UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Igor de Sousa Gonçalves

Criação de painel com dados da FOSP: uma abordagem com análise de sobrevivência

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Igor de Sousa Gonçalves.

Título do trabalho: Criação de painel com dados da FOSP: uma abordagem com análise de sobrevivência.

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Eder Angelo Milani**, **Professor do Magistério Superior**, em 08/12/2025, às 23:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Igor De Sousa Gonçalves**, **Discente**, em 09/12/2025, às 23:59, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5840512** e o código CRC **2F6B0252**.

Referência: Processo nº 23070.060283/2025-87 SEI nº 5840512

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Igor de Sousa Gonçalves

Criação de painel com dados da FOSP: uma abordagem com análise de sobrevivência

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Goiás para aprovação no componente curricular TCC, como parte das exigências para a obtenção do título de bacharel em Estatística. **Orientador**: Eder Angelo Milani

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Gonçalves, Igor de Sousa
Criação de painel com dados da FOSP [manuscrito] : uma abordagem com análise de sobrevivência / Igor de Sousa Gonçalves. - 2025.
43 f.: il.

Orientador: Prof. Dr. Eder Angelo Milani.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.
Bibliografia.
Inclui mapas, fotografias, gráfico, lista de figuras.

1. Análise de sobrevivência. 2. Kaplan-Meier. 3. Dashboards. 4. Fundação Oncocentro de São Paulo. I. Milani, Eder Angelo, orient. II. Título.

CDU 5

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e oito dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado "Criação de painel com dados da FOSP: uma abordagem com análise de sobrevivência", de autoria de Igor de Sousa Gonçalves, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pelo Prof. Dr. Eder Angelo Milani com a participação dos demais membros da Banca Examinadora: Fabiano Fortunato Teixeira dos Santos (IME/UFG) e Marcílio Ramos Pereira Cardial (IME/UFG). Após a apresentação, a banca examinadora realizou a arguição do estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 8,8, tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Marcilio Ramos Pereira Cardial**, **Professor do Magistério Superior**, em 06/12/2025, às 10:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Fabiano Fortunato Teixeira Dos Santos**, **Professor do Magistério Superior**, em 08/12/2025, às 13:52, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Eder Angelo Milani**, **Professor do Magistério Superior**, em 08/12/2025, às 23:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5801942** e o código CRC **C1313405**.

Referência: Processo nº 23070.060283/2025-87 SEI nº 5801942

Agradecimentos

A minha mãe Sônia que é minha rainha e meu exemplo de mulher, ela que sempre me apoiou e me incentivou em todas as decisões em minha vida mas sempre deixando sua opinião como mãe para guiar seu filho, ela que sempre esteve ali nos meus altos e baixos não importando o momento.

Ao meu irmão Breno que é meu exemplo como homem, que assim como minha mãe me ensinou e ensina muito até hoje, ele sempre esteve disponível para me amparar, não importava o que acontecesse ele sempre esteve ali para cuidar de mim e me proteger desde que eu nasci.

A minha namorada Izadora, que chegou em minha vida no mesmo momento da faculdade e que acompanhou toda a minha trajetória desde o inicio, cuidando de mim nos momentos de fraqueza e desespero, mas também estava la nos momentos de alegria e comemoração.

Não é possível descrever em palavras o quão grato eu sou a vocês três por tudo que fizeram e fazem por mim todo esse tempo, talvez vocês não tenham noção do quão importante todos vocês são em minha vida e eu não sei o que seria de mim se tivesse que passar por tudo que passei sem algum de vocês.

Ao meu orientador Eder, não só de TCC como de IC também, que confiou em mim e acreditou que eu era capaz, que mostrou que é possível ensinar tanto em sala de aula quanto na orientação de uma forma leve e divertida, não vejo ele mais como só um professor, vejo como um amigo.

Por fim, obrigado a todos vocês por todo apoio.

Resumo

A Fundação Oncocentro de São Paulo disponibiliza publicamente vários conjuntos de dados, porém, esses não estão em uma forma de fácil visualização, sendo necessário um tratamento desses dados para melhor visualização e compreendimento das informações. Uma forma mais elegante de se apresentar esses dados e que deixa-os de uma forma facilmente entendível é através de painéis ou *dashboards*. O nosso trabalho em questão, extraiu os dados da FOSP e os integrou a um *dashboard* interativo, para assim, proporcionar ao usuário um melhor entendimento e visualização das informações. Para a criação desse *dashboard*, foi utilizado o pacote Shiny do software Rstudio. A versão atual do *dashboard* contem 3 abas, sendo elas: *Dashboard*, Análise de Sobrevivência e Sobre. Na aba *Dashboard*, o usuário pode escolher entre as variáveis: escolaridade, idade, sexo, UF de residência, categoria de atendimento, radioterapia, quimioterapia, cirurgia e idade mediana, então o software exibirá as estatísticas descritivas sobre a variável que foi escolhida. Na aba Análise de Sobrevivência a função de sobrevivência é obtida a partir do estimador de Kaplan-Meier, sendo possível escolher o tipo de Classificação Internacional de Doenças (CID) e também pode-se filtrar os dados utilizando uma variável, sendo elas: sexo, idade mediana, categoria de atendimento, radioterapia, quimioterapia e cirurgia. Além disso, foram implementados modelos de regressão de Cox e Weibull, permitindo avaliar o impacto das covariáveis sobre o risco de morte e compreender o comportamento da função de risco ao longo do tempo. O modelo de Cox, de natureza semiparamétrica, possibilita investigar a influência das covariáveis sem assumir uma forma específica para a função de risco basal, enquanto o modelo de Weibull, de caráter paramétrico, permite descrever explicitamente a variação do risco em função do tempo. Por fim, na aba Sobre é dada uma breve explicação sobre o software, além de explicar sobre o intervalo de tempo que foi usado para as análises.

Palavras-chave: Análise de Sobrevivência, Kaplan-Meier, Dashboards, Fundação Oncocentro de São Paulo.

Abstract

The Fundação Oncocentro de São Paulo publicly provides several datasets related to cancer; however, these datasets are not presented in an easily interpretable format, requiring preprocessing and organization for better understanding. An elegant and intuitive way to display this information is through interactive panels, or *dashboards*. This study extracted data from FOSP and integrated them into an interactive *dashboard* to provide users with a clearer and more dynamic visualization of the information. The dashboard was developed using the *Shiny* package in the RStudio software. The current version contains 3 main tabs: *Dashboard*, Análise de Sobrevivência and Sobre. In the *Dashboard* tab, users can select variables such as escolaridade, idade, sexo, UF de residência, categoria de atendimento, radioterapia, quimioterapia, cirurgia e idade mediana, and the software displays descriptive statistics for the chosen variable. In the *Análise de Sobrevivência* tab, the survival function is estimated using the Kaplan-Meier method, allowing the selection of International Classification of Diseases (ICD) and filtering by variables such as sexo, idade mediana, categoria de atendimento, radioterapia, quimioterapia e cirurgia. Additionally, Cox and Weibull regression models were implemented to assess the impact of covariates on the risk of death and to understand the behavior of the hazard function over time. The Cox model, of semiparametric nature, allows the study of covariate effects without specifying the baseline hazard function, while the Weibull model, of parametric nature, enables explicit modeling of the risk variation over time. Finally, the *Sobre* tab provides a brief explanation of the software and details regarding the time period used in the analyses.

Keywords: Análise de Sobrevivência, Kaplan-Meier, Dashboards, Fundação Oncocentro de São Paulo.

Lista de figuras

Sumário

Introdução

O termo "câncer" abrange um conjunto de mais de 100 doenças distintas, cuja principal característica é o crescimento celular desordenado. Esse processo, conhecido como malignidade, envolve a invasão de tecidos e órgãos adjacentes. Adicionalmente, essas células podem se disseminar para locais distantes, formando metástases. A divisão rápida e incontrolável dessas células resulta no acúmulo de neoplasias malignas, que são agressivas. Isso as diferencia de tumores benignos, que são massas localizadas, de multiplicação lenta e com células semelhantes ao tecido de origem (MARQUES, 2016).

No âmbito nacional, o Instituto Nacional de Câncer (INCA) (Instituto Nacional de Câncer, 2025) é a principal instituição responsável pela coordenação das ações de controle do câncer no Brasil. Vinculado ao Ministério da Saúde, o INCA atua em diversas frentes, incluindo a prevenção, o diagnóstico, o tratamento, a formação de profissionais e o desenvolvimento de pesquisas científicas. Além disso, o Instituto é responsável pela elaboração de estimativas de incidência e mortalidade por câncer no país, bem como pela padronização e monitoramento dos registros hospitalares de câncer (RHC) em nível nacional. Dessa forma, o INCA desempenha um papel estratégico na consolidação das políticas públicas de atenção oncológica, fornecendo diretrizes e dados essenciais para subsidiar ações de vigilância e planejamento em saúde.

O câncer é um dos principais problemas de saúde pública no Brasil, tanto pela sua incidência quanto pelo impacto social e econômico que provoca. No estado de São Paulo, a Fundação Oncocentro de São Paulo (FOSP) (Fundação Oncocentro de São Paulo, 2024) desempenha um papel essencial ao reunir e disponibilizar registros hospitalares de câncer, constituindo uma base de dados ampla e valiosa para estudos epidemiológicos, clínicos e estatísticos. No contexto da saúde, a ciência de dados é utilizada para extrair conhecimento e auxiliar na gestão da informação. Isso é feito através da coleta, do armazenamento e da análise sistemática de grandes volumes de dados (RIBEIRO *et al.*, 2025).

Contudo, apesar da relevância e da crescente transparência desses repositórios, como o da FOSP, persiste uma lacuna entre a disponibilidade dos dados e sua utilização prática. A complexidade, o volume e a falta de mapeamento sobre a estrutura e interoperabilidade dessas bases epidemiológicas dificultam o acesso direto por parte de profissionais da saúde, gestores e pesquisadores. A simples disponibilização dos registros em formato tabular restringe o uso estratégico dos dados para subsidiar políticas públicas, limitando o potencial de abordagens baseadas em Big Data na vigilância em saúde (SHIMAOKA *et al.*, 2025).

Nesse cenário, ferramentas de visualização interativa de dados, como painéis (*dashboards*), surgem como alternativas poderosas para aproximar a informação da prática. Ao transformar grandes bases em representações visuais e interativas, esses *dashboards* permitem identificar

padrões, comparar grupos e facilitar a compreensão por diferentes públicos, independentemente de sua familiaridade com métodos estatísticos avançados (MEDEIROS; NOGUEIRA, 2023).

O presente trabalho se insere nesse contexto, propondo a construção de um painel interativo baseado nos dados da FOSP. Mais do que apenas apresentar estatísticas descritivas, buscou-se implementar recursos analíticos, como a análise de sobrevivência, que ampliam a capacidade de investigação sobre os fatores associados à evolução clínica dos pacientes com câncer em São Paulo. Assim, pretende-se contribuir para tornar os dados mais acessíveis e úteis, favorecendo tanto a pesquisa acadêmica quanto o suporte à gestão em saúde.

Além desta seção introdutória, este trabalho está estruturado da seguinte forma: o Capítulo 1 apresenta a fundamentação teórica, abordando os principais conceitos relacionados ao resumo dos dados e à Análise de Sobrevivência, incluindo o estimador de Kaplan-Meier e os modelos de regressão de Cox e Weibull. O Capítulo 2 descreve a metodologia adotada e o processo de desenvolvimento do *dashboard* interativo, detalhando as etapas de tratamento dos dados e implementação no ambiente *Shiny*. O Capítulo 3 apresenta e discute os resultados obtidos por meio da aplicação desenvolvida, enquanto a conclusão reúne as conclusões do estudo, as limitações encontradas e as perspectivas para trabalhos futuros.

1 Fundamentação Teórica

1.1 Resumo dos Dados

A eficiência na coleta de dados tem melhorado com o decorrer do tempo, juntamente com o aumento na capacidade de armazenamento dos computadores, o que implica diretamente em uma base de dados maior, acarretando assim em mais informações disponíveis para possíveis análises, porém, também significa que é uma base mais complexa para a obtenção de informação mais específicas.

A visualização de dados constitui uma etapa fundamental na análise estatística, pois permite explorar, compreender e comunicar informações complexas de maneira clara e intuitiva. Por meio de tabelas, representações gráficas e medidas descritivas, é possível identificar padrões, tendências, assimetrias e possíveis inconsistências nos dados, auxiliando tanto na etapa exploratória quanto na interpretação dos resultados obtidos por modelos estatísticos.

