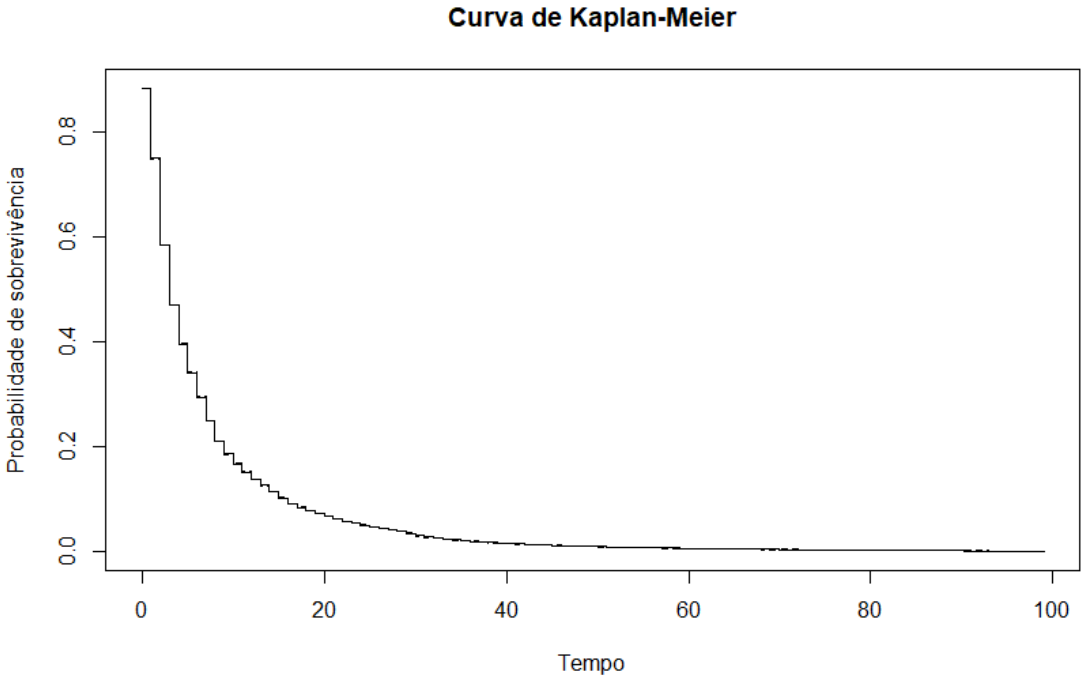
Neste tópico será abordado os resultados das análises. A primeira etapa será análise das variáveis categóricas e numéricas, a segunda será mostrada os resultados dos modelos feitos.

Análise Exploratória dos Dados

Na figura 3 podemos observar a curva de Kaplan-Meier o que mostra uma queda acentuada logo no início, indicando que muitos pacientes com neoplasias experimentam o evento (morte ou progressão da doença) nos primeiros períodos. Isso pode estar relacionado a casos mais agressivos ou avançados da doença. A estabilização da curva após cerca de 40 dias pode sugerir que os pacientes que sobrevivem além desse ponto têm uma chance relativamente

melhor de continuar sobrevivendo, indicando talvez um subgrupo de pacientes com neoplasias menos agressivas ou que respondem melhor ao tratamento.

Figura 3 – Gráfico da curva de Kaplan-Meier.



