Universidade de São Paulo
Instituto de Matemática e Estatística

# Centro de Estatística Aplicada

# Relatório de Análise Estatística

RAE-CEA–21P30

# RELATÓRIO DE ANÁLISE ESTATÍSTICA SOBRE O PROJETO:

## “Fatores prognósticos em carcinoma de canal anal localizado”

**Aline Duarte de Oliveira**

**Augusto Kira Pedroso de Lima**

**Lucas Abrahão de Paiva**

**Lucas Belleza Spina**

**São Paulo, dezembro de 2021**

**CENTRO DE ESTATÍSTICA APLICADA - CEA – USP**

**TÍTULO:** Relatório de Análise Estatística sobre o Projeto: “Fatores prognósticos em carcinoma de canal anal localizado”.

**PESQUISADOR(A):** Dra. Rachel Riechelmann

**ORIENTADOR(A):**

**INSTITUIÇÃO:** AC Camargo Cancer Center

**FINALIDADE DO PROJETO:** Publicação

**RESPONSÁVEIS PELA ANÁLISE:** Aline Duarte de Oliveira
Augusto Kira Pedroso de Lima
Lucas Abrahão de Paiva
Lucas Belleza Spina

**REFERÊNCIA DESTE TRABALHO:** DUARTE, A.; PEDROSO DE LIMA, A.K.; PAIVA, L.A.; SPINA, L.B. Relatório de análise estatística sobre o projeto**: “Fatores prognósticos em carcinoma de canal anal localizado”.** São Paulo, IME-USP, 2021. (RAE–CEA-21P30)

**FICHA TÉCNICA**

**REFERÊNCIAS BIBLIOGRÁFICAS:**

COLOSIMO, E.A.; GIOLO, S.R. (2006). **Análise de Sobrevivência Aplicada**. Editora Blucher

FINE, J.P.; GRAY, R.J.(1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. **Journal of the American Statistical Association, 94**, 496-509.

GRAMBSCH, P.M.; THERNEAU, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. **Biometrika, 81, 3**, 515-526

GRAY, R.J. (1998). A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. **The Annals of Statistics, 16**, 1141-1154

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. (2021). **An Introduction to Statistical Learning with Applications in R**. 2.ed. Springer.

KLEIN, J.P.; MOESCHBERGER, M.L. (2013). **Survival Analysis Techniques for Censored and Truncated Data**, 2.ed. Springer.

MORETTIN, P.A.; BUSSAB, W.O. (2017). **Estatística Básica**. 9.ed. Saraiva Uni.

SCRUCCA, L.; SANTUCCI, A.; AVERSA, F. (2010). Regression modeling of competing risk using R: an in depth guide for clinicians. **Bone Marrow Transplantation, 45**, 1388-1395.

THERNEAU, T.M.; GRAMBSCH, P.M. (2000). **Modeling Survival Data: Extending the Cox Model**. Springer.

**PROGRAMAS COMPUTACIONAIS UTILIZADOS:**

Microsoft Word for Windows (versão 2016)

Microsoft Excel for Windows (versão 2016)

R for Linux, versão 4.1.2

R for Windows, versão 4.1.2

RStudio for Linux, versão 1.4.1717

RStudio for Windows, versão 1.4.1717

**TÉCNICAS ESTATÍSTICAS UTILIZADAS**

Análise Descritiva Multidimensional (03:020)

Análise de Sobrevivência (13:070)

Regressão Logística (07:090)

Imputação (07:990)

**ÁREA DE APLICAÇÃO**

Bioestatística (14:030)

Oncologia (14:990)

**Resumo**

De acordo com o Instituto Nacional do Câncer (INCA), o termo câncer pode abranger mais de 100 tipos de doenças malignas, das quais a principal característica é o

crescimento rápido e desordenado de células agressivas que podem invadir tecidos e órgãos adjacentes, determinando a formação de tumores nessas partes. Neste relatório é considerado apenas pacientes com câncer do canal anal, o qual apresenta como histologia mais comum o carcinoma epidermoide do canal anal, conhecido como CCA.

O conjunto de dados utilizado foi o da Fundação Oncocentro de São Paulo; nele estão disponíveis as informações características da doença do indivíduo e também atributos do paciente, como idade, sexo, setor pelo qual foi tratado (público e privado), tratamento, etc.

O objetivo do estudo é investigar as potenciais relações de tratamento e de setor com a sobrevida do paciente diagnosticado com CCA. Essa relação foi estudada por meio do modelo de riscos competitivos de Fine e Gray, que apontou não existir diferenças estatisticamente significativas na sobrevida de pacientes tratados pelo setor público ou privado, porém apontou diferenças entre os tipos de tratamentos aplicados ao paciente com CCA.

## Sumário

## 1. Introdução

Nesse trabalho, o interesse é estudar o câncer do canal anal, cuja histologia mais comum é o carcinoma epidermóide do canal anal (CCA). O CCA é uma neoplasia rara, que representa aproximadamente 2,6% dos tumores que acometem o trato gastrointestinal. No Brasil, os registros dessa doença variam bastante, dificultando a avaliação do impacto dela nessa população.

A fim de estudar a sobrevida dos pacientes com CCA, foram selecionados 1984 casos referentes a pacientes com diferentes estádios da doença, registrados no banco de dados da Fundação Oncocentro de São Paulo (FOSP), entre os anos de 2000 e 2016. Esses casos são oriundos de 77 hospitais do estado de São Paulo. O objetivo é avaliar fatores associados à menor sobrevida nesses pacientes com ênfase nas comparações dos desfechos por tratamento padrão (pelo menos quimiorradioterapia) ou não-padrão, e categoria de atendimento (setor público ou privado).

## 2. Objetivo(s)

Os principais objetivos desse projeto são:

1. avaliar a sobrevida em pacientes com CCA tratados com pelo menos quimiorradioterapia e comparar com aqueles submetidos a outros tratamentos;
2. avaliar fatores associados à sobrevida em pacientes com CCA com ênfase na comparação daqueles tratados nos setores público e privado.

## 3. Descrição do estudo

O estudo visa avaliar o impacto do câncer do canal anal na população brasileira analisando os casos registrados no estado de São Paulo. Para tal, será utilizado um banco de dados da Fundação Oncocentro de São Paulo (FOSP) que contém registros sobre pacientes com variados tipos de câncer, de 77 hospitais espalhados pelo estado. Nesse trabalho, foram selecionados somente pacientes com carcinoma epidermóide do canal anal (CCA). Vale ressaltar que o banco de dados cadastrou casos diagnosticados a partir de 01/01/2000 e os registros a partir de 2017 até a realização deste estudo são de casos em andamento e possivelmente terão desfechos diferentes dos registrados até o momento. O banco de dados é atualizado trimestralmente adicionando novos casos e dando seguimento aos cadastrados previamente, sendo os dados utilizados para a análise aqueles disponibilizados pela FOSP no dia 31 de março de 2021.

Variáveis como o tratamento aplicado ao paciente e a categoria de atendimento pela qual ele foi tratado serão utilizadas para avaliar o impacto da doença na sobrevida dos pacientes. Em relação às variáveis de interesse, o tratamento foi dividido em duas categorias, tratamento padrão, que inclui todos os pacientes submetidos a quimiorradioterapia com ou sem a inclusão de outro procedimento clínico, e o tratamento não-padrão, que inclui aqueles pacientes submetidos a qualquer outro tratamento sem a presença de quimiorradioterapia. Além dessas variáveis de interesse, também foram registradas datas importantes de acompanhamento do paciente, como a data do diagnóstico do CCA, do início do tratamento e a da última informação obtida do paciente. A sobrevida, em dias, é dada pela diferença entre as datas do início do tratamento e da última informação do paciente. O evento de interesse é o óbito por CCA, sendo óbito por outras causas um evento competitivo.

Um dos desafios relacionados ao conjunto de dados é a presença de informações faltantes em uma das variáveis de interesse, a categoria de atendimento.