As tabelas desempenham um papel essencial na organização e síntese dos dados, devendo ser construídas de forma simples, clara, objetiva e autoexplicativa, de modo que permitam ao leitor compreender rapidamente as informações apresentadas sem necessidade de recorrer a explicações adicionais. Já os gráficos, por sua vez, possibilitam uma interpretação visual imediata das relações entre variáveis e dos comportamentos observados nos dados. Para que cumpram adequadamente esse propósito, devem sintetizar grandes volumes de informação em elementos visuais que favoreçam o raciocínio analítico, preservando a integridade dos dados e evitando distorções interpretativas.

Em contextos de saúde e epidemiologia, a utilização de recursos gráficos é especialmente relevante, uma vez que facilita a compreensão de fenômenos relacionados à incidência, prevalência e mortalidade em populações específicas, permitindo a comunicação mais eficaz dos resultados e o apoio à tomada de decisões baseadas em evidências.

No contexto da Análise de Sobrevivência, a visualização é utilizada tanto para a descrição das características demográficas e clínicas dos indivíduos quanto para a representação gráfica de funções de sobrevivência e risco. Histogramas, gráficos de barras e curvas de Kaplan–Meier são exemplos de ferramentas amplamente empregadas para investigar a distribuição das variáveis e o comportamento do tempo até o evento de interesse.

Além disso, a integração entre métodos estatísticos e ferramentas computacionais interativas tem ampliado as possibilidades de análise. O uso de tecnologias como o pacote *Shiny*, do software R, permite construir interfaces dinâmicas em que os usuários podem selecionar variáveis, ajustar modelos e visualizar resultados de forma imediata. Essa abordagem contribui para a democratização do acesso aos resultados e para uma interpretação mais acessível das

informações, sem comprometer o rigor estatístico.

Dessa forma, a visualização de dados não apenas complementa a análise quantitativa, mas também atua como um instrumento essencial para a comunicação científica, transformando resultados numéricos em representações gráficas que facilitam o processo decisório e a compreensão dos fenômenos estudados.

1.2 Análise de Sobrevivência

A Análise de Sobrevivência é um conjunto de métodos estatísticos cujo objetivo é estudar o tempo até a ocorrência de um determinado evento de interesse. Essa área da estatística é amplamente empregada em estudos clínicos, atuariais e de confiabilidade, permitindo investigar o impacto de diferentes fatores sobre o tempo até a ocorrência do evento. No contexto deste trabalho, o evento de interesse é o óbito de pacientes diagnosticados com câncer, e o tempo é medido em dias desde a data do diagnóstico até o evento ou até o término do acompanhamento.

Uma característica fundamental que diferencia a análise de sobrevivência de outras técnicas estatísticas é a presença de observações censuradas, isto é, indivíduos cujo tempo de falha não é completamente observado. Segundo Colosimo e Giolo (2006), a censura do tipo à direita é a mais comum e ocorre quando o tempo de acompanhamento se encerra antes da ocorrência do evento, seja por término do estudo, perda de seguimento ou manutenção do paciente vivo até o final do período analisado, essa foi a censura adotada neste estudo. Ignorar as observações censuradas pode introduzir viés nas estimativas, resultando em subestimação da função de sobrevivência.

Seja T uma variável aleatória contínua e não negativa representando o tempo até o evento, a quantidade de maior interesse é a Função de Sobrevivência, $S(t)$, definida como a probabilidade de um indivíduo ter um tempo de vida superior a um instante t. Matematicamente, ela é expressa por:

$$S(t) = P(T > t).$$

A função de sobrevivência descreve a proporção esperada de indivíduos que ainda não sofreram o evento até o tempo t. Como $S(0) = 1$ e $S(t)$ é uma função não crescente, ela tende a zero à medida que o tempo aumenta. Outra função relevante é a função densidade de probabilidade $f(t)$, relacionada à função de sobrevivência por:

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)\, du,$$

onde $F(t)$ é a função de distribuição acumulada do tempo até o evento.

Complementarmente, define-se a função de taxa de falha ou função de risco, que expressa a taxa instantânea de ocorrência do evento em um tempo t, dado que o indivíduo sobreviveu até esse instante, formalmente por:

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

A função de risco é útil para caracterizar o comportamento temporal da probabilidade de falha: valores crescentes de $\lambda(t)$ indicam risco crescente com o tempo (como em processos de envelhecimento), enquanto valores decrescentes sugerem risco maior no início do período, típico em doenças com mortalidade precoce.

Outra função conhecida e bastante útil no contexto de Análise de Sobrevivência, é a função taxa de falha acumulada. Tal função é definida como $\Lambda(t)$, ela nos concede a taxa de falha acumulada de um indivíduo até um certo instante t, e sua expressão é dada por:

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

Apesar de $\Lambda(t)$ não possuir uma interpretação direta como probabilidade, ela é útil para avaliar o comportamento da função de risco $\lambda(t)$, sendo particularmente empregada em gráficos diagnósticos, como o gráfico de Cox–Snell.

1.2.1 Estimador de Kaplan-Meier

Para estimar a função de sobrevivência na presença de dados censurados, segundo Kaplan e Meier (1958 apud COLOSIMO; GIOLO, 2006, p. 35) pode-se utilizar o método não paramétrico de Kaplan-Meier, proposto em 1958. Esse estimador, denotado por $\hat{S}(t)$, calcula a probabilidade de sobrevivência de forma condicional a cada instante em que um evento (óbito) ocorre. Sua formulação é dada por:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

sendo que:

- $t_1 < t_2 < \cdots < t_k$ são os k tempos distintos e ordenados de falha;
- d_j é o número de falhas no tempo t_j, para $j = 1, \ldots, k$;
- n_j é o número de indivíduos sob risco em t_j.

O resultado é uma curva em formato de escada que representa a estimativa da probabilidade de sobrevivência ao longo do tempo. Esse método não assume distribuição específica para o tempo de sobrevivência e permite comparações gráficas entre diferentes grupos de pacientes, sendo amplamente utilizado em estudos clínicos (COLOSIMO; GIOLO, 2006).

1.2.2 Modelo de Regressão de Cox

O modelo de regressão de Cox (COX, 1972), é uma abordagem semiparamétrica amplamente empregada para investigar o efeito de covariáveis sobre o tempo de sobrevivência. Esse modelo assume que a função de risco de um indivíduo i no tempo t pode ser expressa como:

$$\lambda(t) = \lambda_0(t) \exp(\beta x_i'),$$

em que $\lambda_0(t)$ é a função de risco basal, β são os parâmetros de regressão e x_i representam as covariáveis associadas ao indivíduo i.

A principal hipótese do modelo é a proporcionalidade dos riscos, segundo a qual o efeito das covariáveis é multiplicativo e constante ao longo do tempo. Assim, a razão entre os riscos de dois indivíduos, i e j, é:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp\{x_i'\beta - x_j'\beta\},$$

independente do tempo t.

O modelo de Cox permite estimar as razões de risco associadas a cada covariável, interpretadas como o fator de aumento ou redução no risco do evento, controlando as demais variáveis. Apesar de não requerer a especificação da forma funcional de $\lambda_0(t)$, o modelo fornece inferências consistentes sobre os parâmetros β, tornando-se uma ferramenta versátil para análise multivariada em estudos de sobrevivência (COLOSIMO; GIOLO, 2006).

1.2.2.1 Resíduo de Schoenfeld

Os resíduos de Schoenfeld propostos por Schoenfeld (1982 apud COLOSIMO; GIOLO, 2006, p. 167) são amplamente utilizados na análise de sobrevivência para verificar a validade da hipótese de proporcionalidade dos riscos no modelo de regressão de Cox. Essa hipótese, é fundamental para a interpretação do modelo, assume que os efeitos das covariáveis sobre o risco são constantes ao longo do tempo.

Formalmente, os resíduos de Schoenfeld para uma covariável x_q associada ao indivíduo i que sofreu o evento no tempo t_i são definidos como a diferença entre o valor observado da covariável e o valor esperado dessa covariável, ponderado pela função de risco estimada pelo modelo de Cox. Assim, tem-se:

$$r_{iq} = x_{iq} - \frac{\sum\limits_{j \in R(t_i)} x_{jq} \exp\{x_j'\hat{\beta}\}}{\sum\limits_{j \in R(t_i)} \exp\{x_j'\hat{\beta}\}},$$

sendo que $q = 1, \ldots, p$ e $R(t_i)$ representa o conjunto de indivíduos sob risco no instante t_i.

O segundo termo da expressão representa o valor esperado ponderado da covariável x_q entre os indivíduos que ainda estão sob risco no momento do evento. Dessa forma, o resíduo r_{iq}

mede o desvio entre o valor observado e o valor esperado da covariável, refletindo o quanto o comportamento do indivíduo i difere do padrão previsto pelo modelo naquele instante de tempo.

Esses resíduos são calculados apenas para os indivíduos que sofreram o evento (não para os censurados) e refletem a diferença entre o comportamento observado e o comportamento esperado das covariáveis segundo o modelo ajustado.

No presente trabalho, a avaliação da suposição de proporcionalidade dos riscos foi realizada por meio do teste de Schoenfeld, que utiliza os resíduos definidos acima para testar a existência de correlação significativa entre os resíduos e o tempo. Valores de p-valor elevados (geralmente acima de 0,05) indicam que não há evidências de violação da hipótese de proporcionalidade dos riscos, confirmando a adequação do modelo de Cox.

Dessa forma, os resíduos de Schoenfeld, por meio de seu teste estatístico, constituem uma ferramenta essencial para verificar a validade dos pressupostos do modelo, assegurando a confiabilidade das estimativas obtidas para os coeficientes β.

1.2.3 Resíduo de Cox-Snell

O resíduo de Cox-Snell é um tipo de resíduo utilizado para avaliar a qualidade do ajuste de modelos de sobrevivência, tanto paramétricos quanto semiparamétricos. Esse resíduo é uma medida da discrepância entre os tempos observados e os tempos esperados de falha, conforme o modelo ajustado. Embora os resíduos de Cox-Snell sejam uma boa ferramenta para nos ajudar a avaliar o ajuste global do modelo, caso seja identificado que o modelo não está bem ajustado através da visualização do gráfico, os resíduos não conseguem apontar qual o tipo de falha. Esses resíduos são determinados por:

$$\hat{e_i} = \hat{\Lambda}(t_i|x_i).$$

Os resíduos de Cox-Snell possuem uma propriedade importante: para um modelo bem ajustado, espera-se que os resíduos r_i sigam uma distribuição exponencial com média 1. Esta distribuição pode ser verificada com o auxílio de um gráfico de resíduos de Cox-Snell, utilizado para avaliar a adequação do modelo de sobrevivência ajustado. Se os resíduos r_i seguem uma distribuição exponencial, isso indica que o modelo de sobrevivência está bem ajustado aos dados. Caso contrário, a inadequação do modelo é sugerida, o que pode levar à necessidade de ajustes ou de utilização de um modelo alternativo.

O gráfico de resíduos de Cox-Snell é uma das ferramentas mais comuns para diagnosticar a qualidade do ajuste do modelo de sobrevivência (HOSMER STANLEY LEMESHOW, 2008). O gráfico exibe os resíduos r_i contra os tempos observados T_i. Idealmente, os resíduos devem seguir uma linha reta (ou próxima disso), indicando que os resíduos estão distribuídos exponencialmente, conforme esperado. Caso contrário, isso pode sugerir que o modelo de sobrevivência não está ajustado adequadamente aos dados.

Como dito anteriormente, os resíduos de Cox-Snell são utilizados para avaliar a qualidade geral do ajuste de modelos de sobrevivência. No caso do modelo de Cox, os resíduos de Cox-Snell são definidos por

$$\hat{e}_i = \hat{\Lambda}_0(t_i) \exp\Big(\sum_{k=1}^{p} x_{ik}\hat{\beta}_k\Big),$$

com $i = 1, \ldots, n$.

Se os resíduos r_i seguem uma distribuição exponencial, isso indica que o modelo de sobrevivência está bem ajustado aos dados. Caso contrário, a inadequação do modelo é sugerida, o que pode levar à necessidade de ajustes ou de utilização de um modelo alternativo.

