

**MODELO PARAMÉTRICO FLEXÍVEL PARA RISCOS COMPETITIVOS:
UMA ANÁLISE DE SOBREVIVÊNCIA EM UMA COORTE ITALIANA**

Reinalda Souza Oliveira

Tese apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” como parte dos requisitos para obtenção do título de Doutora em Biometria.

BOTUCATU
São Paulo – Brasil
Fevereiro – 2019

**MODELO PARAMÉTRICO FLEXÍVEL PARA RISCOS COMPETITIVOS:
UMA ANÁLISE DE SOBREVIDA EM UMA COORTE ITALIANA**

Reinalda Souza Oliveira

Orientadora: Profa. Dra. **Liciana Vaz de Arruda Silveira**

Coorientador: Dr. **Alberto Rubén Osella**

Tese apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” como parte dos requisitos para obtenção do título de Doutora em Biometria.

BOTUCATU
São Paulo – Brasil
Fevereiro – 2019

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Oliveira, Reinalda Souza.

Modelo paramétrico flexível para riscos competitivos :
uma análise de sobrevida em uma coorte italiana / Reinalda
Souza Oliveira. - Botucatu, 2019

Tese (doutorado) - Universidade Estadual Paulista
"Júlio de Mesquita Filho", Instituto de Biociências de
Botucatu

Orientador: Liciane Vaz de Arruda Silveira

Coorientador: Alberto Rubens Osella

Capes: 10202021

1. Análise de sobrevivência (Biometria). 2. Riscos
competitivos. 3. Distribuição espacial da população.

Palavras-chave: Efeitos dependentes do tempo; Modelos de
sobrevivência flexíveis; Subdistribuição.

Dedicatória

Dedico este estudo a minha mãe, Benilce, e aos meus filhos, João Arthur e João Gabriel.

Agradecimentos

Elaborar uma tese de doutorado não é nada fácil, nem me teria sido possível sem a ajuda e o apoio de várias pessoas no decorrer desse longo período.

Inicialmente, agradeço a Deus por ter me dado esta oportunidade e por ter confiado no meu potencial nos momentos em que nem mesmo eu acreditava.

Agradeço à professora doutora Liciane Vaz de Arruda Silveira, minha orientadora, pelas ideias, pela partilha de conhecimentos, pela disponibilidade e pela paciência durante todo esse período. Muito, muito obrigada!

Ao Dr. Alberto Rubens Osella, que, além de co-orientador, coordena a equipe do Laboratório de Epidemiologia e Bioestatística do Hospital Saverio De Bellis, na Itália, grupo que me recebeu com muito carinho e humildade. Muito obrigada pela motivação diária (cafezinhos), pela empatia, pela partilha de conhecimentos e especialmente pela amizade aqui representada pela nutricionista Antonella Mirizzi.

Agradeço também ao pessoal de T. I. do IBB, pelos socorros com instalação de programas e pela empatia e cuidado com cada um dos alunos. A eles o meu muito obrigada!

Agradeço a d. Selda, funcionária do IBB, pelo carinho e cuidado que sempre teve comigo. Muito obrigada!

Aos colegas e amigos, pelo trabalho fantástico que me ajudaram a desenvolver todos os dias, e pelos momentos de descontração e bom convívio.

Agradeço ao colega e amigo Luiz Santos, pelo cuidado, pelo carinho, pelos ensinamentos e pela humildade, pois sei que em vários momentos, com sua privilegiada inteligência, procurou igualar seu nível de conhecimento ao meu, para me ajudar a entender o trivial. Meu muito, mas muito obrigada!

Agradeço à minha irmã Maria Leny, por ter me mostrado os caminhos da ciência.

A minha mãe, Benilce, que é para mim um exemplo de mulher, de força e de superação. A ela o meu muito obrigada por ter me dado a oportunidade de chegar até aqui!

Ao meu esposo, Paulo Ricardo, que desde sempre acreditou em mim e sempre me encorajou nessa aventura que é cursar o doutorado e desenvolver a tese. Sem ele, este estudo definitivamente não estaria concluído. Pela paciência infinita que tem tido comigo, pela amizade, companheirismo e por todas as vezes em que me motivou, incentivou e pressionou a redigir e concluir esta tese, o meu muito obrigada!

Agradeço aos meus filhos amados, João Arthur e João Gabriel, presentes de Deus, por terem suportado e entendido a ausência em período tão importante da vida deles. Amo-os e sempre serei em suas vidas o abrigo para onde podem voltar!

*“Nas grandes batalhas da vida, o primeiro passo
para a vitória é o desejo de vencer.”*

Mahatma Gandhi

Sumário

	Página
LISTA DE FIGURAS	viii
LISTA DE TABELAS	ix
RESUMO	x
SUMMARY	xii
1 INTRODUÇÃO	1
1.1 Contextualização	1
1.2 Objetivos	3
2 FUNDAMENTAÇÃO TEÓRICA	4
2.1 Análise de Sobrevivência	4
2.1.1 Estimador de Kaplan-Meier	5
2.1.1.1 Teste log-rank	6
2.1.2 Modelo de Cox	7
2.1.3 Modelos paramétricos aplicados a dados de sobrevivência	8
2.1.3.1 Modelo Exponencial	9
2.1.3.2 Modelo Poisson	10
2.1.3.3 Modelo Weibull	11
2.1.3.4 Modelo Log-normal	12
2.1.4 <i>Splines</i> cúbicas	12
2.1.5 <i>Splines</i> Cúbicas Restritas	13
2.1.6 Modelos paramétricos flexíveis	14
2.2 Modelos Paramétricos Flexíveis com incorporação de <i>splines</i>	16
2.3 Função de Verossimilhança	17
2.4 Polinômios Fracionários	18
2.5 Riscos Competitivos	19
2.5.1 Dados de sobrevivência na ausência de riscos competitivos	21
2.5.2 Dados de sobrevivência na presença de riscos competitivos	21
2.5.3 Função de risco por causa específica	22
2.6 Estimador da função de incidência acumulada	24

2.7	Função de risco de subdistribuição	25
2.8	Dieta Mediterrânea	26
3	MATERIAL E MÉTODOS	29
3.1	Descrição das coortes	29
3.1.1	Coorte Multicentrica Italiana Colelitiasi (Micol)	29
3.1.2	Coorte Nutrizione Epatopologia (Nutriep)	30
3.1.3	Elaboração da dieta Padrão	31
3.2	Métodos	34
4	RESULTADOS E DISCUSSÃO	37
5	CONCLUSÃO	50
	REFERÊNCIAS BIBLIOGRÁFICAS	53
	APÊNDICES	58

Lista de Figuras

	Página
1 Pirâmide da dieta Mediterrânea	28
2 Efeito da variação do número de nós na função <i>spline</i> na estimação da função de risco basal	40
3 Efeito da variação do número de nós no log do tempo	41
4 Modelo Poisson para o índice de aderência a dieta mediterrânea (IADM) com <i>spline</i> cúbica restrita para a função de risco basal.	43
5 Modelo de Poisson para o índice de aderência a dieta mediterrânea (IADM) com <i>splines</i> cúbicas para a função de risco basal com 95% IC	44
6 Estimação da função de risco de linha de base para um melhor ajuste FP1, FP2 e FP3	46
7 Comparação das funções com polinômios de grau 3 e 4 (FP3 e FP4) para a taxa de risco basal.	46
8 Função de risco de subdistribuição prevista para cada causa e taxa de risco específico por causa.	48
9 Previsão absoluta das diferenças ($IADM \leq 4 - IADM \geq 4$) nas funções de incidência acumulada específica da causa com 95% IC	49

Lista de Tabelas

	Página
1 Posição dos nós internos em modelos <i>splines</i> com diferentes graus de liberdade	17
2 Estatísticas descritivas para dados demográficos das coortes Micol e Nutriep. Dados relacionados as variáveis estudadas, evento de interesse e falha.	38
3 Estimação do parâmetro de boa aderência a dieta mediterrânea ($IADM \geq 4$) do modelo de Cox com diferentes graus de liberdade no tempo e no log do tempo para identificar os efeitos no aumento do número de nós.	39
4 Estimativas obtidas pelo modelo clássico de Cox para inclusão de covariáveis. Razão de risco, erro padrão, valor de z, valor de p e IC.	42
5 Localização de nós selecionados aleatoriamente para verificação da forma da função de risco estimada.	45
6 Modelo de regressão de Fine e Gray sobre o risco de subdistribuição para cada um dos desfechos: morte por câncer gastrointestinal e morte por outras causas – coortes Micol e Nutriep	47

MODELO PARAMÉTRICO FLEXÍVEL PARA RISCOS COMPETITIVOS: UMA ANÁLISE DE SOBREVIDA EM UMA COORTE ITALIANA

Autora: REINALDA SOUZA OLIVEIRA

Orientadora: Profa. Dra. LICIANA VAZ DE ARRUDA SILVEIRA

Coorientador: Dr. ALBERTO RUBÉN OSELLA

RESUMO

Nas pesquisas clínicas, o método usado para analisar os fatores que contribuem para a mortalidade relacionada aos mais diversos tipos de desfecho é a análise de sobrevivência. Na pesquisa clínica há mais de um resultado possível durante o acompanhamento dos dados de sobrevivência, estes são conhecidos como eventos competitivos. Na ocorrência dos eventos em que a morte é considerada um evento censurado, o modelo de risco proporcional de Cox pode ser empregado para estimar os efeitos das covariáveis sobre o risco. Entretanto, o efeito sobre o risco não pode ser diretamente afetado pela função de incidência acumulada (FIA). Assim, de maneira geral, o presente estudo teve por objetivo fazer uma abordagem dos métodos de análise de sobrevida nos aspectos paramétricos de modelos flexíveis para riscos competitivos, e aplicá-los a uma coorte do Sul da Itália, para avaliar os riscos de óbito para câncer gastrointestinal e outras causas. Trata-se de um estudo observacional retrospectivo, com um período de seguimento de dez anos, realizado em duas coortes, totalizando 2.623 indivíduos selecionados segundo a lista de eleitores dos municípios Castellana

Grotte e Putignano, na Itália. Os modelos flexíveis para riscos competitivos foram usados para avaliar as distribuições acumuladas de incidência específica por causa. Observou-se que os indivíduos com câncer gastrointestinal adeptos da dieta mediterrânea apresentam um risco constante e menor de morrer, especialmente nas etapas inicial e final do seguimento, enquanto esse efeito é menos pronunciado em indivíduos que não aderiram à citada dieta. Assim, foi possível observar que a aderência à dieta mediterrânea ajuda a reduzir em cerca de 20% o risco de morte por câncer gastrointestinal, comparativamente à aderência a outras dietas.