## 4. Descrição das variáveis

1. **Escolaridade:** ANALF (Analfabeto), EFI (Ensino Fundamental incompleto), EFC (Ensino Fundamental Completo), EM (Ensino Médio), ES (Ensino superior), Não Informado.
2. **Idade (anos)**
3. **Idade categorizada:** Mais de 70 anos, 70 anos ou menos
4. **Sexo**: Masculino, Feminino
5. **Setor:** Público (SUS), Privado (Convênio e Particular), Não Informado
6. **Sobrevida (dias)**: tempo entre início do tratamento e a última informação obtida do paciente
7. **Tratamento**: Tratamento padrão, Tratamento não-padrão
8. **Desfecho**: Vivo com câncer; Vivo SOE (sem outras especificações); Óbito por câncer; Óbito SOE
9. **Estádio (I, II, III)**: estádio no qual a doença se encontra no paciente
10. **Tempo entre diagnóstico e início do tratamento**: 60 dias ou mais, Menos de 60 dias
11. **Dias entre diagnóstico e início do tratamento (dias)**
12. **Evento**: 0 - não se observou óbito do paciente, 1 - observou-se óbito do paciente por câncer, 2 - observou-se óbito do paciente por outras causas

## 5. Suposições

Por se tratar de uma amostra de conveniência, para realizar a análise, são necessárias algumas suposições. São elas:

- covariáveis constantes ao longo do tempo;
- acompanhamento contínuo dos pacientes;
- não há viés na falta de registro das categorias de atendimento.

## 6. Análise estatística

A análise estatística foi dividida em três partes: análise descritiva, análise inferencial e análises descritiva e inferencial pós imputação dos dados.

No total, são 2314 registros de pacientes, dos quais 1797 apresentam estádio da doença I, II ou III. O estádio IV não é levado em consideração na análise pois o tratamento considerado padrão é diferente nesse estádio em comparação com os demais. Para as análises descritiva e inferencial, foram utilizados esses 1797 registros, enquanto que na parte de imputação, foram considerados todos os registros com categoria de atendimento informada dentre os 2314, um total de 1416 registros.

### 6.1 Análise descritiva e imputação dos dados

Nesta subseção é apresentada a análise descritiva dos dados, que fornece uma visão inicial dos resultados do estudo (Bussab e Morettin, 2013).

Das 13 variáveis estudadas, 10 são qualitativas e 3 são quantitativas. Foram observados números elevados de informações faltantes (*missings*) para duas variáveis, com 672 (37%) *missings* para setor e 504 (28%) para escolaridade. Não há informações faltantes para as demais variáveis. Foi criada a categoria “não informado” para os casos de *missings* da variável escolaridade, visto que não há interesse em fazer a imputação para ela.

A Tabela A.1 dá início à análise descritiva univariada. Ela mostra a frequência da última informação dos pacientes. Dos 1797 indivíduos, 1007 estavam vivos ao final do

acompanhamento, dentre os quais 759 estavam livres do CCA. Também foram observados 790 óbitos, sendo o CCA a causa de 603 deles.

Na Tabela A.2 pode-se observar a proporção de pacientes em cada estádio por setor. Percebe-se que as proporções no setor II e III para pacientes tratados pelo setor privado são semelhantes, enquanto no setor público mais da metade dos pacientes encontram-se no estádio III da doença. Para verificar se há diferença estatisticamente significativa entre as proporções dos estádios nos setores público e privado, foi feito um teste qui-quadrado (Bussab e Morettin, 2013) que não rejeitou a hipótese nula de que as distribuições são equivalentes, ao nível de 5% ($p = 0,119$).

Dos 1797 pacientes registrados no banco de dados, 1359 (75,6%) foram submetidos ao tratamento padrão, que é composto por pelo menos quimiorradioterapia (pode ou não incluir cirurgia) e 438 (24,4%) foram submetidos ao tratamento não-padrão. Ou seja, de uma forma geral, em torno de ¾ dos pacientes se submetem ao tratamento padrão.

A categoria de atendimento mais frequente foi o SUS, com 1013 pacientes registrados; apenas 5 pacientes registrados foram tratados pelo particular, 107 pelo convênio e para 672 pacientes não se tem registro da categoria pela qual foram atendidos (incluídos na categoria não informado). Devido ao baixo número de pacientes atendidos pelo particular, foi criada a variável setor, que une as categorias de atendimento “convênio” e “particular” em uma categoria chamada “privado”; o SUS se enquadra no setor “público” e a setor “não informado” foi utilizado para lidar com os casos em que a categoria de atendimento não foi informada.

A Figura B.1 exibe a frequência dos estádios dos pacientes. Pode-se perceber uma grande quantidade de pacientes com a doença nos estádios II e III (somam mais de 90% do total), com o restante pertencente ao estádio I.

A análise descritiva multidimensional começa com a Figura B.2, em que se verifica as frequências relativas dos estádios da doença pela idade categorizada dos pacientes. O gráfico sugere que as distribuições dos estádios de pacientes com 70 anos ou menos

e com mais de 70 anos são parecidas, com uma concentração de pacientes nos estádios II e III da doença, similar ao que foi visto na Figura B.1.

Na Figura B.3 estão apresentadas as frequências relativas dos desfechos para cada setor. Comparando as frequências dos setores público e privado, verifica-se que o setor privado apresenta proporções menores para todos os desfechos, exceto para pacientes vivos sem CCA (19% maior do que o setor público). A forma das distribuições desses dois setores é similar, isto é, a maior frequência relativa é de pacientes vivos sem a doença, seguida por óbitos por CCA, vivo com câncer e óbito por outras causas. A distribuição dos desfechos quando o setor não é informado tem formato diferente das demais, com uma proporção de óbitos por CCA de 44%, 10% maior do que a de pacientes vivos livres do câncer.

A Figura B.4 apresenta a frequência relativa de cada desfecho nos estádios para os setores não informado, privado e público. Nota-se uma tendência de queda nas proporções de pacientes vivos sem CCA (linha roxa) e uma tendência de crescimento nas proporções de óbito pelo câncer (linha vermelha) conforme o estadiamento da doença avança, independente do setor. As proporções de óbitos por outras causas e de pacientes vivos com câncer têm comportamento quase constante de um estadiamento para o próximo.

A Figura B.5 mostra proporções maiores de pacientes submetidos ao tratamento padrão, independente da idade, exceto para pacientes com mais de 70 anos e doença no estádio I. Nos estádios II e III da doença percebe-se proporções ainda maiores do tratamento padrão, chegando a 82%. As proporções dos pacientes com 70 anos ou menos tratados pelo padrão são maiores do que as dos pacientes com mais de 70 anos em todos os estádios.

Na Figura B.6 não se observa um efeito do tempo entre diagnóstico e início do tratamento nos desfechos por estádio. Ao contrário do que se esperava, as proporções de óbitos por câncer foram maiores quando o paciente aguardou menos de 60 dias para iniciar o tratamento.

As Figuras B.7 a B.13 são curvas da taxa de incidência acumulada (Klein e Moeschberger, 2013) dos dois óbitos observados no conjunto de dados (por câncer e por outras causas) separando por diferentes covariáveis. As curvas de incidência acumulada consideram os dois possíveis eventos na sua construção e por isso são mais apropriadas para o problema do que as curvas de Kaplan-Meier (Colosimo e Giolo, 2006) para o caso de riscos competitivos (Colosimo e Giolo, 2006). Os comentários relativos às figuras mencionadas no início do parágrafo têm como foco as curvas referentes ao óbito por câncer (curvas na cor vermelha), que é o evento de interesse do estudo. Para testar a igualdade entre as curvas, foi utilizado o teste de Gray (Gray, 1988), cuja hipótese nula é de que todas as curvas para um mesmo evento são iguais.

A Figura B.7 mostra o comportamento da taxa de incidência acumulada ao longo do tempo para cada estádio. Ao nível de significância de 5%, o teste de Gray rejeita a igualdade entre as curvas ($p < 0,001$) do óbito por câncer, sugerindo que há associação entre pelo menos um dos estádios e incidência de óbito por CCA. O estádio III apresentou as maiores taxas, seguido pelo estádio II, corroborando o que foi analisado na Figura B.4.

Na Figura B.8 a curva da taxa de incidência acumulada foi separada pelo sexo. Assim como para a covariável estádio, o teste de Gray rejeitou ($p < 0,001$) que as curvas sejam iguais para o sexo, sugerindo que há associação entre sexo e incidência de óbito por CCA. Vale ressaltar que as taxas de incidência do sexo masculino foram maiores do que as do feminino ao longo de todo o tempo observado.