1.2.4 Modelo de Regressão Weibull

O modelo de regressão Weibull é uma abordagem paramétrica na qual o tempo até o evento segue uma distribuição Weibull, amplamente utilizada pela sua flexibilidade em representar diferentes padrões de risco ao longo do tempo. A função de risco da distribuição Weibull é dada por:

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

onde $t \geq 0$ e $\gamma > 0$, sendo que γ é o parâmetro de forma e α o de escala.

De acordo com Colosimo e Giolo (2006), o parâmetro γ controla o comportamento da função de risco:

- se $\gamma = 1$, o risco é constante ao longo do tempo (modelo exponencial);
- se $\gamma > 1$, o risco cresce com o tempo;
- se $\gamma < 1$, o risco decresce com o tempo.

Por se tratar de um modelo paramétrico, o modelo de Weibull permite estimar as funções de risco, sobrevivência e densidade de probabilidade, que são dadas respectivamente por:

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

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

e

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\Big\{-\Big(\frac{t}{\alpha}\Big)^\gamma\Big\},$$

o que o torna apropriado para situações em que há interesse em descrever explicitamente o comportamento temporal do risco.

Como o estudo utiliza o modelo de regressão Weibull via *survreg()* do pacote *survival* (THERNEAU, 2024), é importante destacar que ele segue a parametrização AFT (do ingles *Accelerated Failure Time*). Nesse modelo, assumimos que $Y = log(T)$ (KALBFLEISCH; PRENTICE, 2002), onde sua formulação é dada por:

$$Y = \alpha + X'\beta^* + \sigma W,$$

sendo que $\alpha = -log\lambda$, $\sigma = \gamma^{-1}$ e $\beta^* = -\sigma\beta$.

Além disso, quando o pressuposto de proporcionalidade não é válido no modelo de Cox, o modelo de Weibull pode ser considerado uma alternativa paramétrica, com a vantagem adicional de possibilitar a previsão do tempo médio de sobrevivência e a extrapolação além do período observado (COLOSIMO; GIOLO, 2006).

Para o modelo Weibull os resíduos de Cox-Snell são dados por

$$\hat{e}_i = \left[t_i \exp\{-x_i'\hat{\beta}\}\right]^{\hat{\gamma}}$$

Embora a forma de se obter os resíduos seja diferente, a interpretação continua sendo a mesma, ou seja, para se considerar que temos um bom ajuste do modelo, os resíduos r_i devem seguir uma distribuição exponencial, caso contrário, a indícios de um modelo mal ajustado.

1.3 Desenvolvimento Computacional e Aplicação Shiny

A implementação prática das análises de sobrevivência foi realizada no software *RStudio*, utilizando o pacote *survival* (THERNEAU, 2024), em conjunto com o *survminer* (KASSAMBARA; KOSINSKI; BIECEK, 2024), que emprega o *ggplot2* (WICKHAM, 2016) para a geração das curvas de Kaplan-Meier, Cox e Weibull. Essa ferramenta é útil para investigar diferenças de prognóstico entre subgrupos, como por exemplo, comparar curvas de sobrevivência segundo o sexo ou diferentes modalidades de tratamento. A utilização dessa metodologia possibilita identificar fatores associados a maior ou menor tempo de sobrevida, fornecendo subsídios importantes para a interpretação clínica dos dados.

De forma geral, os resultados obtidos mostram que a aplicação é funcional para análise exploratória e visualização de dados, oferecendo recursos gráficos interativos via *plotly* (SIEVERT, 2020) que auxiliam na compreensão da base. A presença do mapa interativo, gerado com a ajuda do pacote *geobr* (FONSECA; PEREIRA *et al.*, 2021), agrega valor à análise, pois permite verificar a distribuição espacial dos registros, enquanto as curvas de sobrevivência permitem avançar na compreensão de fatores associados à evolução clínica dos pacientes. As ferramentas compu-

tacionais descritas fornecem a base para a exploração interativa dos resultados no *dashboard*, apresentado nos capítulos subsequentes.

Para o desenvolvimento do painel interativo, foi utilizado o pacote *Shiny* (CHANG *et al.*, 2024), disponível no ambiente de programação *RStudio*. Este pacote permite a criação de aplicações web interativas diretamente a partir de códigos em *R*, facilitando a integração entre análises estatísticas e visualização de dados. O *Shiny* é amplamente utilizado em projetos que envolvem análise exploratória, modelagem estatística e comunicação de resultados de forma dinâmica e acessível.

A estrutura de um aplicativo *Shiny* é composta, essencialmente, por dois blocos principais:

- **Interface do Usuário (UI)**: é a parte responsável pela camada visual do aplicativo, onde são definidos os elementos gráficos com os quais o usuário interage, como menus, botões, caixas de seleção e gráficos. A interface determina o layout e o estilo do painel, funcionando como o "rosto" da aplicação.
- **Servidor (Server)**: é o componente responsável pela lógica e processamento do aplicativo. Nesta seção são implementadas as funções e comandos em *R* que processam os dados, executam análises estatísticas, geram gráficos e retornam os resultados para a interface do usuário.

A comunicação entre a *UI* e o *server* ocorre de forma reativa, ou seja, sempre que o usuário realiza uma interação na interface (como selecionar uma variável ou aplicar um filtro), o servidor executa automaticamente as operações necessárias e atualiza os resultados exibidos na tela. Essa característica torna o *Shiny* uma ferramenta poderosa para a construção de *dashboards* dinâmicos e de fácil interpretação, especialmente em contextos de análise de dados em saúde, como o presente estudo.

Por fim, uma característica adicional do *Shiny* é a facilidade de publicação de aplicações na web por meio da plataforma *shinyapps.io*, que permite hospedar o *dashboard* diretamente em servidores da própria *RStudio*. Essa funcionalidade possibilita que o painel desenvolvido seja acessado de qualquer local com conexão à internet, sem a necessidade de infraestrutura própria de hospedagem.

2 Tratamento dos Dados

O desenvolvimento do painel interativo foi dividido em três etapas principais: coleta dos dados, tratamento e organização da base, e implementação do painel em Shiny (CHANG *et al.*, 2024).

A primeira etapa consistiu na coleta dos dados, obtidos diretamente do portal da Fundação Oncocentro de São Paulo (FOSP) (Fundação Oncocentro de São Paulo, 2024), que disponibiliza os registros hospitalares de câncer em formato de data base file (DBF). O arquivo foi importado para o ambiente de desenvolvimento Python e convertido para o formato de valores separados por vírgula (CSV), pois o R apresentou dificuldades em ler o arquivo em sua extensão padrão. O arquivo foi importado para o ambiente de desenvolvimento do RStudio, utilizando a função *read.csv*. O arquivo possui 1.257.217 linhas e 104 colunas, e continha dados do ano de 2000 à 2024, porém, os dados foram filtrados para incluir apenas os pacientes diagnosticados entre os anos de 2014 e 2016, com acompanhamento até 31 de dezembro de 2021, assim definindo o escopo temporal do estudo, período esse que foi estabelecido arbitrariamente para realizar a criação da primeira versão do *dashboard*. Por fim, a base de dados ficou com 199.146 linhas e 107 colunas.

Na segunda etapa, realizou-se um pré-processamento da base utilizando os pacotes *dplyr* (WICKHAM *et al.*, 2023) para manipulação e *lubridate* (GROLEMUND; WICKHAM, 2011) para o tratamento de datas. Este processo incluiu alguns passos, sendo o primeiro a criação das variáveis que serão utilizadas na abordagem de análise de sobrevivência, onde foram geradas as duas variáveis centrais sendo elas: o tempo de seguimento até a morte devido ao câncer (em dias) e o indicador de censura, variáveis essas que foram nomeadas como TEMPO e CENSURA, respectivamente. O tempo foi calculado como o intervalo entre a data do diagnóstico e a data da última informação. Para os pacientes sem registro de óbito até o final do período de acompanhamento, foi estabelecida uma data de censura em 31 de dezembro de 2021. O indicador de censura foi definido como 1 para os pacientes com registro de óbito e 0 para os demais (censurados). O segundo passo foi o tratamento e seleção de variáveis, onde as variáveis de interesse foram selecionadas, sendo elas: (TOPOGRUP, TEMPO, CENSURA, ANODIAG, ESCOLARI, IDADE, SEXO, UFRESID, CATEATEND, RADIO, QUIMIO, CIRURGIA), onde o significado de cada uma de acordo com o dicíonario disponibilizado pela FOSP é

- TOPOGRUP: Grupo da topografia;
- ANODIAG: Ano de diagnóstico;
- ESCOLARI: Código para escolaridade do paciente;
- IDADE: Idade do paciente;

- SEXO: Sexo do paciente;
- UFRESID: UF de residência;
- CATEATEND: Categoria de atendimento ao diagnóstico;
- RADIO: Se o paciente fez ou não radioterapia;
- QUIMIO: Se o paciente fez ou não quimioterapia;
- CIRURGIA: Se o paciente fez ou não cirurgia;

Sendo que ESCOLARI, UFRESID e CATEATEND foram renomeadas para ESCOLARIDADE, UF_RESIDENCIA e CAT._ATENDIMENTO respectivamente. Variáveis categóricas, como ESCOLARIDADE, SEXO e CAT._ATENDIMENTO, foram convertidas de códigos numéricos para o formato de fator (*factor*), com seus respectivos rótulos descritivos. No terceiro e ultimo passo, foi criada uma nova variável binária (IDADE_MED) para categorizar os pacientes em dois grupos: abaixo da idade mediana e acima ou igual à idade mediana, permitindo análises comparativas, a mediana da idade que foi levado em consideração foi usando toda a base, sem nenhum tipo de filtro e o valor não sofrerá nenhum tipo de alteração ao decorrer do trabalho. Adicionalmente, os códigos de topografia do câncer (TOPOGRUP) foram recodificados para seus nomes completos, facilitando a interpretação no painel. Logo, os dados ficaram com 197.304 linhas e 14 colunas, sendo que cada linha representa a informação de um paciente e cada coluna representa as variáveis que foram escolhidas.

Por fim, na terceira etapa, foi desenvolvida a aplicação utilizando o pacote Shiny do software RStudio (R Core Team, 2024). O painel foi estruturado de forma a permitir interação direta com o usuário, possibilitando a escolha das variáveis a serem analisadas e a visualização imediata de estatísticas descritivas, para variáveis numéricas sendo elas: média, mediana, desvio padrão, mínimo e máximo, e para as variáveis categóricas, uma contagem por categoria. Ao escolher a variável também é mostrado uma representação gráfica, em formato de histograma para variáveis numéricas, em barras para variáveis categóricas e um mapa de calor para a variável UF_RESIDENCIA. Esses formatos foram escolhidos pois geram um entendimento rápido e fácil para o usuário.