Palavras-chave: Efeitos dependentes do tempo, Modelos de sobrevivência flexíveis, Subdistribuição.

FLEXIBLE PARAMETRIC MODEL FOR COMPETITIVE RISKS: A SURVEY ANALYSIS IN AN ITALIAN COURTE

Author: REINALDA SOUZA OLIVEIRA

Orientadora: Profa. Dra. LICIANA VAZ DE ARRUDA SILVEIRA

Coorientador: Dr. ALBERTO RUBÉN OSELLA

SUMMARY

In clinical research, the method used to analyze the factors that contribute to mortality related to the most diverse types of outcome is the survival analysis. In clinical research there is more than one possible outcome during monitoring of survival data, these are known as competitive events. In the occurrence of events in which death is considered a censored event, the Cox proportional hazard model can be used to estimate the effects of covariates on risk. However, the effect on risk can not be directly affected by the cumulative incidence function (FIA). Thus, in general, the present study aimed to analyze the methods of survival analysis in the parametric aspects of flexible models for competitive risks, and to apply them to a cohort in the South of Italy to evaluate the risk of death for gastrointestinal cancer and other causes. This is a retrospective observational study with a follow-up period of ten years, carried out in two cohorts, totaling 2,623 individuals selected according to the list of voters of the municipalities Castellana Grotte and Putignano, Italy. Flexible models for competitive risks were used to evaluate cumulative distributions

of specific incidence by cause. It was observed that individuals with gastrointestinal cancer adherents of the Mediterranean diet present a constant and lower risk of dying, especially in the initial and final stages of the follow-up, whereas this effect is less pronounced in individuals who did not adhere to the aforementioned diet. Thus, it was possible to observe that adherence to the Mediterranean diet helps to reduce the risk of death from gastrointestinal cancer by about 20%, compared to adherence to other diets.

KeyWords: Time-dependent effects, Flexible survival models, Subdistribution.

1 INTRODUÇÃO

1.1 Contextualização

Nas pesquisas clínicas, o método usado para analisar os fatores que contribuem para a mortalidade relacionada aos mais diversos tipos de desfecho é a análise de sobrevivência, por meio do método de Kaplan-Meier, que estima as funções de sobrevivência, e do modelo de riscos proporcionais de Cox, que estima os efeitos das covariáveis sobre o risco da ocorrência do evento ((Nelson et al., 2007); (Royston & Parmar, 2002) e (Choodari-Oskoei et al., 2012)).

A análise de sobrevivência engloba um conjunto de procedimentos estatísticos, em que a variável dependente é o tempo, mensurado desde um instante inicial bem definido até a ocorrência de um determinado evento de interesse, conhecido como falha ou desfecho (Colosimo & Giolo, 2006); (Carvalho et al., 2011). O período que vai do início do estudo até a ocorrência do evento de interesse é denominado “tempo de vida” ou “tempo de falha”.

A mensuração da estimativa de probabilidade de uma falha acontecer antes de um determinado tempo é muito importante para estudos que envolvem a análise de sobrevivência, e tem sido exaustivamente discutida nas últimas décadas (Collett, 2014).

Na análise de sobrevivência, os dados são medidos a partir de um momento determinado de início, indo até um evento de interesse ou um ponto final escolhido. Como por exemplo, se quisermos determinar a incidência de morte por câncer do trato gastrointestinal entre pacientes com câncer de intestino, cada paciente será seguido de uma data inicial (como data de diagnóstico ou data da cirurgia), até a data de óbito. Um paciente que morre de câncer de intestino durante o período de estudo seria considerado um “evento” na data da morte enquanto um paciente que está vivo no final do estudo seria considerado “censurado”. Assim, cada paciente fornece duas informações: o tempo de acompanhamento e o *status* (evento ou censurado); ou seja, caso o paciente venha a óbito por uma causa diferente daquela em estudo, esse evento não será considerado.

O método de Kaplan-Meier e o modelo de Cox Cox (1972), consideram como variável resposta apenas um tipo de desfecho, gerando, assim, uma limitação, haja vista que um indivíduo pode sofrer um evento diferente daquele escolhido. O que se observa na análise de sobrevivência clássica é que para cada indivíduo, registra-se apenas o tempo até a

ocorrência de um único evento. Contudo, dados de sobrevivência apresentam falhas devidas a várias causas distintas. Uma alternativa a sobrevivência clássica é o uso da análise de sobrevivência com riscos competitivos.

Nos estudos da oncologia clínica, a taxa de sobrevida, ou seja, a proporção de pacientes que se espera que venham a sobreviver ao câncer, varia de acordo com o tipo e estágio da doença, e a mortalidade relacionada ao câncer pode ser um interesse primário (Rashidkhani et al., 2011). Contudo, o estudo de outras causas de morte, relacionadas ou não ao câncer, pode também ser importante no instante em que influencia os resultados da pesquisa.

Zhang (2017) afirma que quando na pesquisa clínica há mais de um resultado possível durante o acompanhamento dos dados de sobrevivência, estes são conhecidos como eventos competitivos; e que na ocorrência dos eventos em que a morte é considerada um evento censurado, o modelo de risco proporcional de Cox pode ser empregado para estimar os efeitos das covariáveis sobre o risco. Entretanto, o efeito sobre o risco não pode ser diretamente afetado pela função de incidência acumulada (FIA).

A FIA, permite quantificar a probabilidade de um indivíduo experimentar um evento pertencente a um conjunto de causas distintas (Geskus, 2015). Neste contexto, outros métodos de modelagem devem ser empregados como alternativa aos modelos clássicos citados.

Dadas as limitações impostas pelo modelo de Cox, tem sido bastante recorrente o uso de modelos flexíveis com *splines* cúbicas para se obter estimativas distintas das funções de risco de linha de base para cada tipo de evento, assim como para possibilitar a estimação dos efeitos das covariáveis dependentes do tempo. Belot et al. (2010), Lambert & Hinchliffe (2013), e Mozumder et al. (2017) são algumas das obras que fazem uso dos modelos flexíveis para estimação de eventos concorrentes.

Royston & Parmar (2002) desenvolveram uma nova família de modelos paramétricos com o objetivo de aumentar a flexibilidade frente aos modelos então existentes, mantendo as vantagens do modelo de Cox. Os autores propuseram a inclusão de nós internos visando à melhoria do processo de estimação das funções de sobrevivência frente a outros modelos comumente empregados, tais como o Modelo Weibull e o log-logístico. Discutiram também o uso de modelos paramétricos flexíveis de riscos proporcionais e probabilidades proporcionais para dados de sobrevivência censurados, com aplicação na modelagem prognóstica e

na estimativa dos efeitos dos tratamentos.

Existem duas abordagens importantes na modelagem dos riscos competitivos: modelam-se os riscos por causa específica, e faz-se uma transformação para obter a FIA ou modela-se a FIA diretamente (Lau et al., 2009).

A primeira abordagem – risco por causa específica – baseia-se no ajuste de um modelo para cada tipo de falha, sendo o modelo de Cox o mais utilizado. Essa abordagem é criticada pelo fato de não mensurar de forma direta o efeito da covariável sobre a FIA.

A segunda abordagem – a função de subdistribuição do risco – possibilita a mensuração direta dos efeitos das covariáveis sobre a FIA. O modelo ajustado nessa abordagem é o desenvolvido por Fine & Gray (1999), um modelo de regressão de sobrevivência baseado na FIA, que descreve a probabilidade de ocorrer um evento antes de um momento específico. A abordagem de Fine e Gray tornou-se um método padrão para análise de riscos competitivos (Austin et al., 2016).

Ainda que, na modelagem de dados de riscos competitivos faça-se uso do modelo de Fine e Gray, o modelo paramétrico flexível desenvolvido por Royston & Parmar (2002), apresenta-se como alternativa no instante em que possibilita a incorporação de splines, tornando assim a estimação das funções de sobrevivência mais robustas.

1.2 Objetivos

O presente estudo tem como objetivo geral fazer uma abordagem dos métodos de análise de sobrevida nos aspectos paramétricos de modelos flexíveis para riscos competitivos.

Com os dados de duas coortes das cidades de Castellana Grotte e Putignano, ambas na Itália, pretende-se alcançar os seguintes objetivos específicos:

- comparar a abordagem baseada na modelagem do risco por causa específica com a abordagem baseada na modelagem do risco de subdistribuição usando os modelos paramétricos flexíveis; e
- aplicar o mesmo modelo flexível proposto em um estudo prognóstico de mortalidade por câncer, com dois eventos concorrentes: morte por câncer gastrointestinal e morte por outras causas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Análise de Sobrevivência

A análise de sobrevivência engloba um conjunto de procedimentos estatísticos, em que a variável dependente é o tempo, mensurado desde um instante inicial bem definido até a ocorrência de um determinado evento de interesse, conhecido como falha ou desfecho (Colosimo & Giolo, 2006) e (Carvalho et al., 2011). O período que vai do início do estudo até a ocorrência do evento de interesse é denominado “tempo de vida”.

Na análise de sobrevivência, o período até o indivíduo experimentar o evento de interesse pode ser longo, muitas vezes não atingindo o tempo total de seguimento, gerando, assim, uma situação de dados faltantes (Noordzij et al., 2013). Essas informações faltantes ou incompletas são chamadas de informações censuradas. Dados censurados representam a presença de informações incompletas e parciais, decorrente da perda de acompanhamento de um ou mais participantes no decorrer do estudo (Kleinbaum & Klein, 2012). A censura é classificada em: censura informativa, censura não informativa, censura tipo I, censura tipo II e censura tipo III (Carvalho et al., 2011).

O tempo de sobrevivência de um indivíduo é representado pela variável aleatória contínua e não negativa T_i e pela variável indicadora de falha ou censura δ_i , sendo $\delta = 1$ caso seja observado o evento, e $\delta = 0$ no caso contrário, ou seja, ocorrência de censura.