A Figura B.9 apresenta as curvas das taxas de incidência acumulada dos pacientes submetidos ao tratamento padrão e não-padrão. Nota-se que a curva relativa aos pacientes submetidos ao tratamento não-padrão foi maior do que a dos submetidos ao tratamento padrão. O teste de Gray rejeita que as curvas sejam iguais ($p < 0,001$), sugerindo que há associação entre tratamento e incidência de óbito por CCA.

As curvas da Figura B.10 foram separadas pela idade categorizada. O teste de Gray rejeita a igualdade ($p < 0,001$), sugerindo que há associação entre idade categorizada e incidência de óbito por CCA. Os pacientes da categoria acima de 70 anos apresentaram taxas maiores de falha.

Na Figura B.11 observa-se as curvas das taxas de incidência acumulada separadas por tempo entre diagnóstico e início do tratamento (menos de 60 dias ou 60 dias ou mais). As curvas ficaram próximas uma da outra e o teste de Gray não rejeita que elas sejam iguais (p = 0,92), sugerindo que não há associação entre tempo entre diagnóstico e início do tratamento e incidência de óbito por CCA.

A Figura B.12 mostra as curvas das taxas de incidência acumulada por setor. A curva do setor não informado foi maior que a dos demais setores e o teste de Gray rejeita que as curvas três sejam iguais (p < 0,001), sugerindo que há associação entre pelo menos uma das três categorias de setor e incidência de óbito por CCA.

A Figura B.13 exibe as curvas das taxas de incidência acumulada por escolaridade. As seis curvas de escolaridade ficaram próximas entre si, comportamento refletido no valor-p do teste de Gray, que não rejeita que as seis curvas sejam iguais (p = 0,501), sugerindo que não há associação entre escolaridade e incidência de óbito por CCA.

Foi feita uma imputação da categoria do setor para os pacientes que tinham essa informação faltando. Todo o processo está descrito em detalhes no Apêndice C com referência de James et al, 2021. Considerando o modelo de regressão logística com escolaridade, idade, tempo entre diagnóstico e início do tratamento, tratamento, estádio e sexo como covariáveis, a Tabela A.3 mostra a matriz de confusão no conjunto de validação utilizando 0,881 como ponto de corte (encontrado no conjunto de treinamento), pode-se concluir que não há *overfit* e portanto o modelo pode ser ajustado no conjunto de treino completo.

Ao ajustar o modelo final em todos os dados de treino disponíveis, foram obtidas as estimativas de cada coeficiente, que se encontram na Tabela A.4; a curva ROC do ajuste final é apresentada na Figura B.14. O melhor ponto de corte de acordo com a sensibilidade e a especificidade foi 0,846 e a AUC do ajuste final foi 0,822.

Após realizar a imputação das categorias de setor, dos 1797 pacientes, tem-se que 1500 (83,5%) pertencem à categoria pública de setor e 297 (16,5%) à categoria privada.

A Figura B.15 mostra as curvas das taxas de incidência acumulada separando por setor, após a imputação. Vale notar que não há a categoria “não informado” de setor. A curva relativa ao setor público ficou acima da curva do setor privado e o teste de Gray não rejeita ao nível de significância de 5% que as duas curvas sejam iguais (p = 0,051), sugerindo que não há associação entre setor e incidência de óbito por CCA.

### 6.2 Análise inferencial

Nesta subseção, é apresentada a análise inferencial dos dados.

Na seção Descrição do estudo foi mencionado que o conjunto de dados apresentava dois eventos distintos de óbito: por CCA ou por uma causa não especificada. Dessa forma, o paciente que vem a óbito pelo câncer, não pode posteriormente vir a óbito por outra causa e vice-versa, estabelecendo assim um cenário de riscos competitivos. Para lidar com esse tipo de situação, um modelo proposto é o de Fine e Gray (Fine e Gray, 1999), o qual foi utilizado na análise desse estudo.

Inicialmente, foi ajustado um modelo considerando as covariáveis escolaridade, idade categorizada, sexo, setor, tratamento, estádio, tempo entre diagnóstico e inicio do tratamento e interação entre tratamento e estádio do paciente, para avaliar as associações nas taxas de falha do óbito por CCA. O procedimento para a seleção do modelo final foi retirar um a um os coeficientes que não foram estatisticamente significativos ao nível de 5%, com exceção da variável setor por ser de interesse do estudo. A Tabela A.5 mostra as estimativas obtidas para o conjunto de dados sem a imputação do setor (criou-se a categoria “Não informado” para as informações faltantes) do modelo pós seleção de covariáveis. Percebe-se que ao nível de 5% todas foram significativas, exceto o setor privado.

A principal suposição do modelo de Fine e Gray é a proporcionalidade das taxas de falha (Colosimo e Giolo, 2016), a qual pode ser verificada através dos gráficos dos resíduos de Schoenfeld construídos para os níveis de cada covariável categórica e para cada covariável numérica do modelo. Pelas Figuras B.16 e B.17, não se tem evidências de que a suposição de proporcionalidade não seja válida para nenhuma das covariáveis

do modelo final, pois os resíduos de Schoenfeld de cada covariável mostra uma tendência aproximadamente em paralelismo, evidenciada pelo *spline*; portanto, as estimativas obtidas são confiáveis. Os capítulos 4 e 6 do livro *Modeling Survival Data: Extending the Cox Model* (Therneau e Grambsch, 2000) trazem mais informações sobre os resíduos de Schoenfeld, que foram introduzidos no artigo *Proportional hazard tests and diagnostics based on weighted residuals* (Grambsch e Therneau, 1994).

A interpretação das estimativas dos coeficientes do modelo de Fine e Gray é dada a partir das razões de taxa falha específica, no caso, óbito por CCA. Pelo ajuste obtido, observa-se que a taxa de falha específica do óbito por câncer em pacientes submetidos ao tratamento não-padrão é 1,27 vezes a de pacientes submetidos ao tratamento padrão, ou seja, um aumento de 27%; para pacientes do sexo masculino, essa taxa de falha específica foi 1,74 vezes a do sexo feminino, ou seja, um aumento de 74%; a taxa de falha específica para pacientes com mais de 70 anos é 1,63 a de pacientes com 70 anos ou menos, 63% maior. Tratamento padrão, sexo feminino e idade categorizada 70 anos ou menos foram consideradas como as categorias de referência. A taxa de falha específica do óbito por câncer de pacientes com estádio II da doença é 1,68 vezes a de pacientes com estádio I, 68% maior; para pacientes do estádio III, a taxa de falha específica é 3,62 vezes a de pacientes com estádio I, 262% maior. Não foi feita a interpretação dos setores, pois o setor privado não teve coeficiente com diferença estatisticamente significativa em comparação com o setor público ($p = 0,230$) e a comparação do setor público com o setor não informado não faz sentido, visto que a categoria “não informado” foi criada para lidar com as informações faltantes de setor no estudo. A interpretação de cada covariável foi feita considerando valores iguais para as demais covariáveis e sempre considerando o desfecho óbito por câncer.

O mesmo modelo inicial ajustado aos dados sem imputação foi ajustado aos dados com imputação. Novamente, a interação entre tratamento e estádio e as covariáveis escolaridade e tempo entre diagnóstico e tratamento não foram significativas ao nível de 5%. A covariável setor também não foi significativa ao nível de 5%, porém foi mantida no modelo por ser de interesse do estudo. A Tabela A.6 mostra as estimativas obtidas por

máxima verossimilhança parcial do modelo pós seleção de covariáveis para o conjunto de dados com a imputação do setor.

Para verificar a suposição de proporcionalidade das taxas de falha específica do óbito por CCA foram feitos os gráficos dos resíduos de Schoenfeld para os níveis de cada covariável categórica e cada covariável numérica mantida no ajuste, apresentados nas Figuras B.18 e B.19. Assim como para o ajuste aos dados sem imputação, os resíduos de Schoenfeld aos dados com imputação também mostram a tendência de paralelismo, indicando que as estimativas são confiáveis.