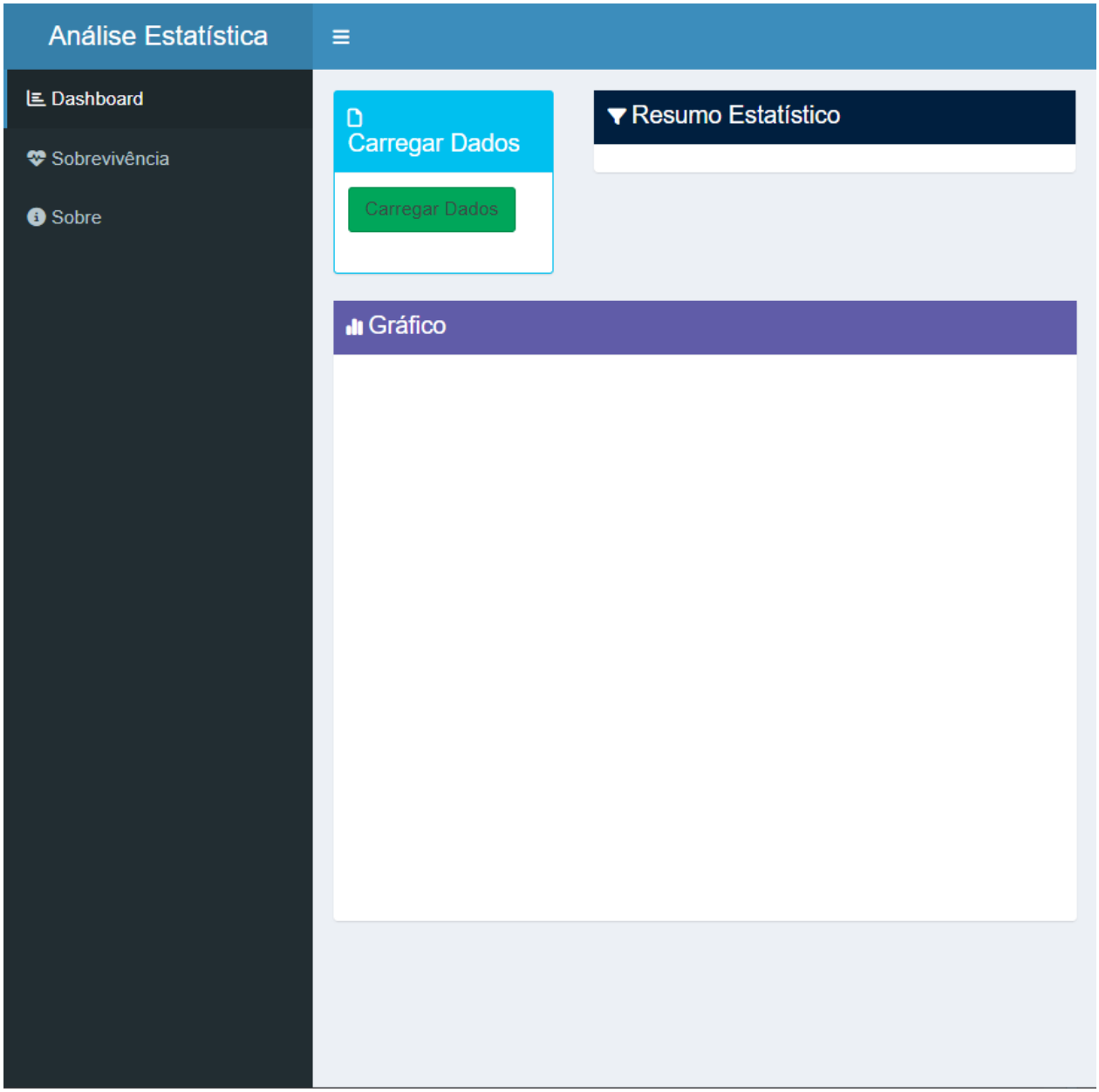
Por fim as variáveis selecionadas e incluídas no painel foram: escolaridade, idade, sexo, Unidade Federativa de residência, categoria de atendimento, radioterapia, quimioterapia e cirurgia. Além disso, na aba de Análise de Sobrevivência, foi incorporado o estimador de Kaplan-Meier, permitindo a avaliação do tempo de sobrevida dos pacientes segundo diferentes características clínicas e demográficas presentes na base de dados.

3 Resultados e discussão

3.1 Dashboard

A aplicação desenvolvida, como pode ser visto na Figura 1, permitiu a análise estatística descritiva e a visualização gráfica dos dados de interesse, possibilitando a exploração de diferentes variáveis relacionadas ao perfil da base utilizada.

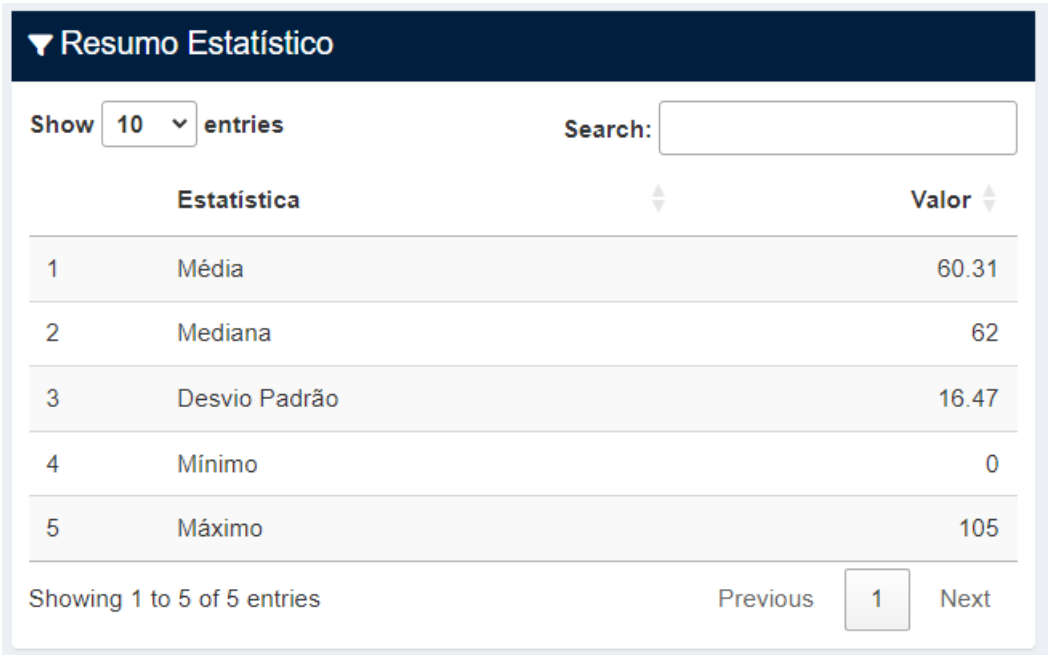
Figura 1 – Dashboard



Fonte: Elaborado pelo autor

Na aba *Resumo Estatístico*, quando selecionadas variáveis numéricas (como idade), o sistema gera automaticamente medidas descritivas, incluindo média, mediana, desvio-padrão, valores mínimos e máximos, como pode ser visto na Figura 2. Isso facilita a compreensão inicial da distribuição dos dados, bem como a identificação de possíveis assimetrias ou valores extremos.

Figura 2 – Resumo Estatístico para Idade



Fonte: Elaborado pelo autor

A Figura 2 apresenta o resumo estatístico da variável idade obtido pelo painel. Observa-se que a idade média dos pacientes é de aproximadamente 60 anos, com mediana de 62, indicando uma distribuição relativamente simétrica em torno desse valor. O desvio-padrão de 16,46 sugere uma variabilidade considerável nas idades, enquanto os valores mínimo (0) e máximo (105) evidenciam a amplitude da faixa etária registrada.

Para variáveis categóricas (como sexo, escolaridade ou clínica), a aplicação retorna tabelas de frequência, que permitem verificar a proporção de indivíduos em cada categoria, como pode ser visto na Figura 3. Esse recurso auxilia na descrição do perfil da população estudada.

Figura 3 – Resumo Estatístico para Escolaridade

Resumo Estatístico

Show 10 entries Search:

	ESCOLARIDADE	Frequência	Freq. Relativa (%)
1	IGNORADA	48372	24.52
2	ANALFABETO	11414	5.78
3	ENS. FUND. INCOMPLETO	59126	29.97
4	ENS. FUND. COMPLETO	34660	17.57
5	ENSINO MÉDIO	25630	12.99
6	SUPERIOR	18102	9.17

Showing 1 to 6 of 6 entries Previous 1 Next

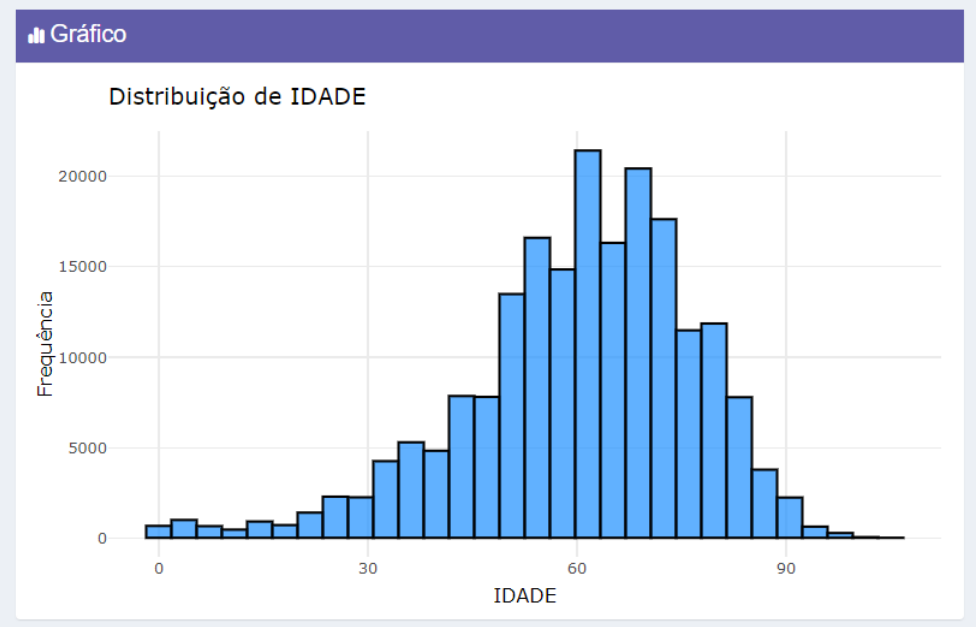
Fonte: Elaborado pelo autor

A Figura 3 apresenta o resumo estatístico da variável escolaridade. Observa-se que a maior parte dos registros corresponde a indivíduos com ensino fundamental incompleto (59.666 casos), seguido por ensino fundamental completo (34.972) e ensino médio (25.860). Nota-se ainda que um número expressivo de registros aparece como ignorado (48.847), o que pode indicar falhas ou ausência de preenchimento adequado no banco de dados. As categorias de analfabeto (11.494) e superior (18.307) concentram menores frequências. Esse resultado sugere tanto uma predominância de baixa escolaridade entre os pacientes atendidos quanto a necessidade de cautela na análise, devido ao volume significativo de informações faltantes.

Na aba de *Gráfico*, foram implementados diferentes tipos de visualização, adaptados ao tipo de variável analisada:

- Para variáveis numéricas, são gerados histogramas a partir do pacote *ggplot2* (WICKHAM, 2016), evidenciando a distribuição dos valores, como pode ser visto na Figura 4.

Figura 4 – Histograma para variável Idade

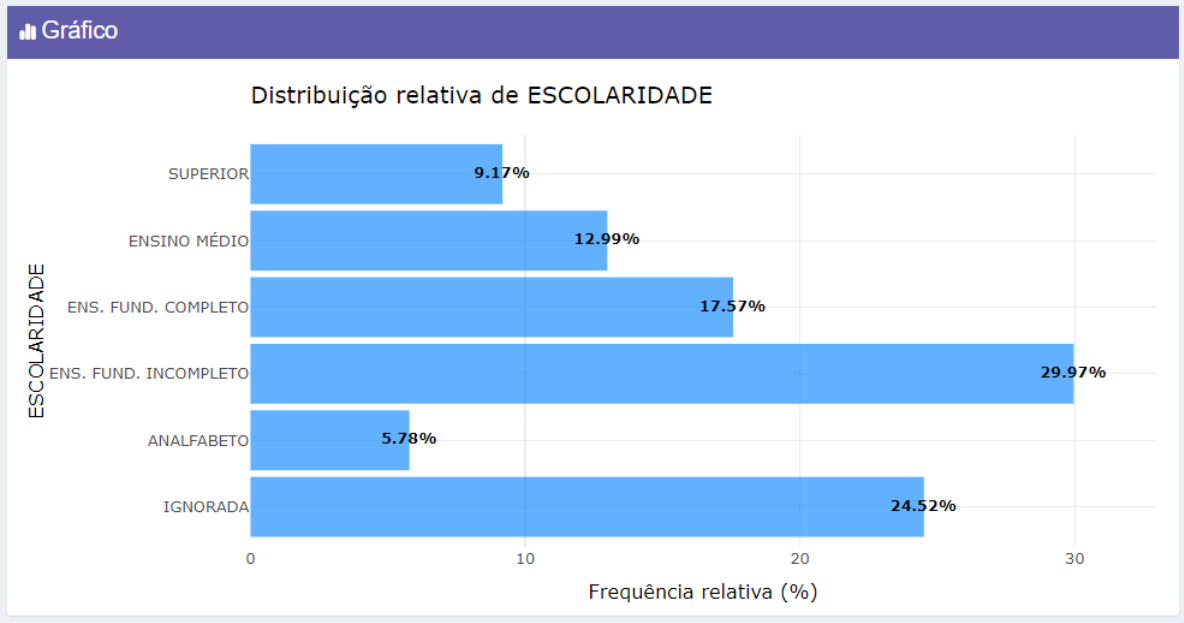


Fonte: Elaborado pelo autor

A Figura 4 apresenta a distribuição da variável idade dos pacientes, representada por meio de um histograma. Observa-se que a maior concentração de registros ocorre entre aproximadamente 50 e 70 anos, com pico próximo aos 60 anos, o que está em consonância com a literatura, que aponta maior incidência de câncer em faixas etárias mais avançadas. Nos extremos da distribuição, nota-se menor frequência de registros em crianças, jovens e idosos acima de 90 anos. Essa visualização confirma o resultado do resumo estatístico (Figura 2), destacando que a população atendida pela FOSP (Fundação Oncocentro de São Paulo, 2024) é majoritariamente composta por adultos e idosos.