A função de sobrevivência que descreve a probabilidade de um indivíduo sobreviver ao instante t é uma função monótona decrescente e contínua, podendo ser expressa por (Colosimo & Giolo, 2006):

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t), \quad (1)$$

em que:

$F(t)$ é a função de distribuição da variável T , para $t > 0$.

Uma função muito importante dentro da análise de sobrevivência é a função de risco. A função de risco é definida pelo limite da probabilidade de um indivíduo falhar num intervalo de tempo $[t; t + \Delta t]$, para Δt tendendo a zero, dado que ele tenha sobrevivido até o tempo t , e é expressa por:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = -\frac{d \log S(t)}{dt}, \quad t > 0. \quad (2)$$

Como a função de risco descreve que a probabilidade instantânea de falha se modifica ao longo do tempo, ela tem sido preferida por muitos autores (Lambert & Royston (2009); Royston & Parmar (2002)), por exemplo, para descrever o comportamento do tempo de sobrevivência.

A função de risco acumulado do tempo de sobrevivência T é expressa por $\Lambda(t)$,

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

sendo, portanto, a função de sobrevivência expressa por:

$$S(t) = \exp\{-\Lambda(t)\}. \quad (3)$$

2.1.1 Estimador de Kaplan-Meier

Colosimo & Giolo (2006) afirmam que, na prática, o conjunto de dados amostrais de tempos de falha apresenta censuras, o que requer técnicas estatísticas especializadas para acomodar a informação contida nessas observações. Assim, é comum utilizar o estimador não paramétrico da função de sobrevivência, conhecido como estimador de Kaplan-Meier. Ele é conhecido também como estimador produto-limite, em que a função de sobrevivência é estimada utilizando-se o produto das probabilidades de sobrevivência até o tempo t .

A abordagem de Kaplan-Meier se baseia na premissa de que a censura é não-informativa, ou seja, independente do evento de interesse. Contudo, na presença de riscos competitivos, os eventos tornam-se censura informativa.

Carvalho et al. (2011) afirmam:

O estimador de Kaplan-Meier utiliza os conceitos de independência de eventos e de probabilidade condicional para desdobrar a condição de sobreviver até o tempo t em uma sequência de elementos independentes que caracterizam a sobrevivência em cada intervalo de tempo anterior a t , cuja probabilidade é condicional aos que estão em risco em cada período.

O estimador de Kaplan-Meier é dado por:

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

considerando-se:

- uma amostra de dimensão n , sendo $t_{(1)}, \dots, t_{(k)}$, $k \leq n$ os tempos distintos até a ocorrência do acontecimento, na amostra;
- d_i o número de ocorrências num determinado $t_{(i)}$, $i = 1, \dots, k$; e
- $n_{(i)}$ o número de indivíduos em risco imediatamente antes de $t_{(i)}$, $i = 1, \dots, k$.

O estimador de Kaplan-Meier considera tantos intervalos quanto o número de acontecimentos distintos que ocorreram, em que a estimativa $\hat{S}(t)$ é uma função em escada com saltos nos instantes de falha observados (Godinho, 2016). A função de sobrevivência estimada $\hat{S}(t)$ tem o formato de escada, porque o risco se mantém constante até a ocorrência do próximo evento.

O gráfico da curva de Kaplan-Meier apresenta a forma como os acontecimentos vêm ocorrendo em diferentes grupos, atingindo o valor zero quando o maior tempo observado não é censurado.

2.1.1.1 Teste log-rank

A verificação de diferenças estatisticamente significativas entre as funções de sobrevivência de dois ou mais grupos é feita utilizando-se testes de hipóteses. Carvalho et al. (2011) e Colosimo & Giolo (2006) ressaltam que apesar de existirem diferentes testes que avaliam as diferenças entre as curvas, o de Mantel-Haenzel, ou *log-rank*, é o mais utilizado em análise de sobrevivência.

O teste *log-rank* compara os valores observados e esperados de cada estrato, sob a hipótese de que o risco é o mesmo em todos os grupos. Ele é considerado o mais robusto quando as funções de risco dos grupos são proporcionais, e, caso não haja proporcionalidade de risco, mesmo assim ele ainda é considerado, desde que as funções de risco não se cruzem.

Ele é apropriado quando a razão das funções de taxa de falha dos grupos a serem comparados é aproximadamente constante, ou seja, as populações têm a propriedade de riscos proporcionais.

A hipótese de riscos proporcionais pode ser avaliada com base na representação gráfica das funções $\ln[-\ln S(t)]$ de cada grupo, que deverão ser aproximadamente paralelas.

2.1.2 Modelo de Cox

O modelo de regressão de Cox é o método estatístico clássico utilizado para modelar dados de sobrevivência. Ele mede a relação entre um conjunto de covariáveis e uma variável-resposta do tipo tempo de vida. O modelo possibilita avaliar o efeito de diferentes tratamentos no tempo de vida dos indivíduos em estudo e o efeito de outras características de base dos indivíduos (Godinho, 2016).

As estimativas produzidas pelo modelo de Cox são razões de risco (RR) que medem o quanto uma covariável de interesse faz aumentar ou diminuir a taxa de ocorrência de um evento em particular, assumindo que ele atua de forma multiplicativa. Uma suposição básica do modelo Cox para estimar os efeitos das covariáveis é a proporcionalidade dos riscos ao longo de todo o período de observação (Kleinbaum & Klein, 2012).

Cox (1972) introduziu uma nova dimensão de flexibilidade na análise de dados censurados com covariáveis. O modelo de regressão de Cox apresenta-se sob a forma de uma função de risco dada pelo produto de dois termos: a função de risco subjacente $h_0(t)$, que é o componente não paramétrico, uma função não negativa do tempo que não é especificada, usualmente chamada de função de risco basal, pois $h(t; \mathbf{x}) = h_0(t)$ quando $x = 0$, e o fator $\exp(\boldsymbol{\beta}'\mathbf{x})$, que representa o efeito multiplicativo das covariáveis na função de risco. Assim, a função de risco no tempo t , dado o vetor $\mathbf{x} = (x_1, \dots, x_p)'$ de covariáveis, é dada por:

$$h(t; \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}), \quad (4)$$

em que:

$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ é o vetor de parâmetros de regressão a ser estimado; e

T é a variável resposta, correspondente ao tempo decorrido até a ocorrência de um evento.

A função de risco basal é uma estimativa muito útil nas ciências biológicas, particularmente na medicina, porque possibilita a obtenção de uma medida clinicamente compreensível da evolução da doença, constituindo a base sobre a qual são estimadas todas as outras medidas.

Por meio da razão de risco, e baseando-se na função de risco, é possível comparar-se indivíduos que sejam diferentes apenas no valor de uma covariável. Considerando-se dois indivíduos com vetor de p covariáveis $\mathbf{x}_1$ e $\mathbf{x}_2$, que diferem apenas no valor da variável x_{kj} , $k = 1, 2$, a razão de risco, conhecida também como risco relativo, é dada por:

$$\frac{h(t; \mathbf{x}_1)}{h(t; \mathbf{x}_2)} = \frac{h_0(t) \exp(\beta_1 x_{11} + \dots + \beta_j x_{1j} + \dots + \beta_p x_{1p})}{h_0(t) \exp(\beta_1 x_{21} + \dots + \beta_j x_{2j} + \dots + \beta_p x_{2p})} = \exp[\beta_j (x_{1j} - x_{2j})].$$

O termo $\exp(\beta_j)$ representa a razão de risco de ocorrência do evento para dois indivíduos que se diferem por uma unidade no valor da covariável x_j .

O modelo de Cox é semiparamétrico. Uma característica importante do modelo é o fato de a razão de risco ($\exp \beta'$) incorporar as covariáveis e não depender do tempo. Como o modelo pressupõe riscos proporcionais, isso significa que o efeito das covariáveis no tempo de vida não sofre influência do tempo (Godinho, 2016).

Apesar de autores como Belot et al. (2010) e Zhang (2017) comentarem sobre as limitações do modelo de Cox, Kleinbaum & Klein (2012) afirmam que, embora seja preferível usar um modelo paramétrico, na maioria das vezes não se tem a certeza quanto ao modelo mais apropriado. E completam, afirmando que, apesar de ser semiparamétrico, o modelo de Cox é bastante robusto, gerando resultados muito próximos do modelo paramétrico correto.

2.1.3 Modelos paramétricos aplicados a dados de sobrevivência

A sobrevivência pode ser influenciada por aspectos internos de inclusão de covariáveis na sua modelagem. Na modelagem de dados de sobrevivência o modelo de Cox é bastante utilizado para descrever dados que envolvem tempo de vida, pois é de fácil interpretação, e, por ser um modelo de regressão semiparamétrico, não é necessário assumir uma distribuição particular para o tempo de vida, característica essa que o torna muito flexível (Godinho, 2016).

O objetivo de um modelo de regressão consiste em estimar o efeito de covariáveis sobre uma variável-resposta de acordo com a distribuição que se supõe que ela possui e conforme a função escolhida que estabelece a relação entre as covariáveis.

Na análise de sobrevivência, aplica-se o mesmo raciocínio, pois o tempo decorrido até

o desfecho de interesse é a variável-resposta, e, por se tratar de um modelo paramétrico, assume-se que esse tempo segue uma família de distribuições com parâmetros desconhecidos (Carvalho et al., 2011).

Quando há a possibilidade de uma distribuição adequada do tempo, é preferível utilizar os modelos paramétricos, haja vista que são mais eficientes do que o modelo de Cox (Royston & Lambert, 2011). Outra vantagem é que modelos paramétricos possibilitam a inclusão de covariáveis cujo efeito não é proporcional em termos de risco, mas pode ser proporcional em outra escala (como, por exemplo, em termos de possibilidade) (Godinho, 2016). Isso, considerando-se, também, que os modelos de sobrevivência paramétricos geralmente fornecem estimativas suavizadas das funções de risco e sobrevivência para qualquer combinação de valores das covariáveis.