Pelo ajuste aos dados com imputação, tem-se que a taxa de falha específica do óbito por câncer de pacientes submetidos ao tratamento não-padrão é 1,33 vezes a de pacientes submetidos ao tratamento padrão, ou seja, um aumento de 33%; as estimativas da categoria “masculino” da covariável sexo e “mais de 70 anos” da covariável idade categorizada foram essencialmente as mesmas que no modelo ajustado aos dados sem imputação. Novamente, tratamento padrão, sexo feminino e idade categorizada 70 anos ou menos foram consideradas como categorias de referência. A taxa de falha específica do óbito por câncer de pacientes com estádio II da doença é 1,75 vezes a de pacientes com estádio I, 75% maior; para pacientes do estádio III, a taxa de falha específica é 3,53 vezes a de pacientes com estádio I, 253% maior. No ajuste aos dados com imputação, o setor privado não teve coeficiente com diferença estatisticamente significativa em comparação com o setor público ($p = 0,430$). A interpretação de cada covariável foi feita considerando valores iguais para as demais covariáveis e sempre considerando o desfecho óbito por câncer.

Os dois ajustes (aos dados com e sem imputação) não mostraram diferenças inferenciais e as estimativas dos coeficientes de ambos também foram muito próximas, sugerindo que a imputação não influenciou o efeito de nenhuma covariável.

## 7. Conclusão

Pela análise feita utilizando o modelo de Fine e Gray, pode-se concluir que não há efeito de interação entre tipo de tratamento estádio, mas há uma diferença estatisticamente significativa ao nível de 5% na taxa de falha específica do óbito por CCA para pacientes submetidos aos tratamentos padrão e não padrão, com os pacientes submetidos ao tratamento não padrão associados a uma maior taxa de falha por CCA.

Não foi possível concluir que há diferença significativa ao nível de 5% entre pacientes tratados pelo setor privado em relação ao público, mesmo com a imputação de setor nos dados faltantes.

# APÊNDICE A

## Tabelas

**Tabela A.1**: Frequência do desfecho observado do paciente.

| Desfecho | N° de ocorrências | Frequência relativa |
|---|---|---|
| Óbito por câncer | 603 | 33,6% |
| Óbito SOE | 187 | 10,4% |
| Vivo com câncer | 248 | 13,8% |
| Vivo SOE | 759 | 42,2% |

**Tabela A.2**: Proporção de estádios por setor.

| | | Estádio | | |
|---|---|---|---|---|
| | | **I** | **II** | **III** |
| **Setor** | **Privado** | 13% | 43% | 44% |
| | **Público** | 10% | 37% | 53% |
| | **Não informado** | 9% | 51% | 40% |

**Tabela A.3**: Matriz de confusão obtida a partir do conjunto de validação.

| **Ponto de corte:** 0,881 | | **Observado** | |
|---|---|---|---|
| | | **Convênio** | **SUS** |
| **Predito** | **Privado** | 33 | 91 |

| | | | |
|---|---|---|---|
| | **Público** | 15 | 287 |

**Tabela A.4**: Estimativas modelo de regressão logística para imputação.

| Coeficientes | Estimativas | Erro Padrão | Valor-P |
|---|---|---|---|
| Intercepto | 4,196 | 1,159 | < 0,001 |
| Tratamento padrão | 0,038 | 0,218 | 0,860 |
| EFC | -1,162 | 1,053 | 0,270 |
| EFI | 0,241 | 1,129 | 0,831 |
| EM | -1,846 | 1,045 | 0,077 |
| ES | -4,127 | 1,031 | < 0,001 |
| Escolaridade Não Informada | -2,691 | 1,020 | 0,008 |
| Masculino | 0,328 | 0,219 | 0,134 |
| Idade | -0,009 | 0,008 | 0,269 |
| Dias entre diagnóstico e início do tratamento | 0,006 | 0,002 | < 0,001 |

**Tabela A.5**: Estimativas modelo Fine e Gray para dados sem imputação.

| Coeficientes | Estimativas | exp(coef) | Erro Padrão | Valor-P |
|---|---|---|---|---|
| Tratamento padrão | 0,24 | 1,27 | 0,10 | 0,016 |
| Sexo masculino | 0,56 | 1,74 | 0,09 | <0,001 |
| Setor privado | -0,29 | 0,75 | 0,238 | 0,230 |
| Setor Não informado | 0,42 | 1,52 | 0,09 | < 0,001 |
| Idade categorizada | 0,489 | 1,63 | 0,10 | 0,001 |

| | | | | |
|---|---|---|---|---|
| mais de 70 anos | | | | |
| estádio II | 0,52 | 1,68 | 0,20 | 0,008 |
| estádio III | 1,29 | 3,62 | 0,19 | <0,001 |

**Tabela A.6**: Estimativas modelo Fine e Gray para dados com imputação.

| **Coeficientes** | **Estimativas** | **exp(coef)** | **Erro Padrão** | **Valor-P** |
|---|---|---|---|---|
| Tratamento padrão | 0,29 | 1,33 | 0,10 | 0,004 |
| Sexo masculino | 0,53 | 1,71 | 0,09 | <0,001 |
| Setor privado | -0,09 | 0,91 | 0,12 | 0,430 |
| Idade categorizada mais de 70 anos | 0,51 | 1,67 | 0,10 | 0,001 |
| estádio II | 0,56 | 1,75 | 0,20 | 0,005 |
| estádio III | 1,26 | 3,53 | 0,20 | <0,001 |