- Para variáveis categóricas, utilizam-se gráficos de barras com auxílio do mesmo pacote, como pode ser visto na Figura 5.

Figura 5 – Gráfico de barras para variável Escolaridade

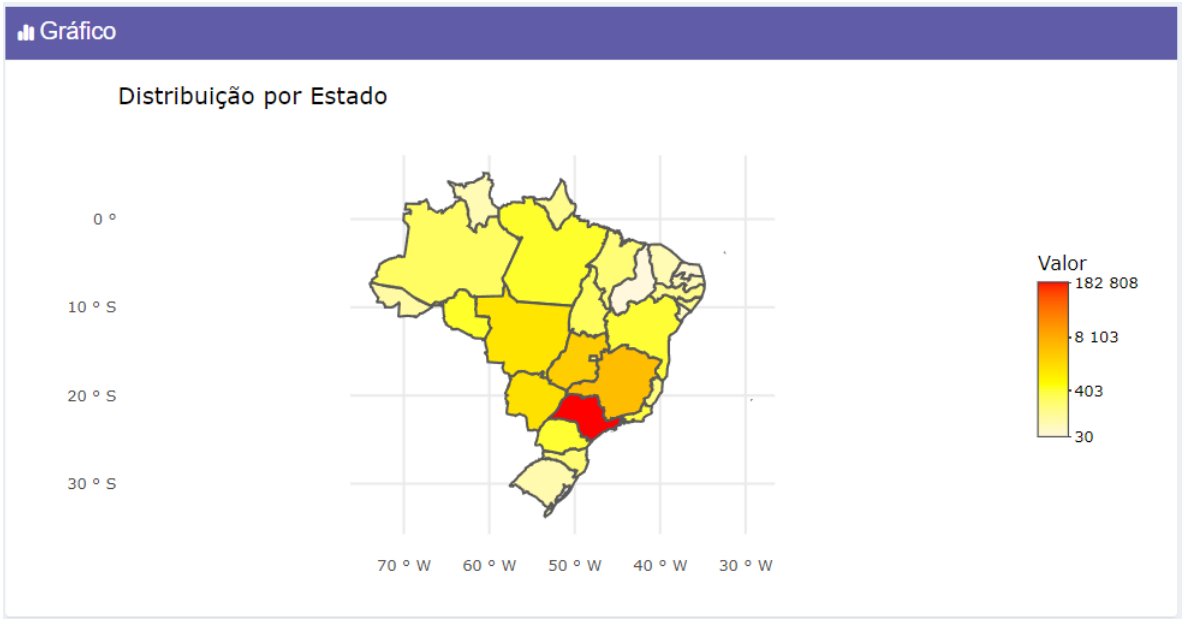


Fonte: Elaborado pelo autor

A Figura 5 apresenta a distribuição dos pacientes segundo a variável escolaridade, representada por meio de um gráfico de barras. Nota-se que a categoria mais frequente é ensino fundamental incompleto, com 29,97% dos registros, seguida por ensino fundamental completo e ensino médio. Chama atenção também o elevado número de registros classificados como "ignorado", o que pode indicar falhas ou ausência de preenchimento no banco de dados. As categorias analfabeto e superior aparecem em menor proporção. Esse padrão sugere predominância de pacientes com baixa escolaridade nos registros da FOSP (Fundação Oncocentro de São Paulo, 2024), além de evidenciar a necessidade de cautela na interpretação devido à elevada proporção de dados faltantes.

- Para a variável UF de residência, foi elaborado um mapa de calor interativo utilizando o pacote *geobr* (FONSECA; PEREIRA *et al.*, 2021), no qual a intensidade da cor representa o número de casos registrados em cada Estado, como pode ser visto na Figura 6.

Figura 6 – Gráfico para UF de residência



Fonte: Elaborado pelo autor

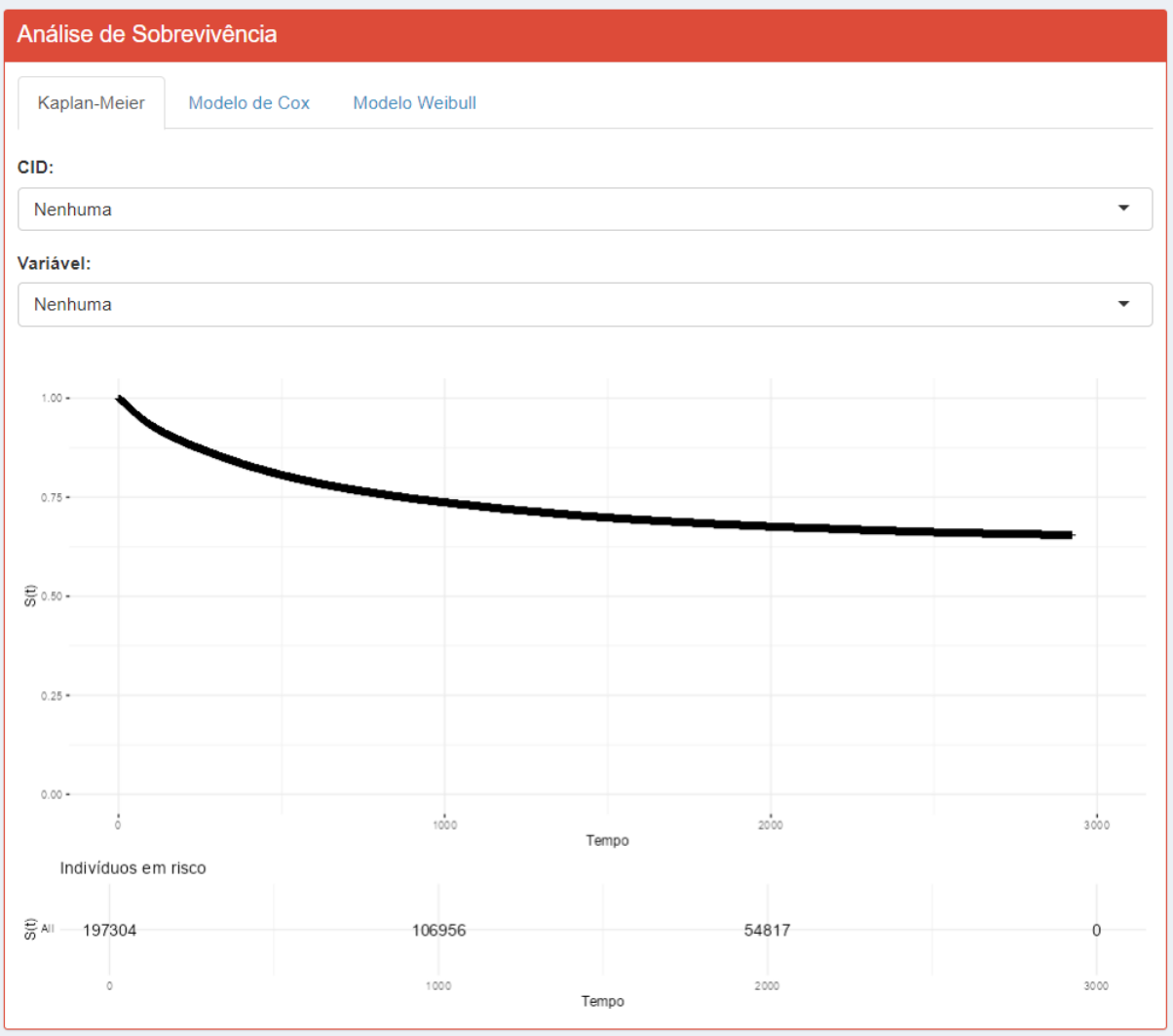
A Figura 6 apresenta o mapa de calor do Brasil, observa-se que o Estado de São Paulo apresenta valores mais elevados no mapa de calor, o que se deve ao fato de a Fundação Oncocentro de São Paulo (FOSP) (Fundação Oncocentro de São Paulo, 2024) concentrar registros hospitalares de câncer de instituições localizadas nesse Estado. Dessa forma, a distribuição espacial reflete principalmente a abrangência e o foco geográfico da base de dados utilizada. Os demais Estados, que aparecem com menor frequência ou até mesmo sem registros, não necessariamente indicam menor incidência da doença, mas sim a ausência de cobertura nacional nos dados disponibilizados pela FOSP.

3.2 Sobrevivência

3.2.1 Kaplan-Meier

Na aba *Sobrevivência*, foi implementada a análise de Kaplan-Meier na sub-aba de mesmo nome, permitindo a estimação da função de sobrevivência ao longo do tempo, como pode ser visto na Figura 7, com possibilidade de aplicar dois tipos de filtro, com o primeiro sendo pelo CID, e o segundo por uma das seguintes variáveis: SEXO, IDADE_MED, CAT._ATENDIMENTO, RADIO, QUIMIO e CIRURGIA, como pode ser visto na Figura 8.

Figura 7 – Gráficos de Kaplan-Meier - Sem nenhum tipo de filtro



Fonte: Elaborado pelo autor

A Figura 7 apresenta exemplos de curvas de sobrevivência obtidas pelo estimador de Kaplan-Meier. As pequenas marcas verticais em cada curva representam as censuras. No gráfico, em que não foi aplicado nenhum filtro, observa-se a curva geral de sobrevivência para toda a base de pacientes, podemos ver uma redução gradual da probabilidade de sobrevivência ao longo do tempo.

Figura 8 – Gráficos de Kaplan-Meier - Filtrado pelo CID C34 e pela variável SEXO

Fonte: Elaborado pelo autor

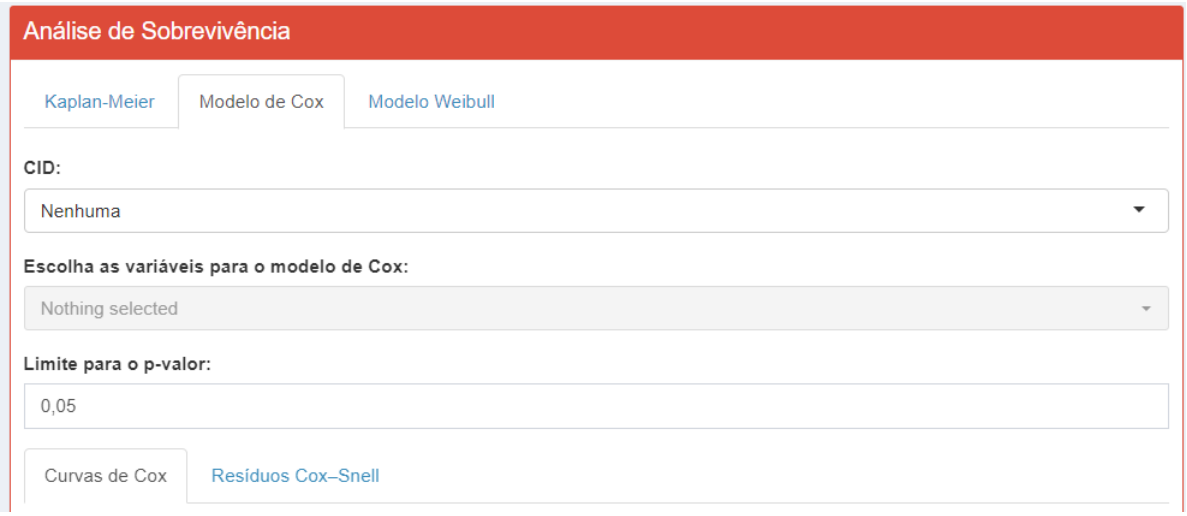
Já no gráfico apresentado na figura 8, ao aplicar o filtro C34 - Brônquios e pulmões e estratificar pela variável sexo, nota-se diferença entre os grupos: pacientes do sexo masculino (representados pela cor azul) apresentam menor probabilidade de sobrevida em comparação às pacientes do sexo feminino (representados pela cor vermelha). Esse resultado ilustra a utilidade da análise de Kaplan-Meier para investigar fatores prognósticos e diferenças entre subgrupos de pacientes. Portanto, podemos concluir que existe uma evidência estatística forte de que o sexo é um fator que influencia o tempo de sobrevida nesta população de estudo, com as mulheres apresentando um sinal de sobrevida mais favorável que os homens.