2.1.3.1 Modelo Exponencial

O modelo exponencial é o modelo de sobrevivência paramétrico mais simples, e se distingue por ser o único modelo paramétrico com função de risco constante, ou seja:

$$h(t) = \lambda, \text{ para } \lambda > 0. \quad (5)$$

Para se verificar o ajuste da função aos dados, pode-se fazer uma análise gráfica, traçando as funções de risco acumulado empíricas e ajustando-as em relação a t :

$$H(t) = \int_0^t \lambda du = \lambda t. \quad (6)$$

Uma aproximação empírica utilizada para $H(t)$ é o estimador de Nelson-Aalen, dado por:

$$H_{NA}(t) = \sum_{j=1}^k \frac{d_j}{n_j},$$

em que, segundo Collett (2014):

$t_{(k)} \leq t < t_{(k+1)}$, quando $k = 1, \dots, r$, e sendo $t_{(1)}, \dots, t_{(r)}$ os r tempos ordenados, com

$t(r+1) = \infty$; e

n_j e d_j são o número em risco e o número de mortes no tempo $t_{(j)}$.

A função de sobrevivência do modelo exponencial é dada por:

$$S(t) = \exp(-\lambda t).$$

O parâmetro λ representa a velocidade de queda da função de sobrevivência ao longo do tempo, sendo λ um valor positivo. Como a função de risco na distribuição exponencial é constante, o risco acumulado é uma função linear no tempo, sendo representada por:

$$\Lambda(t) = -\ln S(t) = \lambda t.$$

2.1.3.2 Modelo Poisson

Dados de sobrevivência podem ser modelados dentro da estrutura de modelagem linear generalizada, usando modelos Poisson, em que a escala de tempo é dividida em várias, possibilitando, assim, alguma flexibilidade na forma da função de risco basal. O modelo de Cox é um caso especial desse tipo de modelo (Royston & Lambert, 2011).

O modelo de Poisson tem muitas aplicações, não apenas na análise de contagem de eventos, mas também no contexto de modelos para tabelas de contingência e análise de dados de sobrevivência.

Uma variável aleatória Y tem uma distribuição de Poisson com parâmetro μ , se tomar valores inteiros $y = 0, 1, 2, \dots$ com probabilidade:

$$Pr(Y = y) = \frac{e^{-\mu} \mu^y}{y!}, \text{ para } \mu > 0.$$

A média e a variância podem ser representadas por:

$$E(Y) = Var(Y) = \mu.$$

Como a média é igual à variância, qualquer fator que afeta um, também afeta o outro. Assim, a suposição usual de homocedasticidade não é apropriada para dados de Poisson (Royston & Lambert, 2011).

Supondo que a taxa de risco $h(t)$ é constante ao longo de algum intervalo de tempo, isto é, $h(t) = \lambda$, uma taxa de risco constante implica que os tempos de sobrevivência têm uma distribuição exponencial, e, portanto, a função de sobrevivência é representada por:

$$S(t) = \exp(-\lambda t). \quad (7)$$

A contribuição do i -ésimo indivíduo para o logaritmo da função de verossimilhança para um modelo de sobrevivência paramétrica é dada por:

$$\ln L_i = d_i \ln h(t_i) + \ln S(t_i),$$

em que:

d_i é o evento indicador (0 para a observação censurada, 1 para o evento de interesse); e t_i é o evento observado ou tempo de censura.

Ao substituir-se as funções de risco e sobrevivência pela distribuição exponencial na fórmula acima, obtém-se:

$$\ln L_i = d_i \ln \lambda - \lambda t_i. \quad (8)$$

Ao eliminar termos que não dependem dos parâmetros, a expressão é idêntica à probabilidade de uma distribuição de Poisson com resultado d_i e média t_i .

2.1.3.3 Modelo Weibull

A distribuição Weibull é uma generalização da exponencial. O modelo de regressão Weibull é bastante aplicado em análise de sobrevivência (Carvalho et al., 2011). Ele possibilita avaliar a variação do risco no tempo de vida.

Apesar de ser um modelo de riscos proporcionais e um modelo de tempo de vida acelerado (Godinho, 2016), ele é muito criticado pela falta de flexibilidade na forma da função de risco basal, que é monótona crescente ou decrescente. Considere-se uma distribuição de Weibull para T com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\gamma > 0$. As funções de sobrevivência, de risco e densidade de probabilidade são dadas respectivamente por:

$$S(t) = \exp(-\lambda t^\gamma) \ln L_i = d_i \ln \lambda - \lambda t_i; \quad (9)$$

$$h(t) = \lambda \gamma (t^{\gamma-1}) \ln L_i = d_i \ln \lambda - \lambda t_i; \quad (10)$$

$$f(t) = \lambda \gamma (t^{\gamma-1}) \exp(-\lambda t^\gamma) \ln L_i = d_i \ln \lambda - \lambda t_i. \quad (11)$$

2.1.3.4 Modelo Log-normal

O uso da distribuição log-normal na modelagem de dados de sobrevivência significa que a variável tempo de vida T segue uma distribuição log-normal, com parâmetros λ e γ , enquanto o $\ln T$ segue uma distribuição normal, com parâmetros λ e γ^2 (Carvalho et al., 2011). As funções densidade de probabilidade, sobrevivência e risco da variável T são dadas por:

$$\begin{aligned} f(t) &= \frac{1}{\gamma\sqrt{2\pi}} t^{-1} \exp \left[-\frac{1}{2} \left(\frac{\ln(t) - \lambda}{\gamma} \right)^2 \right], \\ S(t) &= 1 - \Phi \left(\frac{\ln(t) - \lambda}{\gamma} \right), \quad e \\ h(t) &= \frac{\Phi \left(\frac{\ln(t) - \lambda}{\gamma} \right)}{t\gamma(1 - \Phi(\gamma))}. \end{aligned}$$

em que:

$t > 0$ e;

$\Phi(\cdot)$ é a função de distribuição da normal-padrão.

Pode-se perceber que a distribuição log-normal é um caso particular da distribuição gama generalizada, dada pela função de densidade:

$$f(x) = \frac{kx^{k\alpha-1}}{\beta^{k\alpha}\Gamma(\alpha)} \exp \left[-(x/\beta)^k \right],$$

em que:

$f(x)$ = frequência na classe x ;

k e α são parâmetros de forma ($k > 0, \alpha > 0$);

β é o parâmetro de escala ($\beta > 0$) ; e

Γ é a função gama.

2.1.4 Splines cúbicas

Quando o objetivo é trabalhar com o ajuste e a interpolação de funções, o uso de aproximação por polinômios é muito adequado, pois eles possuem várias propriedades, como a de função analítica (pode ser localmente expandida em séries de Taylor), que torna possível calcular as derivadas de qualquer ordem (Alves et al., 2005).

Quando se faz o uso de muitas derivadas, a função que está sendo aproximada torna-se muito restritiva. Para resolver esse problema, faz-se uso de polinômios por parte, possibilitando-se, assim, a descontinuidade das derivadas de ordem mais elevada em alguns pontos. A função por parte recebe o nome de função *spline*.

O termo *spline* se refere a um tipo particular de função de suavização, definida como polinômios segmentados (Carvalho et al., 2011).

Seja $s(x)$ uma função *spline* não linear de ordem N com covariáveis x , com C nós, em que: $c_1 < \dots < c_k$.

Uma função *spline* sem restrição de continuidade pode ser escrita como:

$$s(x) = \sum_{j=0}^n \beta_{0j} x^j + \sum_{i=1}^C \sum_{j=0}^N \beta_{ij} (x - c_i)_+^j.$$

A notação “+” é definida por:

$$u_+ = \begin{cases} u, & \text{se } u > 0 \\ 0, & \text{se } u \leq 0 \end{cases}$$

2.1.5 Splines Cúbicas Restritas

As *splines* são funções matemáticas flexíveis, definidas por polinômios por parte, com algumas restrições para garantir que a curva geral seja suave. As *splines* cúbicas são as mais utilizadas na prática. Lambert & Royston (2009) afirmam que as *splines* de regressão são úteis porque podem ser incorporadas em qualquer modelo de regressão com um preditor linear.

Uma *spline* cúbica natural ou restrita é definida por funções polinomiais cúbicas, sendo restrita à linearidade para além dos pontos que a delimitam, isto é, para além dos nós-limite (c_{min} e c_{max}) (Godinho, 2016). Os nós-limite – pontos em que os polinômios se juntam – são determinados pelo primeiro e pelo último tempo de vida observado. Além desses nós, existem m nós internos ($c_1, \dots, c_m$ com $c_1 > c_{min}$ e $c_m < c_{max}$), cujo número e localização pode ser definido.

A complexidade da curva depende do número de nós e do número de graus de liberdade ($df = m + 1$, sendo m o número de nós internos). Os citados autores indicam o uso de modelos com $m < 4$, para que as curvas com mais nós não se tornem instáveis. Assim, o número de parâmetros em um modelo de regressão *spline* padrão é $C + 4$.

Uma *spline* cúbica natural pode ser representada por:

$$s(x) = y_0 + y_1 z_1 + y_2 z_2 + \dots + y_{C-1} z_{C-1}, \quad (12)$$

em que:

$y_0, \dots, y_{C-1}$, são parâmetros a serem estimados.

As variáveis derivadas z_j são calculadas como:

$$z_1 = x;$$

$$z_j = (x - c_j)_+^3 - \lambda_j (x - c_1)_+^3 - (1 - \lambda_j) (x - c_K)_+^3, \quad j = 2, \dots, C - 1 \quad e \quad (13)$$

em que:

$$\lambda_j = \frac{c_{max} - c_j}{c_{max} - c_{min}} \quad e \quad (x - a)_+ = \max(0, x - a).$$

Segundo afirmam Royston & Parmar (2002), desde que existam nós suficientes, o aumento do número de nós e a sua localização não são particularmente importantes, haja vista que não alteram a curva ajustada. Para se ajustar uma *spline* cúbica natural para uma variável z , faz-se necessária a criação de novas variáveis, que são transformações de z , obtidas a partir dos nós internos.

2.1.6 Modelos paramétricos flexíveis

O uso de modelos paramétricos apresenta mais vantagens, se comparado ao modelo de Cox, porém existem modelos paramétricos simples, como por exemplo a distribuição normal, que não são tão flexíveis na representação adequada da função de risco quando associada a um grupo de indivíduos Godinho (2016) e ?, o que torna o modelo de Cox um modelo mais usual na análise de sobrevivência, pois como a função de risco subjacente não necessita de definição, ela não influencia as estimativas dos riscos relativos.