# APÊNDICE B

## Figuras

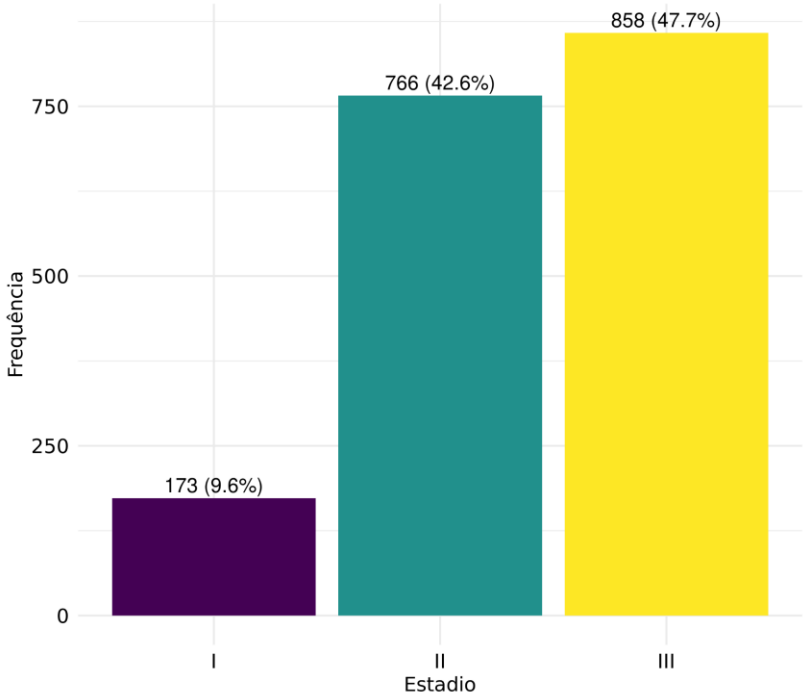


**Figura B.1**: Gráfico de barras do estádio.

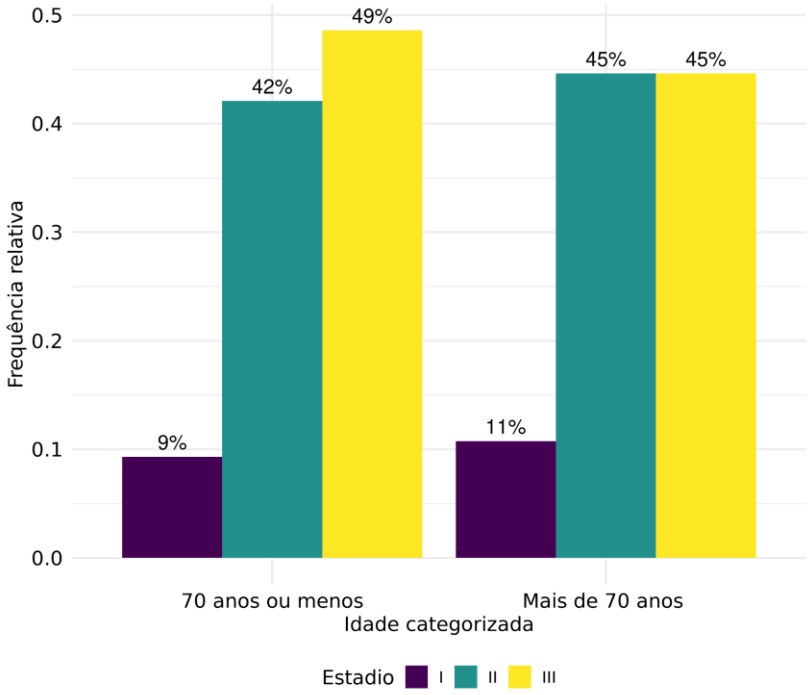

**Figura B.2**: Gráfico de barras da idade categorizada agrupada pelo estádio.

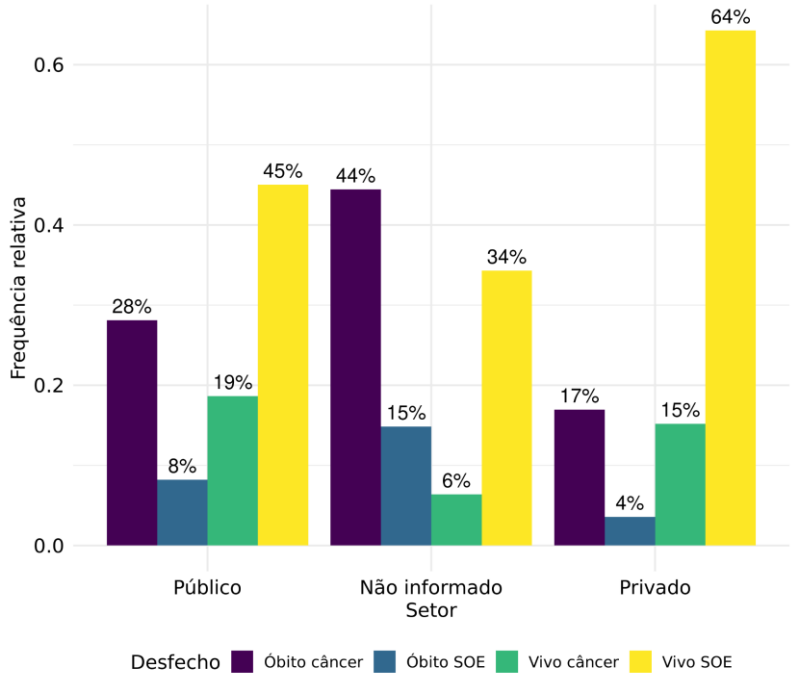


**Figura B.3**: Gráfico de barras do desfecho agrupado por setor.

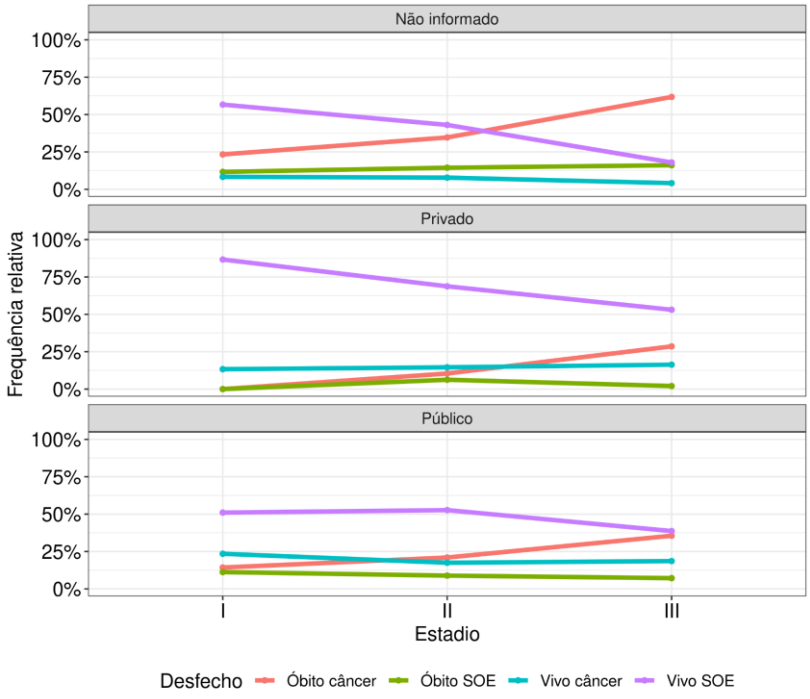


**Figura B.4**: Gráfico de linhas da tendência de cada desfecho nos estádios por setor.

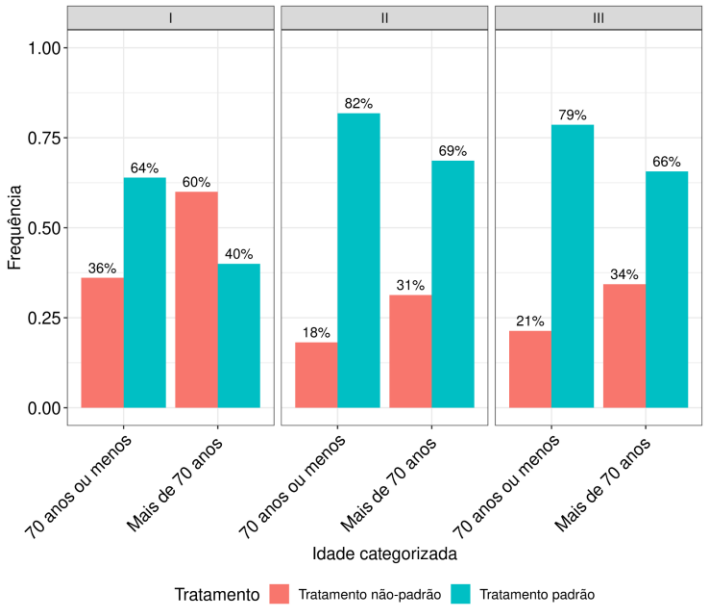


**Figura B.5**: Gráfico de barras do tratamento agrupado pela idade categorizada e estádio.

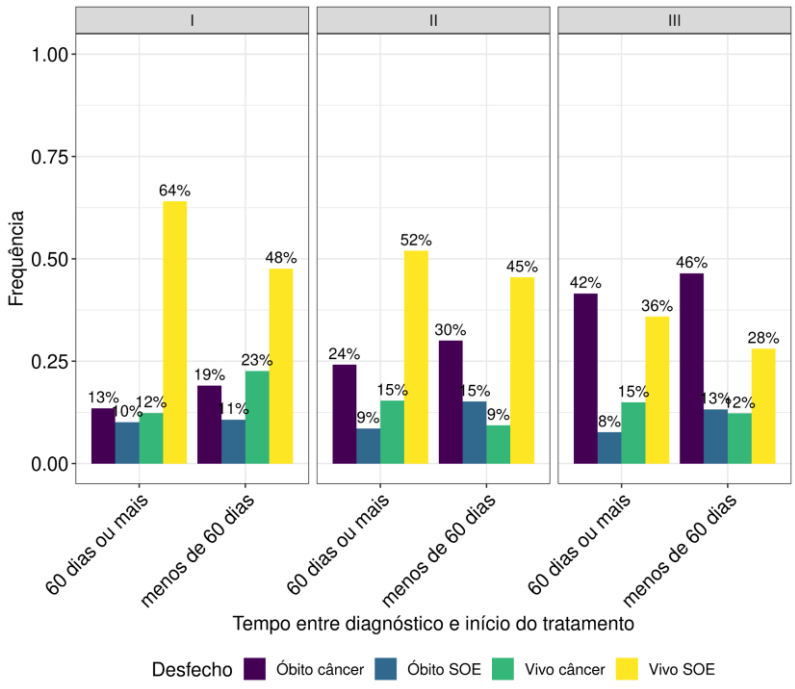


**Figura B.6**: Gráfico de barras do desfecho agrupado pelo tempo entre diagnóstico e início do tratamento e estádio.

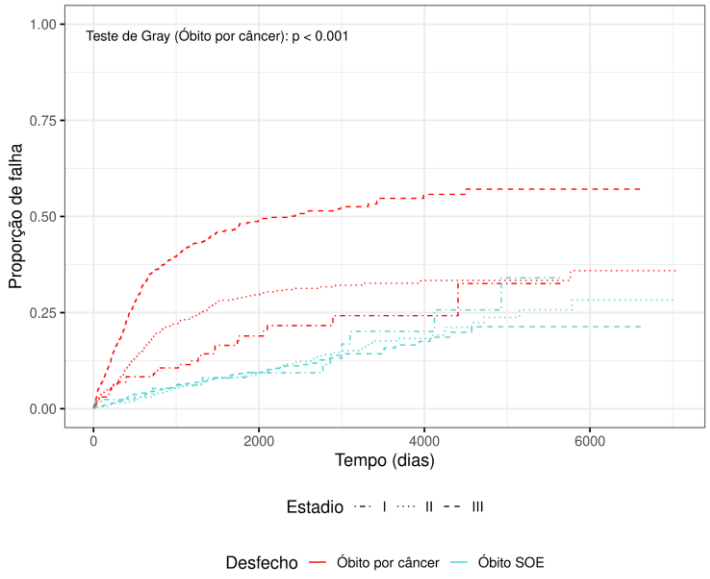