3.2.2 Modelo de Cox

Já na sub-aba de nome *Modelo de Cox*, foi implementado o modelo de regressão de Cox, aqui seguimos podendo aplicar os filtros por CID e por variável, porém, agora também é possível realizar a escolha do p-valor, como pode ser visto na figura 9. Tal modelo nos permite avaliar

o efeito de diferentes covariáveis sobre o tempo de sobrevivência dos pacientes. Esse modelo é amplamente utilizado em estudos clínicos por sua capacidade de estimar a razão de riscos associada a cada variável explicativa, sem a necessidade de especificar a forma funcional da função de risco basal.

Figura 9 – Aba Modelo de Cox - Filtros disponíveis



Fonte: Elaborado pelo autor

Na figura 9 vemos os filtros mencionados anteriormente, porém, diferentemente da aba Kaplan-Meier, aqui, podemos selecionar mais de uma variável para realizar o ajuste do modelo. No campo "Limite para o p-valor" é onde o usuário pode escolher o valor que desejar para o p-valor.

Figura 10 – Teste de proporcionalidade com todas as variáveis proporcionais.

Fonte: Elaborado pelo autor

Como podemos observar na figura 10, para o CID foi escolhido C34 - Brônquios e pulmões,

as variáveis para o modelo foram SEXO, IDADE e CIRURGIA, e o p-valor foi de 0,05, nessas condições o modelo nos retorna o teste de proporcionalidade (Resíduos de schoenfeld) para as variáveis escolhidas, onde, de acordo com o p-valor escolhido a aplicação dirá se as variáveis atendem ou não o pressuposto de proporcionalidade, neste exemplo, todas atenderam.

Figura 11 – Teste de proporcionalidade com algumas variáveis proporcionais.

Fonte: Elaborado pelo autor

Já na figura 11, as variáveis para o modelo foram alteradas, sendo agora: SEXO, IDADE_MED, IDADE, CIRURGIA, CAT._ATENDIMENTO, RADIO e QUIMIO, o CID e o p-valor foram mantidos os mesmo. Para essas condições vemos que as variáveis SEXO, CIRURGIA, RADIO e QUIMIO, não atenderam o pressuposto de proporcionalidade considernado um nível de significância de 5%, portanto, elas são automaticamente retiradas do modelo e o ajuste é realizado com as variáveis restantes.

Figura 12 – Teste de proporcionalidade com nenhuma variável proporcional.

Fonte: Elaborado pelo autor

Agora na figura 12, é mostrado o exemplo em que nenhuma variável atendeu o pressuposto de proporcionalidade, nesse caso, a aplicação retorna uma mensagem para o usuário e o modelo não é ajustado.

Figura 13 – Resumo do modelo ajustado - Filtrado pelo CID C38 e pelas variáveis CIRURGIA, RADIO

```
Call:
coxph(formula = form_init, data = df)

  n= 525, number of events= 230

              coef exp(coef) se(coef)      z Pr(>|z|)
CIRURGIA.L -0.2540    0.7757   0.1223 -2.078  0.03773 *
RADIO.L     0.3434    1.4098   0.1190  2.885  0.00391 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
CIRURGIA.L    0.7757     1.2892    0.6104    0.9857
RADIO.L       1.4098     0.7093    1.1164    1.7802

Concordance= 0.567  (se = 0.016 )
Likelihood ratio test= 12.69  on 2 df,   p=0.002
Wald test            = 11.78  on 2 df,   p=0.003
Score (logrank) test = 11.95  on 2 df,   p=0.003
```

Fonte: Elaborado pelo autor

A Figura 13 apresenta a saída do modelo de regressão de Cox ajustado para as variáveis CIRURGIA e RADIO, considerando o tempo até o óbito como variável resposta. Podemos ver

que o modelo foi estimado com base em 525 observações, das quais 230 correspondem a eventos observados (óbitos).

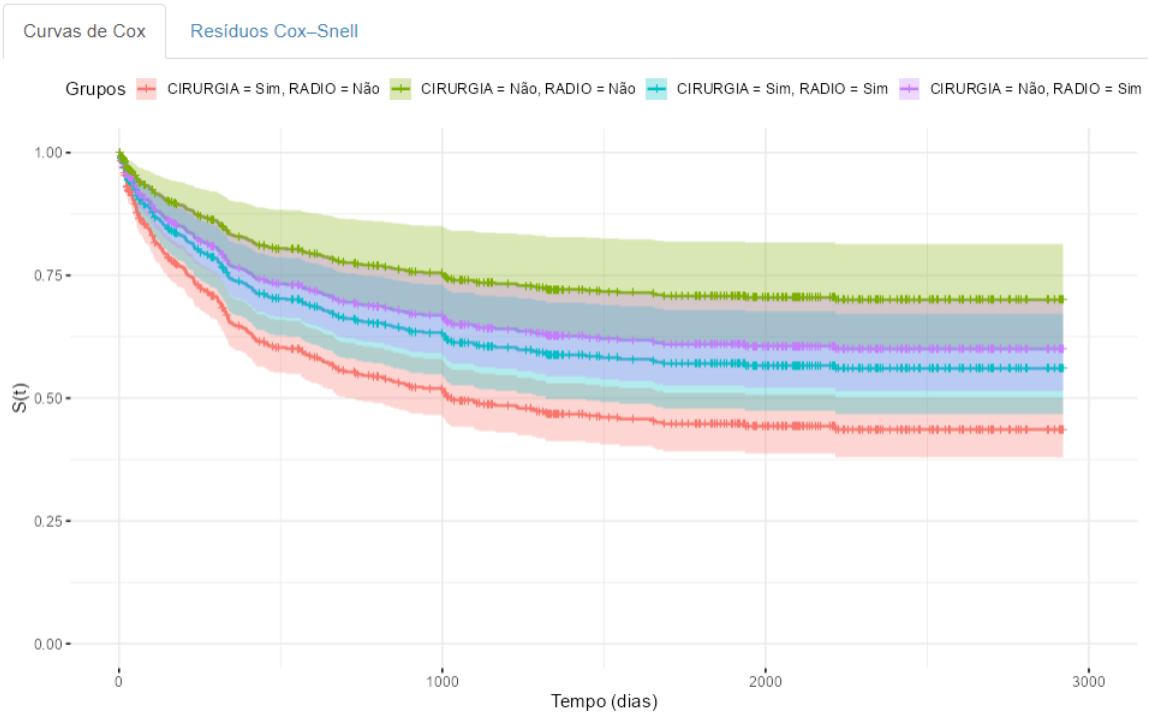
Observa-se que ambas as variáveis apresentam significância estatística, indicando associação com o tempo de sobrevivência. A variável CIRURGIA apresenta coeficiente negativo ($\hat{\beta} = -0{,}2540$, $p = 0{,}0377$), o que implica que pacientes submetidos à cirurgia possuem menor risco de óbito em comparação aos que não realizaram o procedimento, quando controladas as demais variáveis do modelo. O valor de $\exp(\hat{\beta}) = 0{,}7757$ representa a razão de riscos, indicando uma redução de aproximadamente 22,4% no risco de morte para o grupo que realizou cirurgia.

Por outro lado, a variável RADIO apresenta coeficiente positivo ($\hat{\beta} = 0{,}3434$, $p = 0{,}0039$), sugerindo que pacientes que receberam radioterapia possuem maior risco de óbito em relação aos que não receberam. A razão de riscos $\exp(\hat{\beta}) = 1{,}4098$ indica um aumento aproximado de 40,9% no risco de morte associado à radioterapia, o que pode refletir o fato de que esse tratamento é frequentemente aplicado a casos mais graves.

Os testes globais de significância *Likelihood Ratio*, *Wald* e *Score (logrank)* apresentaram valores de $p = 0{,}002$, $p = 0{,}003$ e $p = 0{,}003$, respectivamente, indicando que o modelo, como um todo, é estatisticamente significativo. A concordância de 0,567 sugere que o modelo possui uma capacidade moderada de discriminar entre pacientes com maior e menor risco de óbito.

Em resumo, o modelo de regressão de Cox ajustado aponta que a realização de cirurgia está associada a um efeito protetor sobre a sobrevivência, enquanto a radioterapia, possivelmente por estar relacionada a casos mais avançados, está associada a um aumento no risco de morte.

Figura 14 – Curvas de Cox - Filtrado pelo CID C38 e pelas variáveis CIRURGIA e RADIO



Fonte: Elaborado pelo autor

A Figura 14 apresenta as curvas de sobrevivência estimadas pelo modelo de regressão de Cox para as combinações das variáveis *CIRURGIA* e *RADIO*. Cada curva representa a probabilidade estimada de sobrevivência ao longo do tempo, considerando as categorias das variáveis de tratamento.

Observa-se que os pacientes que realizaram cirurgia e não receberam radioterapia (*CIRURGIA = Sim, RADIO = Não*) apresentam as maiores probabilidades de sobrevivência durante todo o período de acompanhamento, com uma curva consistentemente acima das demais. Esse resultado reforça o efeito protetor da cirurgia identificado na análise dos coeficientes do modelo, indicando que o procedimento cirúrgico está associado a um menor risco de óbito.

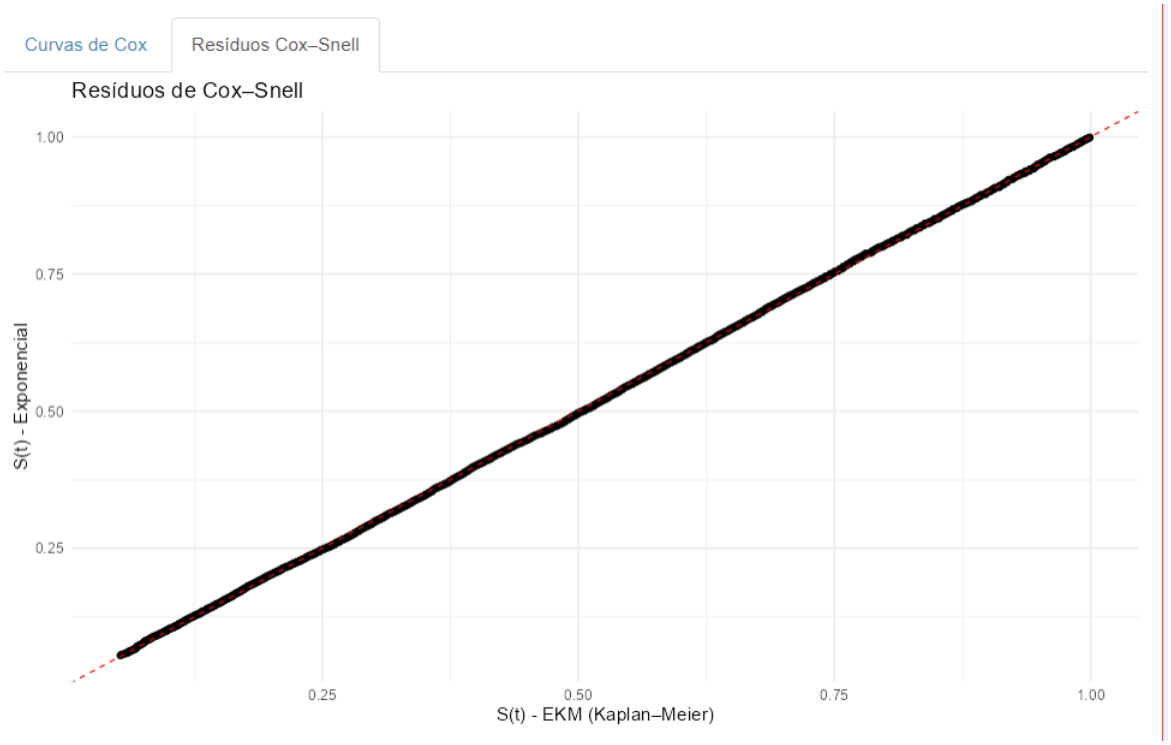
Em contrapartida, o grupo que não realizou cirurgia e também não recebeu radioterapia (*CIRURGIA = Não, RADIO = Não*) apresenta a menor probabilidade de sobrevivência, evidenciando a importância do tratamento ativo no prolongamento da vida dos pacientes. Os grupos que receberam radioterapia, independentemente da realização de cirurgia, apresentam curvas intermediárias, o que sugere que o efeito desse tratamento é mais complexo e possivelmente influenciado pela gravidade do caso clínico, visto que a radioterapia tende a ser indicada para situações mais avançadas.