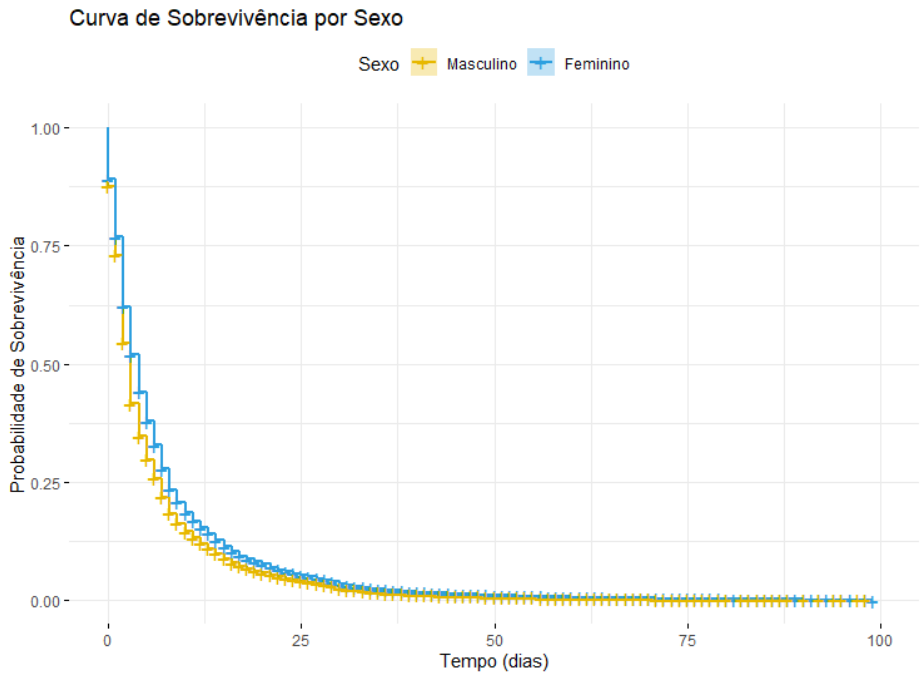
Fonte: Autora (2024).

Analisando o gráfico, investigamos a quantidade de óbitos entre pacientes internados por menos de 20 dias e aqueles com mais de 40 dias. Observamos que uma grande parte dos óbitos ocorre em um período relativamente curto: 65.355 pacientes faleceram em menos de 20 dias, enquanto apenas 1.305 morreram após 40 dias de internação. Isso sugere que a fase inicial da internação é crítica para a maioria dos pacientes. No total, a amostra analisada contém 710.394 pacientes, dos quais 75.536 vieram a óbito. Uma análise mais detalhada revela que 86,52% das mortes ocorreram antes dos 20 dias,

Quando olhamos para os perfis de internação, notamos que, entre os óbitos ocorridos antes de 20 dias, 57.991 foram de pacientes com internações de 'Urgência', enquanto 7.364 foram em caráter 'Eletivo'. Após 40 dias, esse padrão se mantém: 1.023 óbitos ocorreram em internações de 'Urgência' e apenas 282 em internações 'Eletivas'. Esses dados indicam que, para pacientes internados em situação de urgência, as chances de óbito nos primeiros dias são significativamente maiores.

Os tempos medianos de sobrevivência revelaram que pacientes internados por urgência (6,93 dias) têm maior tempo de sobrevivência em comparação aos internados por motivos eletivos (3,91 dias). Além disso, o tempo mediano de sobrevivência foi ligeiramente maior para mulheres (5,76 dias) do que para homens (4,95 dias), sugerindo que o grupo feminino tem uma probabilidade de sobrevivência um pouco maior ao longo do tempo. Como podemos observar na curva de Kaplan-Meier por sexo na figura 4.

Figura 4 – Gráfico da curva de Kaplan-Meier por sexo.