**Figura B.7**: Curvas da taxa de incidência acumulada dos dois óbitos separando por estádio.

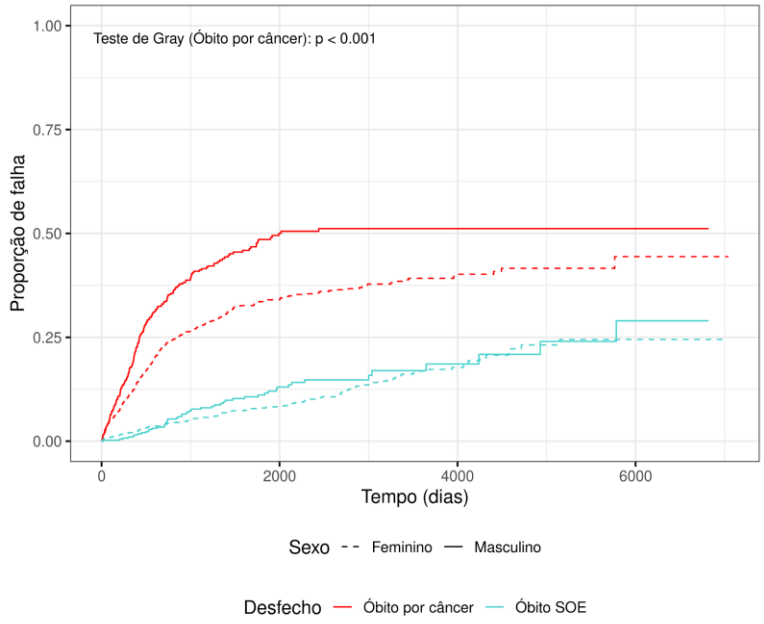


**Figura B.8**: Curvas da taxa de incidência acumulada dos dois óbitos separando por sexo.

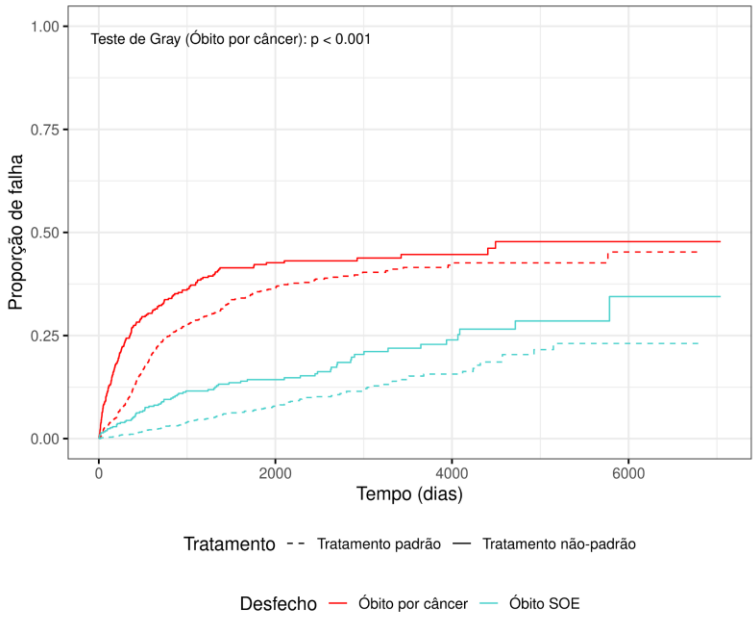


**Figura B.9**: Curvas da taxa de incidência acumulada dos dois óbitos separando por tratamento.

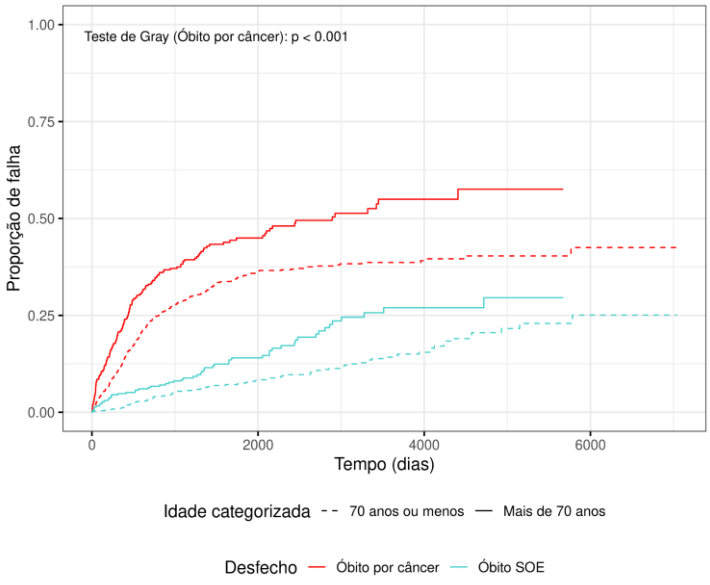


**Figura B.10**: Curvas da taxa de incidência acumulada dos dois óbitos separando pela idade categorizada.

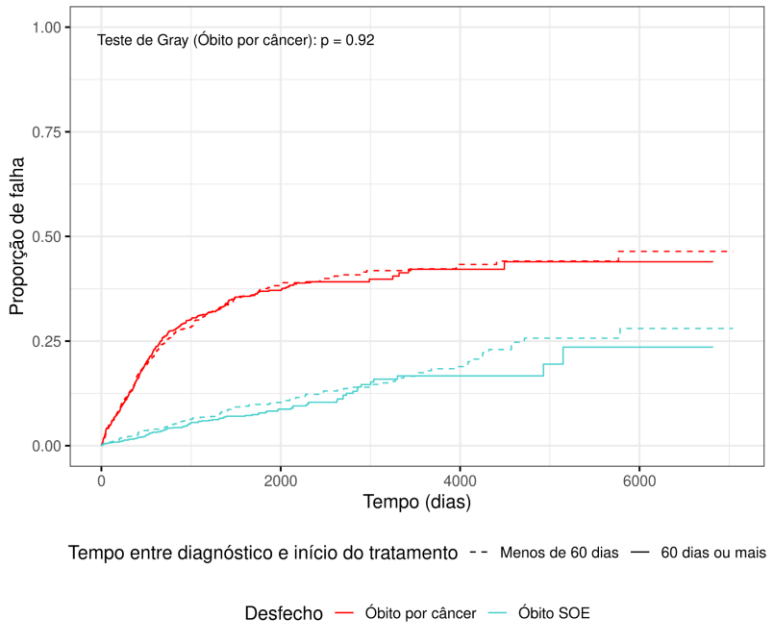


**Figura B.11**: Curvas da taxa de incidência acumulada dos dois óbitos separando por tempo entre diagnóstico e início do tratamento.

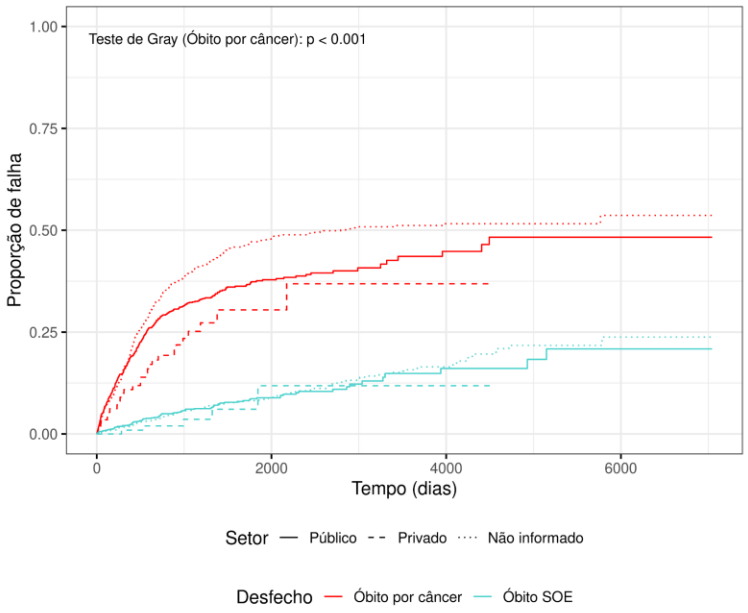


**Figura B.12**: Curvas da taxa de incidência acumulada dos dois óbitos separando por setor (considerando setor não informado).