As faixas coloridas ao redor de cada curva representam os intervalos de confiança de 95% para as estimativas de sobrevivência, indicando o grau de incerteza associado às estimativas em cada ponto do tempo. Nota-se que as curvas permanecem bem separadas entre si, o que indica

diferenças estatisticamente relevantes entre os grupos analisados.

De forma geral, o comportamento das curvas é coerente com os resultados obtidos no modelo de regressão de Cox, sugerindo que a cirurgia exerce um efeito benéfico sobre a sobrevivência, enquanto a radioterapia está associada a uma maior taxa de mortalidade, possivelmente por refletir casos mais graves da doença.

Figura 15 – Resíduos de Cox-Snell - Filtrado pelo CID C34 e pela variável SEXO



Fonte: Elaborado pelo autor

A Figura 15 apresenta o gráfico dos resíduos de Cox–Snell obtidos a partir do modelo de regressão de Cox ajustado. Esse gráfico é utilizado para avaliar a adequação geral do modelo de sobrevivência, verificando se os resíduos seguem a distribuição esperada sob um bom ajuste, isto é, uma distribuição exponencial com média igual a 1.

No gráfico, os resíduos estimados são comparados com a curva teórica da distribuição exponencial (representada pela linha tracejada em vermelho). Quando o modelo se ajusta bem aos dados, os pontos observados tendem a alinhar-se próximo à linha de referência, indicando que o comportamento empírico dos resíduos é consistente com o esperado pelo modelo teórico.

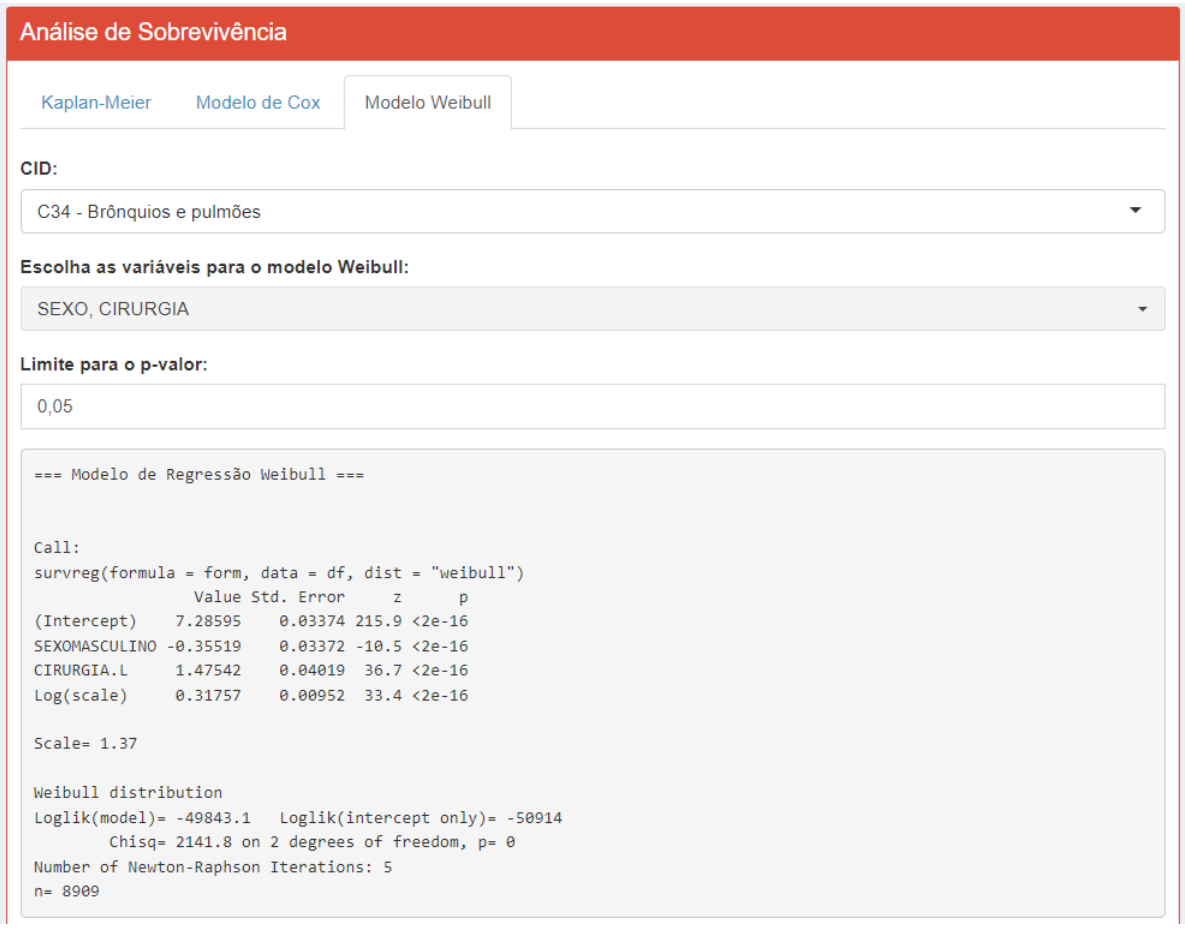
Observa-se que os pontos seguem de forma bastante próxima a linha de 45 graus, o que sugere que o modelo de Cox apresenta um ajuste adequado aos dados, sem evidências de grandes desvios sistemáticos. Em outras palavras, as suposições do modelo especialmente a proporcionalidade dos riscos, parecem razoavelmente satisfeitas.

Esse resultado reforça a validade do modelo ajustado e indica que as estimativas obtidas para as variáveis *CIRURGIA* e *RADIO* são confiáveis para interpretação e inferência estatística.

3.2.3 Modelo de Weibull

Já na sub-aba de nome *Modelo Weibull*, foi implementado o modelo de regressão Weibull, que permite estimar a função de risco e de sobrevivência assumindo que os tempos até o evento seguem uma distribuição paramétrica do tipo Weibull. Esse modelo possibilita descrever diferentes padrões de risco ao longo do tempo e, assim como o modelo de Cox, avaliar o efeito de covariáveis sobre o tempo de sobrevivência. Além disso, a sub-aba apresenta as estimativas dos parâmetros, gráficos de curvas ajustadas e o diagnóstico de ajuste por meio dos resíduos de Cox–Snell, permitindo verificar a adequação do modelo aos dados analisados.

Figura 16 – Resumo do Modelo Weibull Filtrado pelo CID C34 e pelas variáveis SEXO, CIRURGIA



Fonte: Elaborado pelo autor

A Figura 16 apresenta a saída do modelo de regressão Weibull ajustado para o grupo de pacientes com câncer de brônquios e pulmões (*CID C34*), considerando as variáveis *SEXO* e *CIRURGIA* como covariáveis explicativas. O modelo foi estimado com base em 8.090 observações, sendo implementado por meio da função *survreg()* do pacote *survival*.

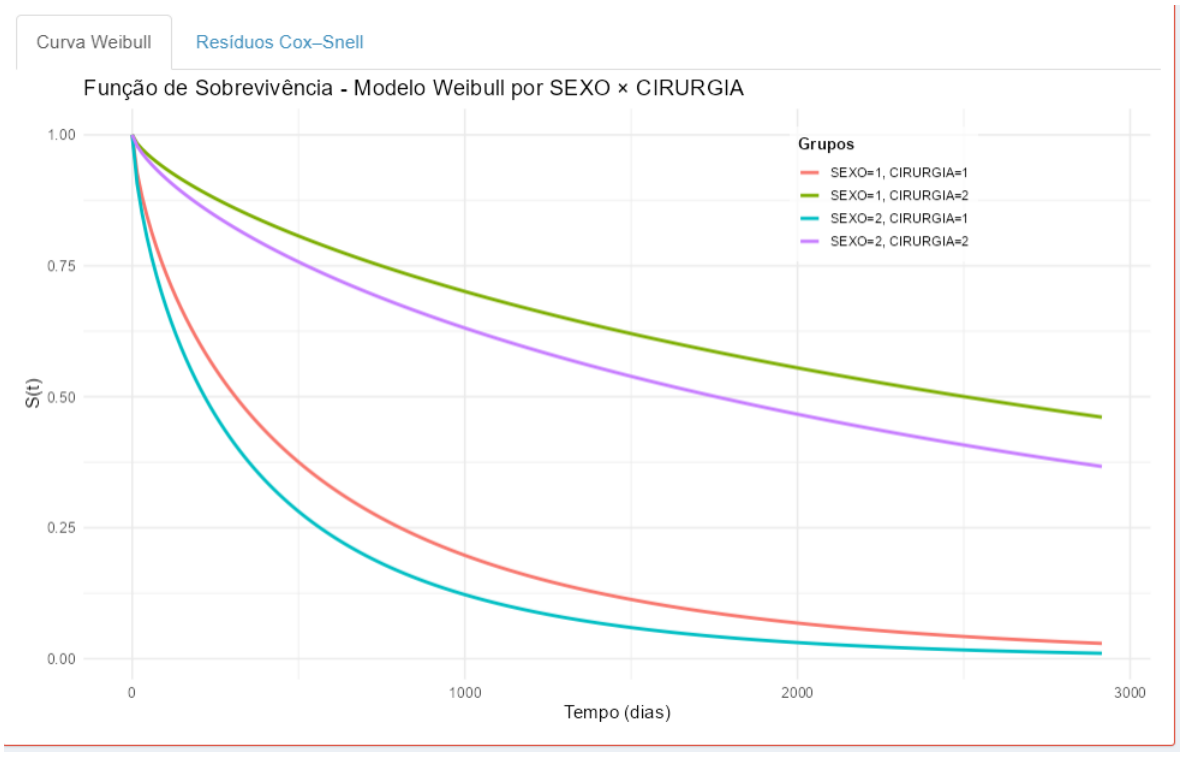
Observa-se que ambas as variáveis apresentam significância estatística ao nível de 5%, indicando associação com o tempo de sobrevivência. O coeficiente negativo associado à variável *SEXO MASCULINO* ($\hat{\beta} = -0{,}35519$, $p < 0{,}001$) sugere que pacientes do sexo masculino

possuem menor tempo médio de sobrevivência em comparação às pacientes do sexo feminino. Já a variável *CIRURGIA.L* apresenta coeficiente positivo ($\hat{\beta} = 1{,}47542$, $p < 0{,}001$), indicando que a realização de cirurgia está associada a um aumento no tempo de sobrevivência, ou seja, um efeito protetor em relação ao risco de óbito.

O parâmetro *scale* do modelo estimado é 1,37, o que implica que o parâmetro de forma da distribuição Weibull é aproximadamente $\gamma = 1/1{,}37 \approx 0.73$. Esse valor menor que 1 indica uma função de risco decrescente ao longo do tempo, ou seja, o risco de morte tende a ser maior nos períodos iniciais após o diagnóstico e diminui à medida que o tempo passa.

O resultado do teste de razão de verossimilhança apresentado na saída indica $\chi^2 = 2141{,}8$ com 2 graus de liberdade e $p < 0{,}001$, o que confirma que, de forma conjunta, as covariáveis incluídas no modelo contribuem significativamente para explicar o tempo de sobrevivência. Além disso, os valores de log-verossimilhança mostram uma melhora substancial no ajuste do modelo completo (-49843.1) em relação ao modelo nulo (-50914), evidenciando que a inclusão das variáveis *SEXO* e *CIRURGIA* aumenta a capacidade explicativa do modelo.

Figura 17 – Curvas de Weibull - Filtrado pelo CID C34 e pelas variáveis SEXO, CIRURGIA



Fonte: Elaborado pelo autor

A Figura 17 apresenta as curvas de sobrevivência estimadas pelo modelo de regressão Weibull, considerando a interação entre as variáveis *SEXO* e *CIRURGIA*. Cada curva representa a probabilidade estimada de sobrevivência ao longo do tempo para um grupo específico de pacientes, de acordo com a combinação dessas covariáveis.