Os modelos paramétricos flexíveis são uma extensão dos modelos Weibull, log-logístico e log-normal. Para os modelos citados, assume-se que o efeito das covariáveis

é proporcional na escala apropriada, ou seja, no risco, na probabilidade de falha ou no *probit* da probabilidade de falha, respectivamente.

A suposição de linearidade nos modelos paramétricos flexíveis implica a transformação da função de sobrevivência e do logaritmo do tempo de sobrevivência. Lambert & Royston (2009) afirmam que a ideia básica da abordagem paramétrica flexível consiste em relaxar a suposição de linearidade do logaritmo do tempo de seguimento, usando *splines* cúbicas restritas.

A abordagem seguida em Royston & Parmar (2002), na obtenção de modelos mais flexíveis, consiste em modelar uma transformação da função de sobrevivência $S(t)$ por uma função de ligação $g(\cdot)$, apresentada na Equação 14.

$$g[S(t; \mathbf{x})] = g[S_0(t)] + \boldsymbol{\beta}' \mathbf{x} \quad (14)$$

o que consiste na modelagem da transformação de $S(t)$ como uma função *spline*, em vez de uma função linear do logaritmo do tempo,

em que:

$S_0(t) = S(t; \mathbf{0})$ é a função de sobrevivência basal; e

$\boldsymbol{\beta}$ é um vetor de parâmetros a serem estimados para as covariáveis $\mathbf{x}$.

Ao implementar o alisamento de $g[S_0(t)]$ na escala do logaritmo do tempo, a função ajustada se torna tipicamente curvada ou quase linear e muito suave (Royston et al., 2006). A suavidade tende a reduzir a chance de ruídos na estimativa de risco. A estimativa de $g[S_0(t)]$ deve ser monótona em t , enquanto a *spline* cúbica natural, que é restrita a ser linear além de algumas observações extremas, não é globalmente monótona.

Segundo (Royston & Parmar, 2002, p.1), “a restrição de linearidade impõe a monotonicidade nas regiões da cauda onde os dados observados são escassos, ao passo que em regiões onde os dados são densos (e desde que o tamanho da amostra não seja pequeno), a monotonicidade é imposta pelos próprios dados”.

Assim, o uso de *splines* cúbicas garante a monotonicidade em conjuntos de dados com tamanho razoável, como, por exemplo, um conjunto com aproximadamente 50 observações não censuradas.

2.2 Modelos Paramétricos Flexíveis com incorporação de *splines*

Royston & Parmar (2002) propuseram um modelo paramétrico flexível usando *splines* cúbicas restritas, para a modelagem de dados de sobrevivência censurados, possibilitando, desse modo, que o risco basal fosse modelado diretamente, introduzindo, assim, maior flexibilidade na forma da distribuição de sobrevivência.

Nelson et al. (2007) afirmam que as *splines* cúbicas restritas oferecem maior flexibilidade na forma da função de risco, quando comparadas com modelos paramétricos padrão, e que modelos com *splines* podem ser facilmente estendidos para efeitos dependentes do tempo. Afirmam, também, que a principal vantagem do método é que as funções de sobrevivência e risco são obtidas analiticamente, o que acelera o tempo computacional, se comparado com métodos que adotam integração numérica.

Para obter um modelo flexível de risco proporcional, Royston & Parmar (2002) modelaram o logaritmo da função de risco acumulado subjacente, como uma função *spline* cúbica do logaritmo do tempo:

$$\ln[H(t|\mathbf{x}_i)] = \ln[H_0(t)] + \mathbf{x}_i\beta, \quad (15)$$

em que,

$\mathbf{x}_i$ são covariáveis; e

β é um vetor de parâmetros.

Uma função *spline* cúbica restrita de $\ln(t)$, com nós $c_0 = (c_1, \dots, c_c)$, pode ser escrita como $s(\ln(t)|\gamma, c_0)$, com localização de nó c_0 e covariáveis x_i . Isso é então usado para o logaritmo do risco basal acumulado no modelo de risco acumulado proporcional (Lambert & Royston, 2009):

$$\ln[H(t|\mathbf{x}_i)] = s[\ln(t)|\gamma, \mathbf{c}_0] + \mathbf{x}_i\beta. \quad (16)$$

Transformando na escala de sobrevivência e risco, tem-se:

$$S(t|X_i) = \exp\{-\exp(\eta_i)\},$$

$$h(t|X_i) = \frac{ds[\ln(t)|\gamma, c_0]}{dt} \exp(\eta_i).$$

A função de risco envolve as derivadas das funções splines cúbicas restritas. Assim:

$$s'(x) = \gamma_1 z'_1 + \gamma_2 z'_2 + \dots + \gamma_{K-1} z'_{K-1}$$

em que:

$$z'_1 = 1 \text{ e}$$

$$z'_j = 3(x - k_j)_+^2 - 3\lambda_j(x - c_{c1})_+^2 - 3(1 - \lambda_j)(x - c_{cc})_+^2.$$

As posições dos nós internos $c_1, \dots, c_k$, obtidos com base nos percentis empíricos da distribuição do logaritmo dos tempos observados, são apresentadas na Tabela 1.

Tabela 1: Posição dos nós internos em modelos *splines* com diferentes graus de liberdade

gl	Percentis			
2	50			
3	33	67		
4	25	50	75	
...				
9	10	20	...	90

Fonte: Royston & Lambert (2011).

2.3 Função de Verossimilhança

A função de verossimilhança é muito usada na modelagem estatística. Através da função de verossimilhança se avalia o quanto os dados apoiam, concordam ou suportam cada valor possível do parâmetro a ser estimado (Carvalho et al., 2011). Baseando-se nos resultados obtidos pela amostra, pode-se determinar através da verossimilhança qual é a distribuição dentre todas que foram definidas pelos valores de seus parâmetros, com maior possibilidade de ter gerado a amostra.

O logaritmo da função de verossimilhança para o *i-ésimo* indivíduo para um modelo paramétrico flexível na escala de risco acumulado pode ser escrito como (Lambert & Royston, 2009):

$$\ln L_i = d_i[\ln(s'[\ln(t_i)|\gamma, \mathbf{k}_0]) + \eta_i] - \exp(\eta_i), \text{ para } i = 1, \dots, n,$$

em que:

d_i é o evento indicador, sendo $d_i = 1$ para o evento morte; e

$d_i = 0$, no caso contrário.

Para se maximizar a verossimilhança, deve-se definir uma equação adicional para as derivadas da função *spline* e restringir os parâmetros para serem iguais às funções *spline* equivalentes no preditor linear.

2.4 Polinômios Fracionários

Os polinômios fracionários são empregados como alternativa na modelagem da função de risco basal, e constituem uma extensão dos polinômios-padrão. A ideia básica de um polinômio fracionário consiste em possibilitar potências negativas e não inteiras, como também lineares, quadradas e cúbicas. Historicamente, têm sido adaptados polinômios fracionários quadrados, ou têm sido aplicadas transformações simples como o $\log(x)$ ou a $\sqrt{x}$.

Royston & Sauerbrei (2008) apresentam o resultado de um estudo sobre gordura corporal e o uso de estatísticas de qualidade do ajuste para cinco modelos (linear, quadrático, cúbico, polinômio fracionário de potência 1 (FP1) e polinômio fracionário de potência 1 (FP2)).

Nesse estudo, eles mostram que em termos de variação explicada R^2 , há pouca diferença quando se usa um modelo quadrático, cúbico, FP1 ou FP2, enquanto para o modelo linear o valor é bem inferior. No tocante ao desvio $D = -2l$ (onde l é a máxima verossimilhança logarítmica), os modelos cúbico, FP1 e FP2 são superiores ao quadrático, mas diferem pouco entre si. Como, por definição, o modelo FP1 é monotônico, então, desde que o ajuste seja adequado, ele representa bem variáveis do tipo tempo de vida.

Nos polinômios fracionários, é definido um conjunto de potências $S = \{-2, -1, -0, 5, 0, 0, 5, 1, 23\}$, onde uma ou mais dessas transformações nas covariáveis são escolhidas para inclusão no preditor linear (Royston & Sauerbrei, 2008).

Os polinômios de primeira ordem incluem um termo, p_1 , tomado da série S . Polinômios de primeira ordem (FP1) incluem polinômios-padrão da seguinte forma:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

A flexibilidade adicional é obtida pela adição de outras potências. Por exemplo:

$$Y = \beta_0 + \beta_1 x^{-2} + \beta_2 x^{0,5}$$

Um aspecto importante é a inclusão de potências repetidas, possibilitando $p_1 = p_2 = p$. Desse modo, obtém-se um modelo FP2:

$$Y = \beta_0 + \beta_1 x^p + \beta_2 x^p \ln x$$

Um modelo FPM inclui até m potências de $\mathbf{S}$, algumas das quais podem ser repetidas. Por exemplo, a $FP3(-1, 2, 2)$ envolve duas potências de $\mathbf{S}(-1 \text{ e } 2)$, uma das quais é repetida:

$$Y = \beta_0 + \beta_1 x^{-1} + \beta_2 x^2 + \beta_2 x^2 \ln x$$

2.5 Riscos Competitivos

Na área de saúde, é comum a ocorrência de eventos competitivos, especialmente em estudos referentes a oncologia e medicina cardiovascular (Noordzij et al., 2013). Um risco competitivo é um evento cuja ocorrência impede a ocorrência do evento primário (Belot et al., 2010).

Diversos autores abordaram a aplicação da análise de sobrevivência e a presença de riscos competitivos na área de saúde (Wolbers et al. (2014); Noordzij et al. (2013); Coviello & Bogges (2004); Belot et al. (2010)). Austin et al. (2016) estudaram um modelo de predição de mortalidade com a presença de riscos competitivos para avaliar o efeito do relato público do desempenho hospitalar na qualidade do atendimento prestado a pacientes com doença cardiovascular em Ontário, Canadá. Obtiveram dados clínicos detalhados por meio de revisão retrospectiva de prontuários de pacientes hospitalizados com insuficiência cardíaca (IC) entre 1º de abril de 1999 e 31 de março de 2001 (fase 1) e entre 1º de abril de 2004 e 31 de março de 2005 (fase 2) em 103 hospitais em Ontário, Canadá. Zhang (2017) usaram a teoria de riscos competitivos para analisar o efeito dos fatores de risco

para mortalidade a partir do câncer de mama e para a mortalidade decorrente da presença de outras causas.