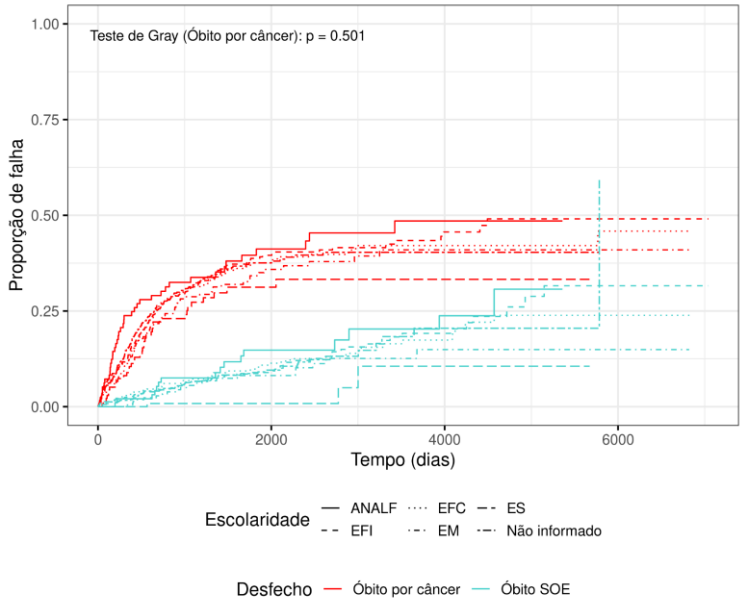


**Figura B.13**: Curvas da taxa de incidência acumulada dos dois óbitos separando por escolaridade.

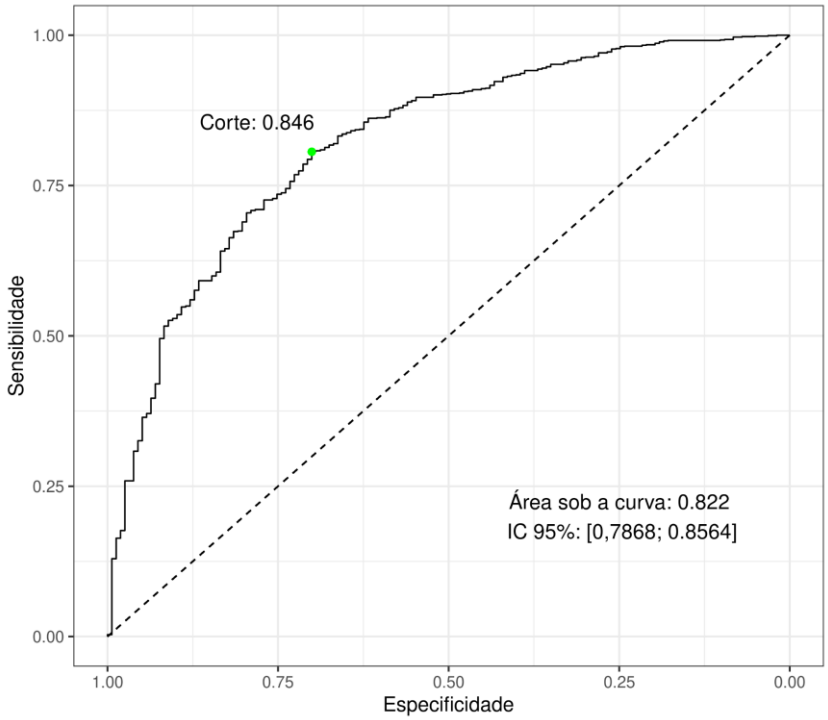


**Figura B.14**: Curva ROC da classificação do modelo final de regressão logística.

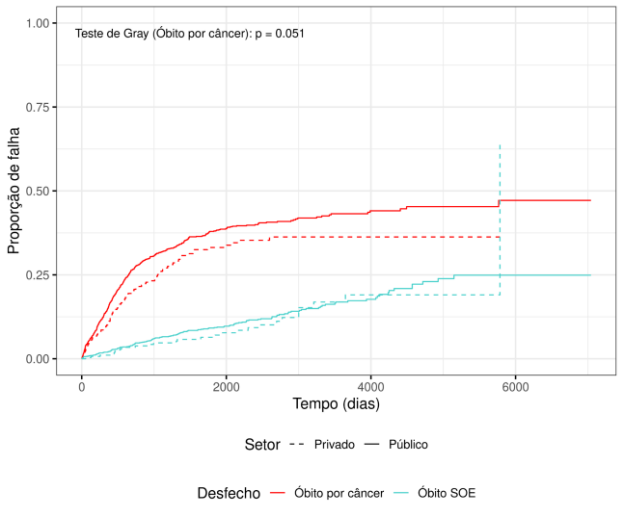


**Figura B.15**: Curvas da taxa de incidência acumulada separando por setor após a imputação.

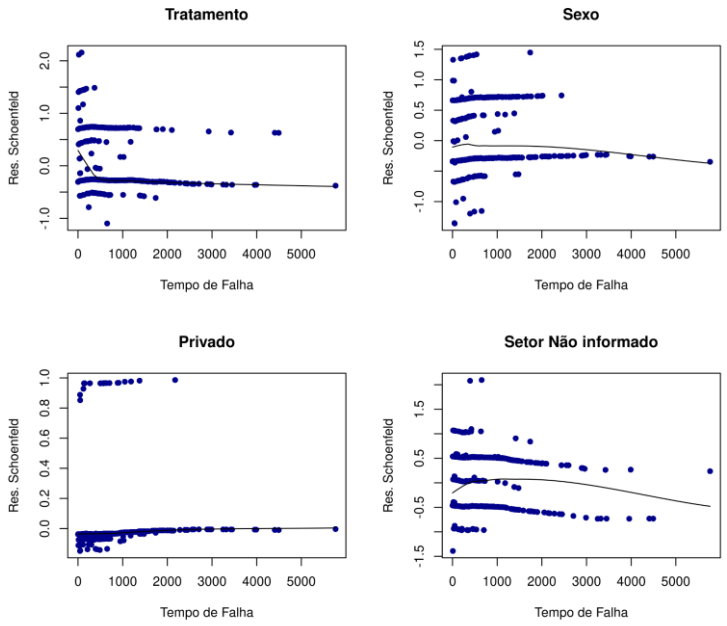


**Figura B.16**: Gráficos dos resíduos de Schoenfeld das covariáveis tratamento, sexo, setor privado e setor não informado do modelo de Fine e Gray ajustado aos dados sem imputação.

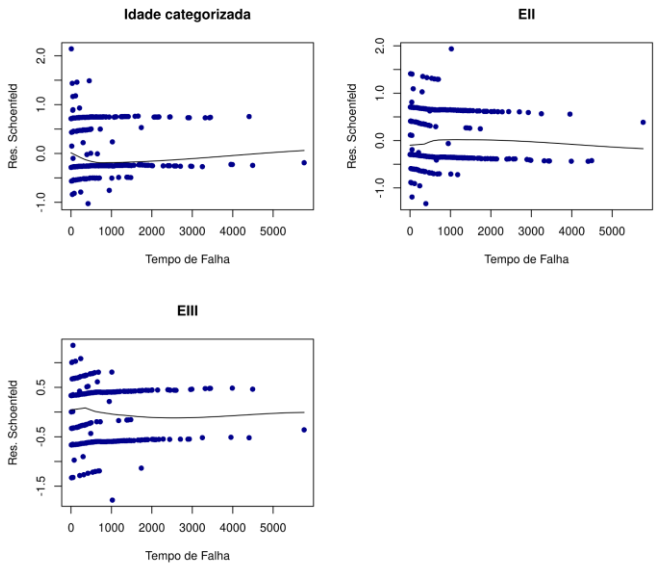


**Figura B.17**: Gráficos dos resíduos de Schoenfeld das covariáveis idade categorizada, estádio II (EII) e estádio III (EIII) do modelo de Fine e Gray ajustado aos dados sem imputação.