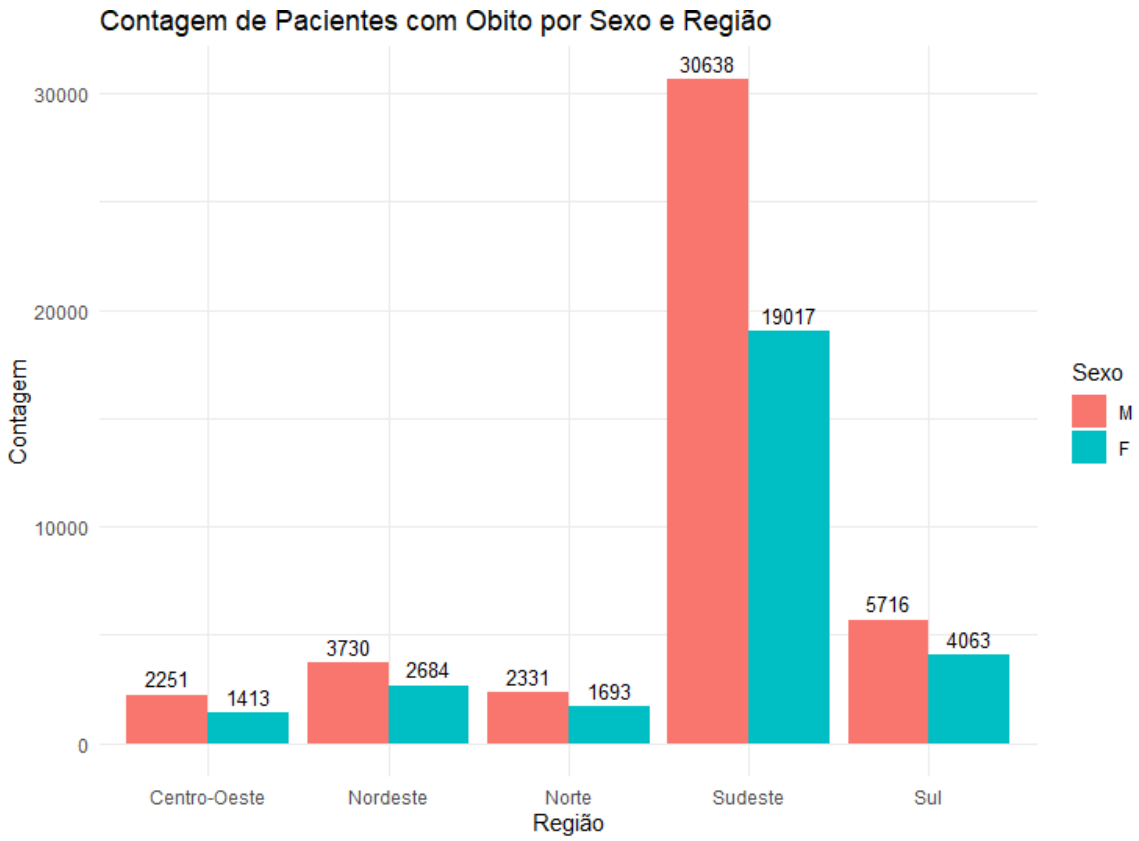


Fonte: Autora (2024).

Nos resultados obtidos, observamos que o sexo masculino é ligeiramente mais acometido por neoplasias, representando 52,7% dos casos, com um total de 374.717 pacientes, em comparação com 335.677 pacientes do sexo feminino. Além disso, o percentual de óbitos é maior entre os homens (11,9%) do que entre as mulheres (8,6%).

Em relação às regiões do Brasil, o Sudeste registrou o maior número de mortes do sexo masculino, com 30.638 óbitos em homens (6,54%) e do sexo feminino a região com maior óbito foi Sul 4,57% em mulheres. Esse padrão se mantém, embora com menor intensidade, nas outras regiões. Na figura 5 pode ser observado a quantidade de acometidos por neoplasia por sexo.

Figura 5 – Contagem de pacientes com óbito por sexo e região.



Fonte: Autora (2024).

Apesar da base de dados incluir apenas pacientes do estado de São Paulo, foram identificados casos de pessoas de outras regiões. Isso pode indicar que muitos pacientes vêm a São Paulo em busca de tratamento, possivelmente devido à melhor infraestrutura hospitalar e à presença de especialistas renomados em oncologia no estado.