Na área de saúde, aplica-se a análise de sobrevivência para avaliação das taxas de letalidade dos diversos tipos de câncer. As abordagens mais utilizadas, já citadas anteriormente, são o método de Kaplan-Meier e o modelo de Cox (?; Belot et al. (2010); Jeong & Fine (2006); Zhang (2017)).

O método de Kaplan-Meier é usado para descrever o tempo de sobrevivência da amostra, enquanto o modelo de Cox ajuda a explorar o relacionamento das covariáveis com a sobrevivência, por meio da estimação dos seus efeitos sobre a variável-resposta. Esses dois métodos são limitados a partir do momento em que consideram que a morte por qualquer tipo de câncer está relacionada a um único evento, ou seja, eventos independentes. Gooley et al. (1999), apresentaram em seu estudo a inapropriedade do uso do complemento do estimador de Kaplan-Meier para a probabilidade de falha quando associado a riscos competitivos.

Na realidade, o que se verifica na área de saúde é a existência de eventos múltiplos, ou seja, eventos diferentes decorrentes de um mesmo fator de risco, como, por exemplo, um paciente vir a óbito por um tipo de câncer diferente daquele diagnosticado no início, situação desprezada pelos métodos supracitados (Andersen et al., 2012).

A metodologia de riscos competitivos avalia a incidência, a letalidade e os fatores de risco associados ao óbito de pacientes com câncer, sendo, portanto, uma alternativa ao método de Kaplan-Meier e ao modelo de Cox.

Os eventos competitivos podem ser assim classificados: (1) sobrevivência até o primeiro desfecho – utilizada quando se estudam vários possíveis efeitos de um fator de risco; (2) risco por causa específica – utilizado quando se deseja estimar o efeito de uma variável explicativa sobre um desfecho específico, considerando-se os demais desfechos como censura não informativa; e (3) modelagem de subdistribuição do risco – baseada na função de incidência acumulada dos riscos (subdistribuição), e que leva em conta a totalidade dos desfechos, e não apenas o desfecho de interesse.

Na presença de riscos competitivos, o risco por causa específica fornece a taxa de mortalidade por causa específica, enquanto a função de incidência acumulada (FIA) fornece a proporção de pacientes em qualquer momento que tenham vindo a óbito por causas

específicas.

2.5.1 Dados de sobrevivência na ausência de riscos competitivos

Na ausência de riscos competitivos, a função de sobrevivência descreve a distribuição dos tempos do evento, dada por:

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t) \text{ ou } S(t) = \exp(-H(t)).$$

O complemento da função de sobrevivência, $F(t) = 1 - S(t)$, descreve a incidência do evento ao longo do período de seguimento.

Na ausência de riscos competitivos, existe uma correspondência de “um para um” entre $S(t)$ e a função de risco $H(t)$ (Zhang, 2017). A FIA pode ser escrita como:

$$F(t) = 1 - \exp(-H(t))$$

2.5.2 Dados de sobrevivência na presença de riscos competitivos

Na presença de riscos competitivos, a FIA não pode ser diretamente ligada à função de risco. Assim, não existe uma correspondência “um para um” entre a FIA e a função de risco por causa específica. A incidência acumulada derivada do estimador de Kaplan-Meier é sempre maior do que aquela obtida pela contagem dos riscos competitivos.

Segundo Zhang (2017):

Na estimativa de Kaplan-Meier, um indivíduo é removido do conjunto de risco quando ele experimenta um evento concorrente. Dentro da estrutura de risco competitivo, o indivíduo é um evento no cálculo da probabilidade global de sobrevivência. Portanto, a sobrevivência global de qualquer evento é menor quando considerados os riscos concorrentes.

Ao se usar a função de distribuição acumulada de um evento na presença de riscos competitivos pelo método de Kaplan-Meier ou pelo modelo de Cox, corre-se o risco de superestimar a função, e os efeitos das covariáveis podem não ser adequadamente estimados (Noordzij et al., 2013). Assim, a abordagem que propõe a utilização da FIA ou subdistribuição, que resolve esses problemas, é o método adequado usualmente empregado na presença de riscos competitivos (Mozumder et al., 2017).

Existem duas abordagens importantes na modelagem dos riscos competitivos: (1) modelam-se os riscos por causa específica, e faz-se uma transformação para obter a FIA; e (2) modela-se a FIA diretamente (Lau et al., 2009).

A primeira abordagem – risco por causa específica – baseia-se no ajuste de um modelo para cada tipo de falha, sendo o modelo de Cox o mais utilizado. Essa abordagem é criticada pelo fato de não mensurar de forma direta o efeito da covariável sobre a FIA.

A segunda abordagem – a função de subdistribuição do risco – possibilita a mensuração direta dos efeitos das covariáveis sobre a FIA. O modelo ajustado nessa abordagem é aquele desenvolvido por Fine & Gray (1999), um modelo de regressão de sobrevivência baseado na FIA, que descreve a probabilidade de ocorrer um evento antes de um momento específico.

De acordo com as duas abordagens, risco por causa específica e função de distribuição, duas formas de análise com a presença de eventos competitivos são possíveis.

2.5.3 Função de risco por causa específica

Os dados observados para cada indivíduo no modelo de riscos competitivos são representados por um par de variáveis aleatórias (T, K) , em que K é uma variável aleatória que assume valor zero se a observação é censurada. No caso contrário, $K = k$, sendo k o tipo da primeira causa de falha observada ($K = 1, 2, 3, \dots, k$), e T o período decorrido até a ocorrência da falha por causa k . No caso contrário, T será o período decorrido até a censura.

Na modelagem de riscos específicos, a função de risco por causa específica é definida como a representação da taxa instantânea da ocorrência da k -ésima falha no tempo t , condicionada à sobrevivência até o tempo t , na presença de todas as outras causas de falhas possíveis Lambert & Hinchliffe (2013), sendo matematicamente representada por.

$$h_k(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t, K = k | T \geq t)}{\Delta t} \quad (17)$$

em que:

$k = k$ -ésima falha.

A taxa de risco total é dada pela soma dos riscos específicos por causa, isto é

$$\lambda(t) = \sum_{K=1}^k \lambda_K(t) \quad (18)$$

Sobre a suposição de riscos proporcionais, a taxa de risco por causa específica para a causa k para um paciente com covariáveis x_k é calculada por.

$$h_k(t|x) = h_{k,0}(t) \exp(\beta_k x_k) \quad (19)$$

em que:

$h_{k,0}(t)$ é o risco por causa específica basal para a causa k ; e

β_k são os efeitos das covariáveis (logaritmos das taxas de risco).

A função de risco por causa específica pode ser transformada para que se obtenha uma função de sobrevivência $S_k(t)$, que pode ser expressa por:

$$S_k(t) = \exp[-\Lambda_k] \quad (20)$$

em que:

$\Lambda = \int_0^t h_k(u) du$ é a função de risco acumulado da causa específica.

A função de sobrevivência $S_k(t)$, estimada pela Equação 20, pode ser interpretada como a probabilidade de o indivíduo morrer da causa k , sendo que ele não pode morrer de nenhuma outra causa. Essa situação não é possível nos eventos que envolvem outras causas de morte; ou seja, essa interpretação só é possível em uma situação em que as outras causas de falha não estão presentes, e as distribuições dos tempos para os eventos competitivos são independentes.

Derivada da função de risco por causa específica, a FIA fornece uma estimativa da probabilidade marginal de um evento na presença de eventos competitivos, e não exige a suposição de que os riscos competitivos sejam independentes.

Se os dados não contiverem indivíduos censurados, a FIA no tempo t pode ser estimada como a proporção do número de sujeitos que experimentaram esse tipo de evento – evento competitivo – até o tempo t dividido pelo número total de sujeitos no conjunto de dados. À medida que o tempo aumenta, a FIA aumenta de zero para a proporção total de eventos desse tipo de dados.

A FIA para a causa k é definida como a probabilidade de ocorrência de falha por causa k num tempo menor ou igual a t , e pode ser expressa matematicamente por (Kalbfleisch & Prentice, 2002).

$$F_k(t) = P(T \leq t, K = k), k = 1, \dots, K. \quad (21)$$

Se considerarmos os riscos para todas as causas de falha, a FIA pode ser expressa por:

$$F_k(t) = \int_0^t h_k(u) S(u) du = \int_0^t h_k(u) \exp \left[- \sum_1^k \Lambda_k(u) \right] du. \quad (22)$$

A função de distribuição global, que representa a soma de todas as FIAs para todas as causas de falha, é expressa por:

$$F(t) = P(T \leq t) \sum_{k=1}^K P(T \leq t, K = k) = \sum_{k=1}^K F_k(t).$$

A FIA incorpora os riscos específicos por causa para os eventos competitivos, não sendo, portanto, apenas uma função de risco específico por causa para um evento de interesse.

2.6 Estimador da função de incidência acumulada

Na presença de riscos competitivos, a função de sobrevivência não deve ser utilizada para descrever a ocorrência de eventos, pois estes são descritos pela FIA.

Para se estimar a FIA, considerem-se: (1) m tempos distintos ordenados de ocorrência de qualquer um dos eventos $0 < t_1 < t_2 < \dots < t_m$; (2) uma amostra com n indivíduos; (3) um conjunto de indivíduos em risco $R(t_j)$; (4) d_{ij} o número de eventos do tipo i que ocorreram no tempo t_j ; e (5) n_j o número de indivíduos sob risco em t_j , em que $i = 1, \dots, k$ e $j = 1, \dots, m$. Carvalho et al. (2011) afirmam que a FIA pode ser obtida pela soma de todos os t_j das probabilidades de se observar um evento k entre os indivíduos que não experimentaram qualquer dos eventos possíveis.