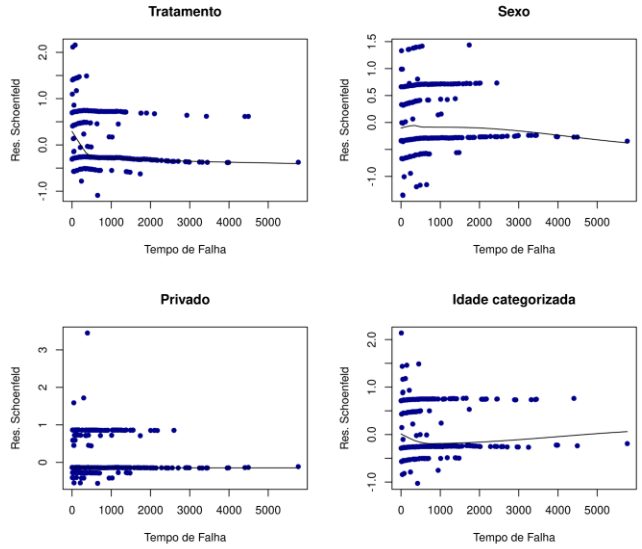


**Figura B.18**: Gráficos dos resíduos de Schoenfeld das covariáveis tratamento, sexo, setor privado e idade categorizada do modelo de Fine e Gray ajustado aos dados com imputação.

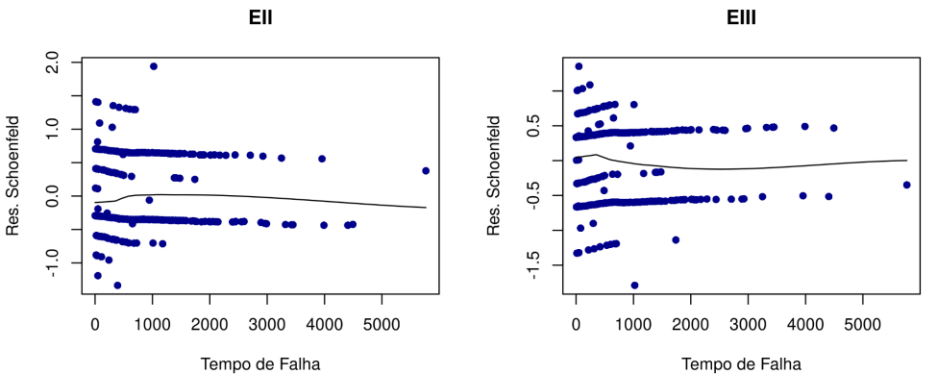


**Figura B.19**: Gráficos dos resíduos de Schoenfeld da covariável estádio II (EII) e III (EIII) do modelo de Fine e Gray ajustado aos dados com imputação.

# APÊNDICE C

## Técnica de imputação

No conjunto de dados disponível há uma grande quantidade de informações faltantes em um dos atributos de interesse dos pacientes com CCA: setor, com duas possíveis categorias, público e privado. Por ser um dos objetivos principais, ao invés de descartar as observações em que esse atributo não foi registrado, foi feita a imputação (classificação) dessa característica para os pacientes em que ela não foi registrada, por meio de uma regressão logística (James et al, 2021), que é um modelo utilizado para lidar com variável-resposta categórica.

A pedido da pesquisadora, a construção e aplicação dessa técnica será descrita em detalhes a seguir.

Para a imputação, foram utilizadas todas as 2314 observações disponíveis, dentre essas, 898 não se têm a indicação de setor, sobrando 1416 observações “completas”, das quais 990 (aproximadamente 70%) foram aleatoriamente selecionadas como conjunto de treinamento e as demais 426 foram selecionadas como conjunto de validação. O objetivo dessa separação é avaliar se o modelo apresenta *overfit* (James et al, 2021); como não houve *overfit* no ajuste com os dados de treinamento, o modelo final foi ajustado aos dados de treino completos (inclui os dados treinamento e validação), e então este modelo final foi utilizado para fazer as predições. Ao utilizar todas as 2314 observações, há a ocorrência de três categorias da variável estádio que foram inicialmente desconsideradas, categorias IV, X e Y, visto que a categoria IV considera tratamento padrão de uma forma diferente das categorias I, II e III e as categorias X e Y são utilizadas quando não se sabe ou não se especifica o estádio do paciente. Uma vez ajustado o modelo de imputação e feitas as predições, os registros dos pacientes com estádios IV, X e Y foram desconsiderados novamente. Como há mais registros de pacientes que são tratados pelo setor público, na divisão em conjuntos de treinamento e validação, a proporção de pacientes dos setores público e privado foi mantida.

A variável em que há interesse de se prever é o setor, que foi a variável resposta da regressão logística. Como um dos objetivos finais é avaliar a associação dessa variável com a sobrevida dos pacientes, não foi utilizado o tempo de sobrevida ou os eventos de interesse para prever qual o setor. Dessa forma, as variáveis utilizadas foram:

escolaridade, idade, dias entre diagnóstico e início do tratamento, tratamento, estádio e sexo, denominadas variáveis explicativas ou covariáveis.

Com as 7 variáveis explicativas, modelou-se a probabilidade do paciente ser do setor público, atribuindo-se um “peso” para cada variável explicativa, denominado coeficiente. Os coeficientes foram estimados pelo método de máxima verossimilhança (Bussab e Morettin, 2013) e como o objetivo dessa regressão logística é a predição, não foi levada em consideração a significância estatística desses coeficientes para o modelo.

Como descrito em James et al (2021), após o treinamento do modelo, é necessário definir um ponto de corte, em que valores com probabilidades estimadas iguais ou acima desse ponto são classificados como “público” e probabilidades estimadas abaixo dele, “privado”. O ponto de corte foi escolhido com base no conjunto em que o modelo foi treinado, ou seja, é o ponto com a melhor combinação de sensibilidade e especificidade para predizer as categorias do conjunto de treino. A escolha dessa métrica para decidir o ponto de corte foi feita por causa da disparidade entre pacientes dos setores público e privado observada no conjunto de dados. Geralmente se monta uma matriz de confusão para visualizar como é a classificação do modelo em relação ao observado, no conjunto de validação; para visualizar o ponto de corte escolhido, pode-se utilizar a curva ROC. O ponto (1, 1), isso é, com sensibilidade e especificidade iguais a 1 seria o de um classificador perfeito; o ponto na curva ROC mais próximo ao (1, 1) é o que tem o melhor ponto de corte do classificador que gerou aquela curva.

Ao ajustar o modelo no conjunto de treinamento, definir o ponto de corte com base no conjunto de treino e realizar as predições no conjunto de validação, não havendo indícios de *overfit*, ajusta-se esse mesmo modelo no conjunto de treino completo (treinamento + validação).

Outra estatística comumente utilizada para avaliar a qualidade de um classificador é a AUC (*Area Under the Curve*) (James et al, 2021), em que valores próximos a 1 são indicadores de um bom classificador. Em geral, há uma linha pontilhada (bissetriz) no gráfico da curva ROC que representa um classificador aleatório, que prediz as categorias ao acaso para cada observação; quando se tem um problema de classificação binária, a

AUC deste classificador é 0,5. Espera-se que um classificador com AUC abaixo de 0,5 tenha poder preditivo pior do que um classificador aleatório. Classificadores com AUC abaixo de 0,8 são, em geral, considerados ruins.

Uma outra forma comum de escolher esse ponto de corte quando há apenas duas classes a serem preditas é pela acurácia (James et al, 2021). A acurácia é calculada como a soma das classificações corretas dividida pelo número total de classificações feitas; usando a matriz de confusão, o cálculo passa a ser a soma da diagonal principal dividido pelo total. Devido ao grande desbalanceamento das duas classes a serem preditas nos dados desse projeto, a acurácia não será utilizada.

Por fim, com o modelo final, dos 898 pacientes a terem o setor imputado, 225 tinham estádio IV, X ou Y, e portanto foram descartados. Utilizando a regressão logística ajustada como classificador e aplicando-a aos 672 pacientes cujo setor não foi informado e o estádio era diferente de IV, X ou Y, 487 deles foram classificados como do setor público e os 185 restantes, do setor privado.