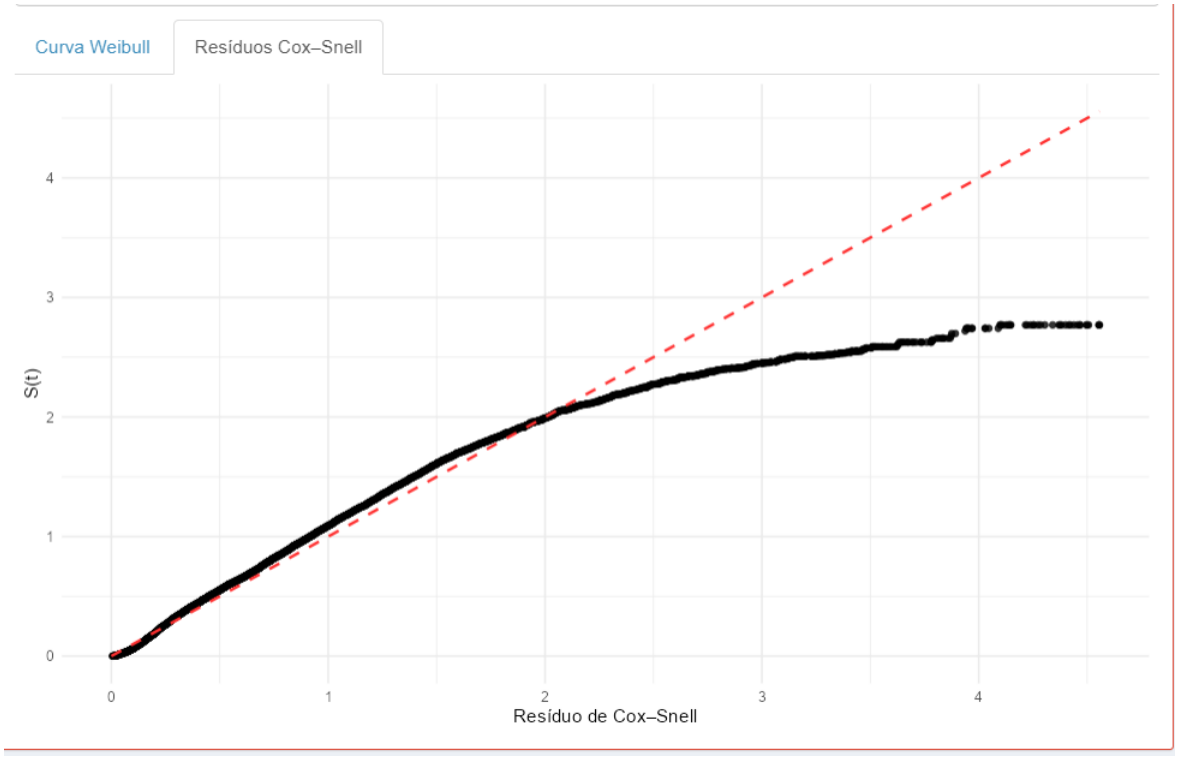
Observa-se que os pacientes do sexo feminino que realizaram cirurgia (*SEXO = 1, CI-*

RURGIA = 2) apresentam as maiores probabilidades de sobrevivência durante todo o período analisado, com uma curva consistentemente acima das demais. Esse resultado indica que a realização de cirurgia exerce um efeito protetor, prolongando o tempo de sobrevivência, especialmente entre as mulheres.

Em contrapartida, os pacientes do sexo masculino que não realizaram cirurgia (*SEXO = 2, CIRURGIA = 1*) exibem as menores probabilidades de sobrevivência, sugerindo maior risco de óbito nesse grupo. Os demais grupos homens operados e mulheres não operadas apresentam curvas intermediárias, refletindo o efeito conjunto do sexo e da cirurgia sobre o tempo de vida.

O formato decrescente das curvas, característico da distribuição Weibull com parâmetro de forma menor que 1, indica que o risco de morte é mais elevado nos períodos iniciais e tende a diminuir com o tempo. Assim, as estimativas obtidas reforçam os resultados apresentados na saída do modelo, confirmando que o sexo e a realização de cirurgia estão significativamente associados à sobrevida dos pacientes com câncer de brônquios e pulmões.

Figura 18 – Resíduo de Cox-Snell - Filtrado pelo CID C34 e pelas variáveis SEXO, CIRURGIA



Fonte: Elaborado pelo autor

A Figura 18 apresenta o gráfico dos resíduos de Cox–Snell obtidos a partir do modelo de regressão Weibull ajustado. Esse gráfico tem como objetivo avaliar a qualidade do ajuste global do modelo, comparando os resíduos empíricos com o comportamento esperado sob um modelo bem ajustado.

A linha tracejada em vermelho representa a referência teórica de um modelo idealmente ajustado, na qual os resíduos deveriam seguir uma distribuição exponencial com média igual a 1.

Quando o modelo é adequado, os pontos empíricos (em preto) tendem a se alinhar próximo a essa linha.

Observa-se, entretanto, que na região inicial as observações acompanham razoavelmente a linha de referência, mas à medida que o valor dos resíduos aumenta, ocorre um afastamento gradual da linha vermelha. Esse comportamento sugere que o modelo Weibull apresenta um ajuste razoável, mas com pequenas discrepâncias nos valores mais altos de tempo de sobrevivência, o que pode indicar a presença de heterogeneidade não capturada ou de covariáveis não incluídas no modelo.

De forma geral, o gráfico indica que o modelo é satisfatório para descrever o comportamento global dos dados, ainda que apresente leve perda de aderência nas extremidades, o que é comum em modelos paramétricos aplicados a populações com alta variabilidade no tempo de sobrevivência.

Conclusão

A presente pesquisa teve como principal objetivo desenvolver um painel interativo baseado nos dados da Fundação Oncocentro de São Paulo (FOSP), integrando técnicas de Análise de Sobrevivência com ferramentas computacionais de visualização de dados. O painel foi implementado utilizando o pacote *Shiny* do software *RStudio*, permitindo ao usuário explorar de forma dinâmica as informações da base de dados, realizar análises descritivas e aplicar modelos estatísticos de sobrevivência, como Kaplan-Meier, Cox e Weibull.

Os métodos de análise de sobrevivência apresentados Kaplan-Meier, Cox e Weibull constituem a base teórica fundamental para o estudo do tempo até a ocorrência de um evento de interesse, como no contexto deste trabalho, que investiga a sobrevivência de pacientes com câncer. A escolha entre modelos paramétricos e semiparamétricos depende das características dos dados e dos objetivos da pesquisa, sendo que ambos oferecem vantagens complementares. O método de Kaplan-Meier é crucial para estimativas não paramétricas da função de sobrevivência, especialmente em dados censurados. Por sua vez, os modelos de regressão de Cox e Weibull permitem incorporar covariáveis, possibilitando avaliar o impacto de fatores clínicos e sociodemográficos sobre o risco de morte. O modelo de Cox, de natureza semiparamétrica, proporciona flexibilidade ao não exigir especificação da forma da função de risco, enquanto o modelo Weibull, de caráter paramétrico, permite descrever explicitamente o comportamento do risco ao longo do tempo.

Os resultados obtidos com o painel confirmam a aplicabilidade dessas metodologias. A análise de Kaplan-Meier possibilitou estimar e comparar curvas de sobrevivência entre grupos, revelando diferenças significativas entre os sexos e entre pacientes submetidos ou não à cirurgia. O modelo de regressão de Cox permitiu identificar fatores de risco de forma multivariada, enquanto o modelo de Weibull forneceu uma alternativa paramétrica para representar a variação temporal do risco. A inclusão dos resíduos de Schoenfeld e Cox–Snell possibilitou avaliar a validade dos pressupostos e a qualidade de ajuste dos modelos, garantindo maior confiabilidade às inferências realizadas.

Do ponto de vista computacional, o painel desenvolvido demonstra o potencial de ferramentas interativas na democratização do acesso à informação, tornando a análise estatística mais acessível a pesquisadores e profissionais da saúde. A integração entre estatística e tecnologia promove maior transparência e compreensão dos resultados, mesmo para usuários sem conhecimento técnico avançado, ampliando o alcance e a aplicabilidade das análises realizadas.

Como contribuição prática, o painel transforma uma base de dados complexa e extensa em um ambiente visualmente intuitivo e interativo, capaz de evidenciar padrões, comparar grupos e compreender fatores associados à sobrevida de pacientes com câncer. Essa iniciativa contribui

para o fortalecimento de práticas baseadas em evidências e pode servir de suporte tanto à pesquisa científica quanto à gestão em saúde pública.

Como perspectivas para trabalhos futuros, recomenda-se expandir o painel com novas variáveis clínicas e sociodemográficas, considerar outros modelos paramétricos como: exponencial, gaussiano, logistico, lognormal e loglogistico, além de explorar metodologias mais avançadas, como as de fração de cura, riscos competitivos e de fragilidade. Melhorias na interface incluindo a geração automática de relatórios e a exportação de resultados também podem ampliar sua utilidade para diferentes públicos.

Em síntese, o desenvolvimento deste painel interativo representa uma contribuição relevante para a análise e visualização dos dados da FOSP, unindo rigor estatístico, acessibilidade e inovação tecnológica. O trabalho evidencia o potencial da Análise de Sobrevivência como ferramenta essencial para o estudo de desfechos clínicos e destaca a importância de soluções computacionais interativas no apoio à pesquisa e à tomada de decisão em saúde.

Referências

CHANG, W. *et al.* **shiny: Web Application Framework for R**. [S.l.], 2024. R package version 1.10.0. Disponível em: <https://CRAN.R-project.org/package=shiny>. Citado 2 vezes nas páginas 20 e 21.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006. Citado 3 vezes nas páginas 15, 16 e 19.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado na página 16.

FONSECA, A.; PEREIRA, R. H. M. *et al.* **geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil**. [S.l.], 2021. R package version 1.7.0. Disponível em: <https://CRAN.R-project.org/package=geobr>. Citado 2 vezes nas páginas 19 e 27.

Fundação Oncocentro de São Paulo. **Registro Hospitalar de Câncer – Download de Arquivos**. 2024. <https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/download-de-arquivos>. Acesso em: 26 ago. 2025. Citado 5 vezes nas páginas 11, 21, 26, 27 e 28.

GROLEMUND, G.; WICKHAM, H. Dates and times made easy with lubridate. **Journal of Statistical Software**, v. 40, n. 3, p. 1–25, 2011. Disponível em: <https://www.jstatsoft.org/v40/i03/>. Citado na página 21.

HOSMER STANLEY LEMESHOW, S. M. D. **Applied Survival Analysis: Regression Modeling of Time to Event Data**. 2nd. ed. [S.l.]: Wiley, 2008. Citado na página 17.

Instituto Nacional de Câncer. **Institucional: Sobre o INCA**. 2025. <https://www.gov.br/inca/pt-br>. Acesso em: 04 nov. 2025. Citado na página 11.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. [S.l.]: John Wiley & Sons, 2002. Citado na página 19.

KASSAMBARA, A.; KOSINSKI, M.; BIECEK, P. **survminer: Drawing Survival Curves using 'ggplot2'**. [S.l.], 2024. R package version 0.5.0. Disponível em: <https://CRAN.R-project.org/package=survminer>. Citado na página 19.

MARQUES, C. **Oncologia: uma abordagem multidisciplinar**. [S.l.]: Carpe Diem, 2016. Citado na página 11.

MEDEIROS, J. P.; NOGUEIRA, M. C. Construção de um dashboard para análise dos dados de câncer na macrorregião de saúde do sudeste de minas gerais. **Revista Brasileira de Cancerologia**, v. 69, n. 4, 2023. Citado na página 12.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>. Citado na página 22.

RIBEIRO, F. de A. *et al.* Investigação de padrões epidemiológicos em oncologia pediátrica a partir de integração e visualização de dados. In: SBC. **Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)**. [S.l.], 2025. p. 677–688. Citado na página 11.

SHIMAOKA, A. M. *et al.* Big data na saúde pública: Análise do ecossistema das bases epidemiológicas no brasil: Big data in public health: Analysis of the epidemiological database ecosystem in brazil. **Revista de Epidemiologia e Saúde Pública-RESP**, v. 3, n. 1, p. 167–177, 2025. Citado na página 11.

SIEVERT, C. **Interactive Web-Based Data Visualization with R, plotly, and shiny**. Chapman and Hall/CRC, 2020. ISBN 9781138331457. Disponível em: <https://plotly-r.com>. Citado na página 19.

THERNEAU, T. M. **A Package for Survival Analysis in R**. [S.l.], 2024. R package version 3.8-3. Disponível em: <https://CRAN.R-project.org/package=survival>. Citado na página 19.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>. Citado 2 vezes nas páginas 19 e 25.

WICKHAM, H. *et al.* **dplyr: A Grammar of Data Manipulation**. [S.l.], 2023. R package version 1.1.4. Disponível em: <https://CRAN.R-project.org/package=dplyr>. Citado na página 21.