Nesse estudo foi observado 706 diferentes tipos de CIDs e dentre esses os que mais apareceram foi o D259 referente a Leiomioma do útero, não especificado com 42869 casos, C61 referente a Neoplasia Maligna da Próstata com 39332 casos e C20 com 30312 casos. Na tabela 2 podemos observar que os casos de Neoplasia maligna do cólon (C20), há mais casos de "Sem óbito" em ambos os sexos. No entanto, os homens (15.884) superam as mulheres (11.685) tanto em sobrevivência quanto em óbito, com uma contagem ligeiramente maior de óbitos entre homens (1.546) em comparação com as mulheres (1.197). A Neoplasia maligna da próstata (C61) diagnóstico que afeta

exclusivamente homens, com um número considerável de casos "Sem óbito" (35.384) em relação aos óbitos (3.948), observamos que é a maior proporção entre sobrevivente e óbitos (11,15 %). E por fim Tumor benigno do útero (D259) que é uma neoplasia que acomete apenas mulheres. A grande maioria não evoluiu para óbito, com 42.849 casos "Sem óbito" e apenas 20 casos "Com óbito", sugerindo uma baixa taxa de mortalidade associada a esse diagnóstico.

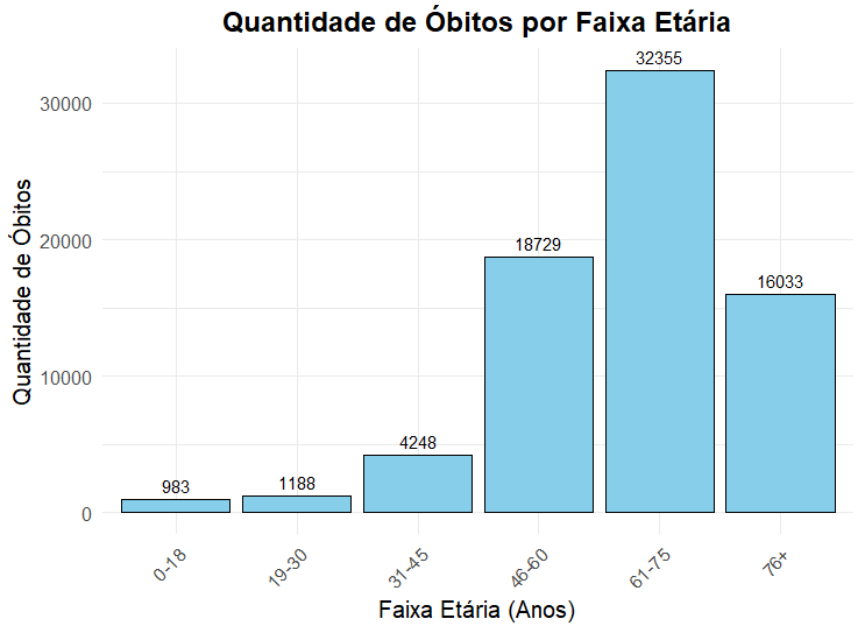
Tabela 2 – Relação dos CIDs com mais incidência por sexo óbito e sem óbito.

CID	SEXO	Sem óbito	Com óbito	Percentual
C20	M	15884	1546	9,73%
C20	F	11685	1197	10,24%
C61	M	35384	3948	11,15%
D259	F	42849	20	0,04%

Fonte: Autora (2024).