Sendo t_{j-1} a probabilidade de estar livre de qualquer evento antes do tempo t_j , e experimentar um evento do tipo k , no tempo t_j , a probabilidade acumulada é dada por:

$$\hat{F}_k(t) = \sum_{\forall t, t_j \leq t} \frac{d_{kj}}{n_j} \hat{S}(t_{j-1}) \quad (23)$$

em que:

k representa a causa da falha;

d_{kj} é o número de indivíduos que experimentaram o evento k no tempo t_j ;

n_j é o número de indivíduos em risco no tempo t_j ; e

$\hat{S}(t)$ é a probabilidade global de sobrevivência, ou seja, o estimador de Kaplan-Meier.

2.7 Função de risco de subdistribuição

A Teoria de Riscos Competitivos propõe várias formas de análise de problemas para eventos desse tipo, sendo as mais comuns: sobrevivência até o primeiro entre os eventos; risco específico da causa; e modelagem da subdistribuição dos riscos (Carvalho et al., 2011). Como o modelo clássico de riscos proporcionais de Cox não possibilita estimar adequadamente o efeito de covariáveis quando os eventos são competitivos, Fine & Gray (1999) apresentaram uma forma alternativa de estimar o efeito das covariáveis sem supor independência entre os tempos dos eventos, modelando diretamente as FIAs. Por meio da relação entre o risco e a sobrevivência, eles definiram uma função de risco de subdistribuição, que pode ser expressa pela Equação 24.

$$\begin{aligned} \bar{\lambda}_k(t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P[t \leq T < t + \Delta t, K = k | T \geq t \cup (T \leq t \cap K \neq k)] \\ &= - \frac{d}{\Delta t} \log[1 - F_k(t)], \end{aligned} \quad (24)$$

interpretada como a taxa instantânea de falha no tempo t de causa k entre os que ainda estão vivos ou aqueles que morreram por qualquer das outras causas concorrentes de $K - 1$, exceto a causa k (Mozumder et al., 2017).

Usando-se a suposição de riscos proporcionais sobre o risco de subdistribuição, a função pode ser assim expressa (Fine & Gray, 1999):

$$\bar{\lambda}_k(t|\mathbf{x}) = \bar{\lambda}_{k,0}(t) \exp(\beta_k^T \mathbf{x}). \quad (25)$$

Para o risco por causa específica $\lambda_k(t|x)$, o conjunto de risco diminui a cada instante

em que ocorrer a falha por outra causa ($\neq k$). Para a função de subdistribuição sob suposição de proporcionalidade, a verossimilhança é dada por:

$$L(\beta_k) = \prod_{k=1}^K \prod_{i=1}^{d_j} \frac{\exp(\beta_k^T x_{(k)i})}{\sum_{l \in R(t_{j(i)})} w_{il} \exp(\beta_k^T x_l)}. \quad (26)$$

O conjunto de risco $R(t_{j(i)})$ é composto por todos os indivíduos que não sofreram qualquer evento em t e todos os que falharam por outros eventos competitivos em t . Como os indivíduos contribuem de formas diferentes para o risco, atribui-se um peso $W_i l$ a cada indivíduo l que sofreu um evento competitivo, e que muda a cada ocorrência do evento de interesse k no tempo t_j . Esse peso é assim definido por Carvalho et al. (2011):

$$w_{l(t_j)} = \begin{cases} 1, & \text{se } l \text{ não tiver sofrido evento ou censura;} \\ \frac{\hat{G}_{km}(t_j)}{\hat{G}_{km}(t_i)}, & \text{se } l \text{ sofreu evento competitivo em } t_i < t_j; \\ 0, & \text{quando o indivíduo } l \text{ é retirado do estudo,} \end{cases}$$

em que:

G é a estimativa de Kaplan-Meier da função de sobrevivência da distribuição de censura; ou seja, no cálculo de $G(\cdot)$, censura se torna evento, e evento (qualquer que seja) se torna censura.

A FIA fornece uma probabilidade marginal, que não utiliza a formulação de Kaplan-Meier, e rejeita a suposição de que os riscos competitivos sejam independentes. Carvalho et al. (2011) e Kleinbaum & Klein (2012) afirmam que, apesar dessas vantagens, a FIA requer que o risco global seja igual à soma de todos os possíveis riscos, condição esta só satisfeita em situações nas quais os riscos competitivos sejam mutuamente excludentes e os eventos não sejam concorrentes.

2.8 Dieta Mediterrânea

A dieta mediterrânea foi a base dos hábitos alimentares ao longo do século XX em todos os países da região do mediterrâneo, originalmente com base em modelos mediterrânicos agrícolas e rurais. No entanto, a prática da dieta mediterrânea tradicional agora está sendo reduzida devido à economia ocidental, à globalização da produção, distribuição

e consumo de alimentos, relacionadas à homogeneização dos hábitos alimentares atuais (Martínez-González et al., 2008).

Estudos epidemiológicos mais recentes, tais como: Kesse-Guyot et al. (2013), Mischiagna et al. (2017), Cummings & Bingham (1998), Kontou et al. (2011), e Lichtenstein et al. (2006) estabeleceram os benefícios para a saúde associados à adesão ao modelo de dieta mediterrânea, especialmente em termos de redução do risco de desenvolver síndrome metabólica, diabetes tipo 2, doenças cardiovasculares, algumas doenças neurodegenerativas e câncer.

Dentro da comunidade científica, há uma visão compartilhada do que constitui a pirâmide tradicional da dieta mediterrânea, pirâmide essa formada por alimentos ricos em grãos integrais, consumo moderado de queijo e vinho e consumo de azeite doce como única fonte de gordura. Verifica-se sempre a necessidade de atualização as recomendações presentes, considerando o novo estilo de vida. Especialmente os grupos de baixa renda, para os quais a prevalência de doenças cardiovasculares aumentou juntamente com o desvio de hábitos alimentares da dieta mediterrânea tradicional.

Essa pirâmide atualizada resulta do consenso científico entre os especialistas, e se baseia na mais recente pesquisa em nutrição e saúde e evidências científicas sobre o efeito saudável do modelo de dieta mediterrânea e seu papel na prevenção de muitas doenças crônicas, documentada em uma grande quantidade de estudos publicados em centenas de artigos científicos nas últimas décadas.

A representação gráfica do modelo da pirâmide representada na Figura 1 é constituída pela base de alimentos que devem suportar a dieta e fornecer um alto consumo de energia e níveis mais altos de alimentos a serem consumidos em quantidades moderadas, como as de origem animal e / ou ricas em açúcares e gordura, que deve ser consumida de forma moderada e alguns deixados para ocasiões especiais. As refeições têm um papel essencial na dieta mediterrânea, e, portanto, a importância das refeições e sua composição é enfatizada na nova representação. Uma composição equilibrada das principais refeições deve incluir frutas, vegetais, pratos de cereais suplementados por outros alimentos vegetais, produtos lácteos e fontes de proteína.

Mediterranean Diet Pyramid: a lifestyle for today
Guidelines for Adult population

Serving size based on frugality
and local habits

Wine in moderation
and respecting social beliefs



Fundación
Dieta Mediterránea

ICAF
International Commission on the
Anthropology of Food and Nutrition

FORUM ON
MEDITERRANEAN
FOOD CULTURES

Predimed
Promoción una Dieta Mediterránea

Ciisam
Centro de Investigación en
Salud y Alimentación
Mediterránea

HMF
Hellenic Mediterranean
Foundation

IUNS
International Union of
Nutritional Sciences

CIHEAM
International Centre for
Horticulture and
Food and Nutrition

fen
Federation of
European Nations

Figura 1: Pirâmide da dieta Mediterrânea

Fonte: Public Health Nutre (Alwan et al., 2011).

3 MATERIAL E MÉTODOS

Os dados utilizados neste trabalho referem-se a um estudo observacional retrospectivo, durante um período de seguimento de 10 anos, realizado em duas coortes, totalizando 2.623 indivíduos selecionados segundo a lista de eleitores dos municípios de Castellana Grotte e Putignano, ambos na Itália. Para a pesquisa, foram selecionados indivíduos que atenderam à convocação (Croxford et al., 2017). A variável resposta é o tempo até a primeira falha e o tipo de primeira falha, onde o evento de interesse principal é a falha decorrente do óbito por câncer gastrointestinal enquanto o evento competitivo é a falha devido ao óbito por outras causas.

3.1 Descrição das coortes

3.1.1 Coorte Multicentrica Italiana Colelitiasi (Micol)

A pesquisa foi realizada em uma amostra aleatória da lista de eleitores do município de Castellana Grotte (Osella et al., 1997). A cidade se localiza ao sul da província de Bari, a cerca de 50 *km* da capital, a 14 *km* do mar Adriático, e cerca de 250 *m* acima do nível do mar. Em 1984, o território media 68 *km*² e abrigava 17.334 habitantes, 30% dos quais trabalhavam na agricultura.

A coorte Micol teve início em 1985-1986. Da amostra de 3.500 indivíduos selecionados pelo padrão eleitoral, responderam 2.472 (70,6%), sendo 1.429 (57,8%) do gênero masculino e 1.043 (42,2%) do feminino. Esses indivíduos foram seguidos (primeiro acompanhamento) em 1992-1993 e 2005-2006 (segundo acompanhamento). Em 2005-2006, 1.705 (69% das coortes originais) responderam ao convite para acompanhamento, sendo 948 (55,6%) homens e 757 (44,4%) mulheres. Ao mesmo tempo, a população foi compilada entre 30 e 50 anos de idade, para alimentar a coorte com pessoas dessa faixa etária, devido ao envelhecimento da coorte. Com uma amostragem sistemática de 1:5, foram extraídos 1.906 indivíduos dessa faixa etária. Desses, 1.268 (66,5%) foram recrutados e aceitos, sendo que nessa nova amostragem o número de homens passou a 731 (57,6%) e o de mulheres ficou em 537 (42,4%).

Os entrevistados receberam novamente um formulário destinado a detectar dados demográficos, clínicos e nutricionais, bem como um questionário semiquantitativo de consumo

alimentar (*Food Frequency Questionnaire* – FFQ ou Questionário de Frequência Alimentar – QFA), preparado *ad hoc* e validado, além de realizarem exames de ultrassonografia, de sangue e hepático-hemorragicos.