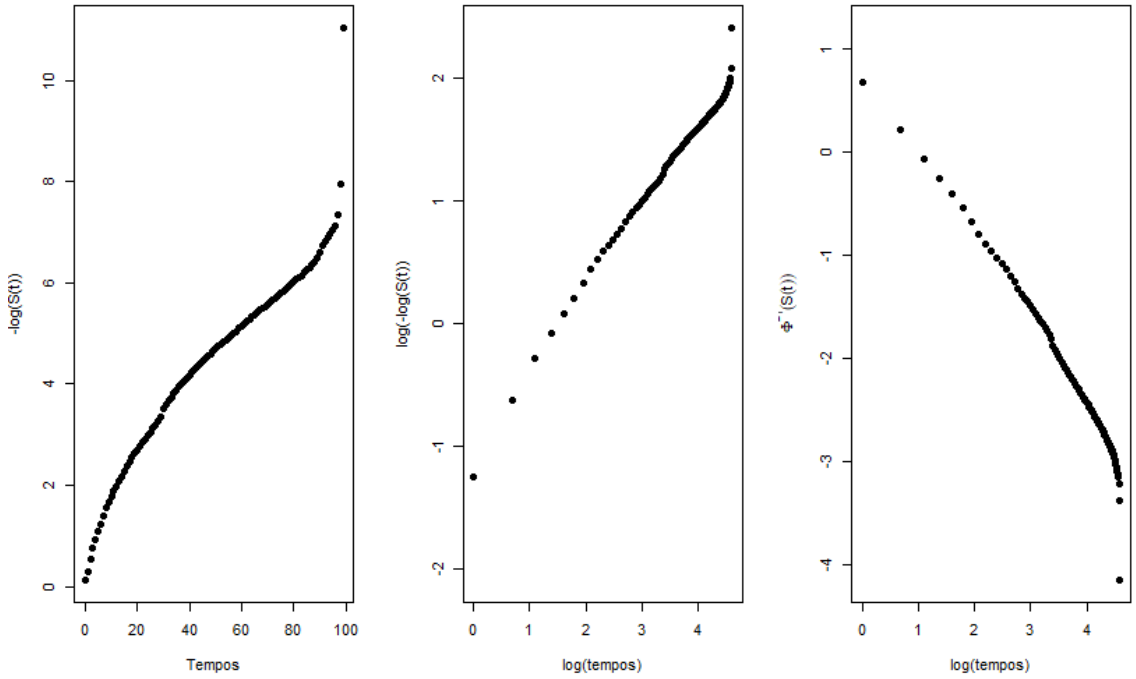
O CID que mais matou foi C349 referente à neoplasia maligna dos brônquios ou pulmões com 5257 casos, C61 (Câncer de Próstata) com 3948 e C169 referente à neoplasia maligna do estômago, não especificado com 2914 casos. Na figura 6 podemos observar que os óbitos estão concentrados na faixa etária de 46 anos em diante.

Figura 6 – Quantidade de óbitos por faixa etária.



Fonte: Autora (2024).

Figura 7 – Gráficos $t\ x - log\left(\hat{S}(t)\right), log\ t\ x\ log\left(-log\left(\hat{S}(t)\right)\right)$ $e\ log\ t\ x\ \phi^{-1}\left(\hat{S}(t)\right)$



Fonte: Autora (2024).

As distribuições exponencial e de Weibull, que pode ser observado na Figura 7, apresentam-se visualmente como as melhores candidatas, dentre as consideradas, para a análise dos dados desse estudo. Foram obtidas as estimativas dos parâmetros apresentadas nas Tabela 3 e 4 respectivamente.

Tabela 3 – Estimativas dos parâmetros Regressão Exponencial.

Regressão Exponencial			
	Valor	EP	Pvalor
Intercepto (Beta0)	1,95	0,004126	<2e-16
Idade	0,002	0,000062	<2e-16
Motivo Internação	-0,732	0,002542	<2e-16
Gênero	0,14	0,002515	<2e-16

Fonte: Autora (2024).

Tabela 4 – Estimativas dos parâmetros Regressão Weibull.

Regressão Weibull			
	Valor	EP	Pvalor
Intercepto (Beta0)	1,858049535	0,00594	<2e-16
Idade	0,002935645	0,0000896	<2e-16
Motivo Internação	-0,795784665	0,00386	<2e-16
Gênero	0,140869122	0,00114	<2e-16

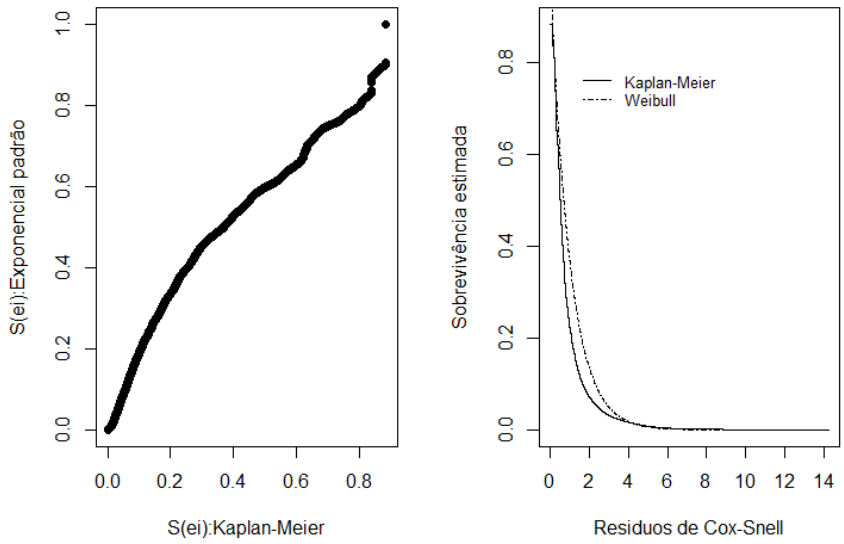
Fonte: Autora (2024).

No coeficiente Idade é positivo e indica uma tendência que à medida que a idade aumenta, o tempo de sobrevivência pode aumentar, mantendo as outras variáveis constantes. A variável motivo internação é negativo, indica que urgência e eletivo que são as categorias de motivo internação, tem um impacto negativo significativo no tempo de sobrevivência. A covariável gênero também é positiva indicando associação positiva. Notamos que o pvalor são baixos indicando que são altamente significativas no modelo, ou seja, têm um efeito estatisticamente significativo no tempo de sobrevivência.

Temos para a regressão exponencial *log-likelihoods* do modelo intercepto= -1712021 e do modelo completo= -1750308, temos que um valor de *log-likelihood* maior (menos negativo) indica um melhor ajuste do modelo. Neste caso, o modelo com as covariáveis ajusta melhor os dados do que o modelo com apenas o intercepto.

Para a regressão Weibull, temos que o valor para gama é $\gamma = 0{,}83375$ indicando que o risco de falha diminui com o tempo. Isto é, os óbitos têm maior probabilidade de ocorrer no início do período de observação, e o risco vai decrescendo à medida que o tempo passa.

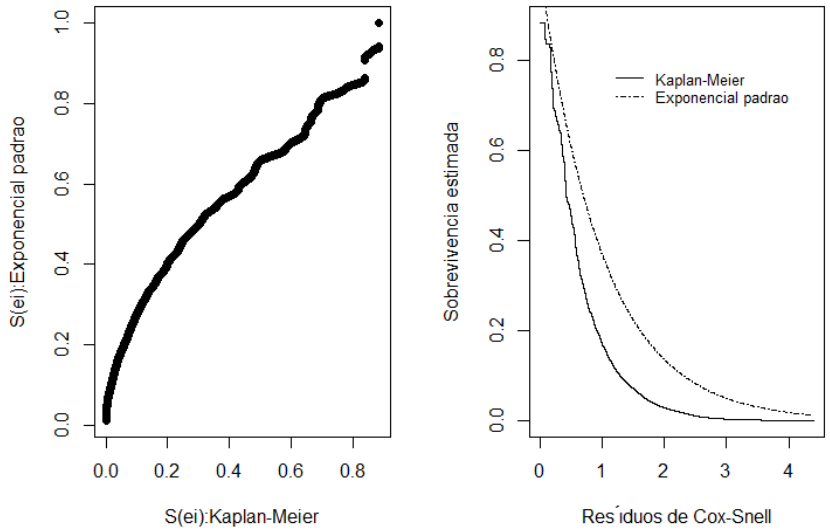
Figura 8 - Análise dos resíduos de Cox-Snell do modelo de regressão weibull ajustado para os dados de neoplasia com covariáveis.



Fonte: Autora (2024).

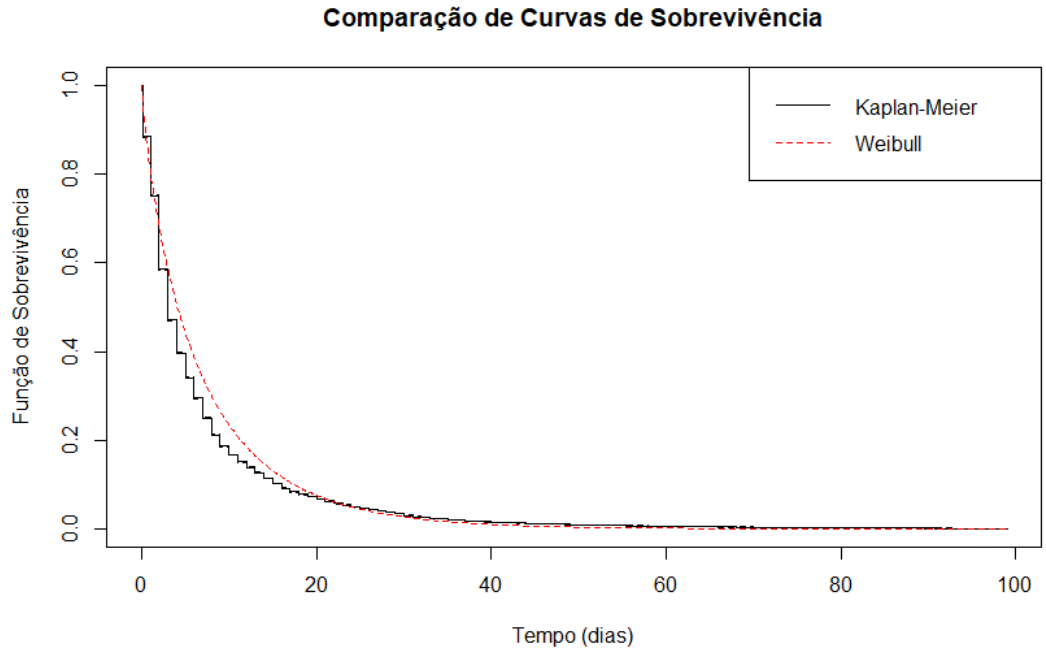
Se o modelo for adequado, esses resíduos, devem ser considerados como provenientes de uma amostra aleatória da distribuição exponencial padrão. Assim, as estimativas das curvas de sobrevivência desses resíduos, obtidas por Kaplan-Meier $\left(\hat{S}\left(\widehat{e_{\iota_{KM}}}\right)\right)$e pelo modelo exponencial padrão $\left(\hat{S}\left(\widehat{e_{\iota_{EXP}}}\right)\right)$, devem estar próximas. Além disso, o gráfico dos pares de pontos $\left(\hat{S}\left(\widehat{e_{\iota_{KM}}}\right), \hat{S}\left(\widehat{e_{\iota_{EXP}}}\right)\right)$, deve formar aproximadamente uma reta, para que o modelo ajustado possa ser considerado satisfatório. A Figura 8, que apresenta ambos os gráficos mencionados, mostra que o modelo Weibull parece aceitável, mas ainda não perfeito. Enquanto na Figura 9, temos a mesma comparação para o modelo de regressão exponencial e tivemos um ajuste ruim.

Figura 9 - Análise dos resíduos de Cox-Snell do modelo de regressão exponencial ajustado para os dados de neoplasia com covariáveis.



Podemos notar que embora tenhamos rodado o modelo com as covariáveis idade, motivo e gênero, o modelo de regressão de Weibull apresentou um desempenho superior em termos de ajuste ao incluir apenas o intercepto, conforme mostrado na Figura 10.

Figura 10- Curvas de sobrevivência estimadas pelos modelos de Weibull versus a curva de sobrevivência estimada por Kaplan-Meier apenas com intercepto.



Fonte: Autora (2024).

Conclusões

Este estudo revelou que os tipos de neoplasias com maior incidência são o leiomioma do útero, que afeta exclusivamente mulheres, a neoplasia maligna da próstata, restrita aos homens, e a neoplasia maligna do cólon, que acomete ambos os sexos. No entanto, o câncer de próstata foi responsável pela maior taxa de óbitos, com 11,15% das mortes quando comparado ao total de óbitos. Em contrapartida, o leiomioma do útero, apesar de ser o tipo mais incidente, apresentou a menor taxa de mortalidade, com apenas 0,04%. A neoplasia maligna do cólon mostrou maior incidência em homens, porém com uma menor taxa de mortalidade entre as mulheres.

Outro ponto relevante foi a identificação da faixa etária de maior risco. Pacientes entre 61 e 75 anos representaram a maioria dos óbitos. Cerca de 86,52% das mortes ocorreram antes dos primeiros 20 dias de internação. E destes, 88,73% eram classificados como casos de 'Urgência', o que destaca a importância de cuidados intensivos imediatos nesse período crítico.

Embora o estudo tenha se focado no estado de São Paulo, foi possível identificar pacientes de outras regiões, especialmente do Sul do Brasil, que somaram cerca de 10 mil pessoas. Esse dado sugere que São Paulo é um polo de referência em tratamentos oncológicos, atraindo pacientes de todo o país, provavelmente devido à excelência de seus hospitais e ao nível de especialização médica disponível.

Insights importantes que surgiram deste estudo incluem a necessidade de uma resposta rápida em casos de internação por 'Urgência', dado o elevado número de óbitos em até 20 dias. Além disso, a baixa mortalidade associada ao leiomioma do útero, apesar de sua alta incidência, pode indicar sucesso em tratamentos ou diagnósticos precoces, o que poderia ser mais investigado para aplicar aos outros tipos de neoplasias.

Os próximos passos envolvem expandir o estudo para outras regiões do Brasil, a fim de verificar se essas tendências se repetem em outras populações. Também seria interessante investigar os fatores que contribuem para a baixa mortalidade em neoplasias como o leiomioma e explorar mais a fundo o impacto

do tempo de internação sobre as taxas de óbito. Por fim, a aplicação de modelos preditivos mais complexos, como o de Cox, pode proporcionar comparações mais robustas com os resultados da regressão de Weibull, oferecendo novas perspectivas sobre a dinâmica da sobrevivência em diferentes tipos de neoplasia.

Agradecimentos

Agradeço ao meu orientador Prof Dr. Mário Hissatmitsu Tarumoto por todo o apoio e à Fundacte por acreditar no trabalho.

Referências

Brasil, 2020. A estrada para a transformação digital do SUS: Book das realizações. Disponível em: https://datasus.saude.gov.br/wp-content/uploads/2020/05/DATASUS-29-ANOS-Book-das-realiza%C3%A7%C3%B5es-de-2019-a-2020-A-Estrada-para aTransforma%C3%A7%C3%A3o-Digital-do-SUS.pdf. Data de acesso: 24 /05/2024.

BUSSAB, Wilton; MORETTIN, Pedro. Estatística Básica. 9 ed. Editora Saraiva. 2017.

COLOSIMO, E; GIOLO, S.R. Análise de Sobrevivência Aplicada. Edgard Blucher, São Paulo. 2006.

JOHNSON, Richard; WICHERN, Dean. Applied Multivariate Statistical Analysis. 6 ed. Pearson Education, Inc. 2018.

LAWLESS, J.F. Statistical models and methods for lifetime data, John Wiley & Sons. 1982

NELSON, W. Applied Life Data Analysis, Wiley, New York, 2003.

ELANDT-JOHNSON, R.C.; JOHNSON, N.L. Survival Models and Data Analysis.

Wiley, 2014

LEE, E.T.; WANG, J. Statistical Methods for Survival Data Analysis. 2003.

FioCruz, 2019. Sistema de Informações Hospitalares do SUS – SIHSUS. Disponível em: https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacoes-hospitalares-do-sus-sihsus/documentacao/. Data de acesso: 25/05/2024