3.1.2 Coorte Nutrizione Epatologia (Nutriep)

A pesquisa foi realizada pelo Instituto de Hospitalidade e Cuidados em Escrita Científica Saverio De Bellis (Castellana Grotte, Bari, Itália), com a colaboração de doze médicos de família que trabalham em Putignano. O estudo foi realizado no período de julho de 2005 a janeiro de 2007 (Osella et al., 1997).

Putignano é uma pequena cidade, com população de aproximadamente 27.000 habitantes na região de Puglia, no Sul da Itália. Fica a 40 *km* de Bari, a maior cidade de Puglia, com aproximadamente 322.000 habitantes. Putignano foi escolhida para a realização do estudo pelos seguintes motivos: 1) a renda *per capita* situa-se no ranking médio; 2) a distribuição da população por gênero e faixa etária se assemelha à do Sul da Itália; e 3) verifica-se uma grande disponibilidade de colaboração por parte dos médicos de família.

Usando uma amostragem sistemática de 1:5 (Osella, 2000), extraiu-se uma amostra de pacientes assistidos pelos médicos de família, com idade de 18 anos ou mais. As listas dos médicos de família foram utilizadas no lugar das listas eleitorais, pois não houve diferenças significativas entre a distribuição por gênero e idade da população em geral e a lista dos médicos de família. Assim, supõe-se que os pacientes distribuem-se casualmente entre os médicos.

Na Itália, o médico de família é obrigatório. Portanto, as listas dos médicos de família correspondem às listas eleitorais. Um possível viés de seleção é que subgrupos específicos de pacientes (idosos, com doenças crônicas, com doença hepática e outras) podem frequentar médicos de família em um período limitado. Para minimizar esse potencial elemento de confusão, uma análise estatística foi realizada para testar se a idade média da população era a mesma que a do elenco de médicos de família. Assim, usaram-se a Anova e o teste de Bonferroni de comparações múltiplas. A idade média da população em geral foi calculada em 47,5 anos ($\pm 18,6$), enquanto a idade média dos indivíduos inscritos e agrupados em quatro consultórios de 12 médicos de família foram calculadas em 47,5 ($\pm 15,8$), 47,02 ($\pm 16,3$), 46,61 ($\pm 15,8$) e 46,05 ($\pm 14,09$) anos, respectivamente ($p = 0,61$). A Anova foi

usada também para testar a hipótese de que a média específica gênero-idade entre os pacientes dos médicos de família é igual à média da população geral. Não houve evidência significativa de diferença estatística ($p = 0,15$).

Portanto, 2.550 indivíduos que compõem a amostra foram convidados a participar do estudo, 2.301 dos quais (90% dos participantes) aceitaram participar, dando o consentimento de acordo com a Declaração de Helsinski. Os indivíduos participantes responderam a um questionário pré-codificado que registra informações sociodemográficas e a história médica pessoal. Em cada um dos participantes foram tomadas as medidas antropométricas, assim como foi realizada uma ecografia hepática e colhida uma amostra de sangue.

Os indivíduos também responderam ao questionário *European Prospective Investigation into cancer and Nutrition* (EPIC). Trata-se de um instrumento de coleta de dados usado em toda a Europa para registrar informações sobre os hábitos alimentares da população levando em conta a quantidade e a frequência das porções consumidas, podendo também ser vistas por meio de figuras apresentadas no mesmo questionário.

3.1.3 Elaboração da dieta Padrão

O *Food Frequency Questionnaire* (FFQ) relativo ao estudo na coorte Micol foi introduzido e elaborado com um *software ad hoc* para obter as quantidades diárias (gr/dia) de 87 alimentos diferentes consumidas por cada indivíduo. FFQ EPIC foi enviado para sua elaboração no Instituto Nazionale dei Tumori, em Milão. Esse órgão possui uma sede nacional única, onde os questionários são elaborados e se obtém o consumo diário (gr/dia) de 123 alimentos diferentes.

Os alimentos referidos nos questionários foram agrupados em quantidades para construir grupos similares. Foram considerados os seguintes grupos de alimentos: água, bebidas alcoólicas, bebidas não alcoólicas, carnes e ovos, carnes processadas, chá-café, doces-biscoitos-pães, frutas *in natura*, frutas secas, gorduras, grãos refinados, grãos integrais, legumes, leite e iogurte, peixes e mariscos, produtos com açúcar agregado, produtos doces à base de leite, queijos frescos, queijos temperados e vegetais não amiláceos.

Os grupos de alimentos foram submetidos a uma análise fatorial, usando-se o método de componentes principais, extraíndo-se três fatores, aos quais aplicou-se posteriormente uma rotação *varimax* (Osella, 2000).

Os três fatores extraídos foram os seguintes: padrão mediterrâneo (frutas *in natura*, grãos integrais, vegetais não amiláceos), padrão ocidental (carnes, carnes processadas) e padrão com alto consumo de açúcar (bebidas não alcoólicas, doces, etc). Ao final, foi extraído um escore padronizado para cada padrão.

Após esse procedimento, foi construído um índice de aderência para a dieta mediterrânea (Buckland et al. (2008); Contreras (1993)), com o objetivo de se obter um escore categorizado. Atribuiu-se o valor “0” ou “1” a cada um dos nove grupos de alimentos da dieta, usando-se a mediana como ponto de corte. Os componentes considerados benéficos (cereais, frutas *in natura*, frutas secas, legumes, peixes e mariscos, vegetais) foram categorizados em “1” para as pessoas com consumo igual ou superior à mediana, e “0” para aquelas com consumo inferior. Para aqueles alimentos considerados prejudiciais (carnes, frangos e produtos lácteos), atribuiu-se o valor “1” se o consumo era inferior à mediana, e o valor “0” se o consumo era igual ou superior. Para o etanol, foi atribuído o valor “1” se o consumo diário ficava entre 10 e 50 gramas para os homens e entre 5 e 25 gramas para as mulheres. Assim, o índice de aderência à dieta mediterrânea variaria de “0” (aderência mínima) até “9” (adesão máxima). Esse índice foi dicotomizado em aderência mínima (0-3) e boa aderência (4-9) (Contreras, 1993).

Foi solicitada ao Setor Anagrafe dos municípios de Castellana Grotte e Putignano. A lista dos indivíduos que faleceram entre janeiro de 2005 e dezembro de 2016. Posteriormente, as pessoas falecidas no intervalo mencionado foram identificadas juntamente com as respectivas coortes.

O próximo passo consistiu na identificação da *causa mortis* de cada um dos sujeitos pertencentes às coortes. O setor responsável do Laboratório de Epidemiologia e Bioestatística, em colaboração com os setores correspondentes do Sistema Nacional de Saúde da Itália, identificou as causas de óbito, que foram codificadas como Classificação Internacional de Doenças Versão 10 (ICD10).

A seleção de variáveis a serem incluídas no modelo, como em outros modelos de regressão, pode ser realizada por meio de vários métodos de seleção, como, por exemplo, o método *stepwise* e os critérios de seleção AIC e BIC.

Royston & Parmar (2002) afirmam que possivelmente as estimativas dos coeficientes são semelhantes entre o modelo Weibull, o modelo de Cox e os modelos flexíveis de riscos

proporcionais. Logo, as variáveis a serem incluídas no modelo podem ser selecionadas pelo modelo de Cox ou pelo modelo Weibull. No presente estudo, a seleção dessas variáveis foi realizada com o auxílio do modelo de Cox.

Nesta pesquisa foram consideradas características de saúde, sociodemográficas e aderência às dietas, cujos valores foram obtidos por meio da aplicação do questionário no início do estudo.

Selecionaram-se as variáveis consideradas mais apropriadas para a análise, as quais foram ao mesmo tempo relacionadas ao evento da morte. Foram consideradas as seguintes variáveis: idade na matrícula, gênero, nível educacional, trabalho realizado, Índice de Massa Corporal (IMC), pontuação padronizada do padrão dietético ocidental, pontuação padronizada do padrão dietético mediterrâneo, pontuação padronizada pelo índice elevado de carboidratos e Índice de Aderência a Dieta Mediterrânea (IADM). A seleção baseou-se em critérios biológicos e sociais, bem como na exposição a fatores relacionados ao estilo de vida (nutricionais). As variáveis relacionadas ao histórico de saúde dos sujeitos foram excluídas da análise, porque se considerou que todos os eventos patológicos da história pessoal estavam contidos na causa específica da morte.

A idade na inscrição (em anos) foi calculada automaticamente como a diferença entre a data de inscrição e a data de nascimento encontradas no documento em questão. A idade na inscrição foi utilizada como uma variável contínua nos modelos estatísticos e como uma variável categórica para a descrição da amostra (50-59, 60-69, 70 ou mais). A categoria de referência para a variável gênero foi o gênero masculino em todos os casos. A variável nível educacional foi categorizada de acordo com o sistema educacional italiano, na forma a saber: Escola primária, 1º colegial, 2º colegial, universidade, analfabetos e sem informação. A variável tipo de trabalho foi categorizada de acordo com a classificação-padrão internacional de profissões, em: Gerentes e Profissionais, Artesãos, Trabalhadores Agrícolas e Vendedores, Obras Primárias, Aposentados, Sem Trabalho e Sem Informação.

O IMC (Índice de Massa Corporal) foi utilizado como uma variável contínua e categórica com base em dados sobre peso e altura. O peso (kg) foi medido com uma balança eletrônica digital, e a altura (m) foi obtida com um altímetro de parede. O IMC foi estimado como $\text{peso}/(\text{altura})^2$. As categorias de IMC foram construídas de acordo com as diretrizes da OMS, sendo distribuídas em três faixas, a saber: peso normal (até 24,9),

sobrepeso ($25 - 29,9$) e obesidade (≥ 30).

Os escores padronizados das dietas ocidental, mediterrânea e alto teor de açúcares foram utilizados como variáveis contínuas. A variável IADM foi dicotomizada em baixa aderência (< 4) e boa aderência (≥ 4).

Dois diferentes modelos de regressão foram ajustados: o modelo de Cox para o risco específico por causa e o modelo Fine e Gray para o risco de subdistribuição.

As análises dos dados foram realizadas no *software* Stata 15 Stata (2017), utilizando as funções *stpm2*, *stcox*, *trt*, *stset* e *stcompt* do pacote *survival*.

3.2 Métodos

Neste estudo foram analisados os eventos competitivos óbito por câncer gastrointestinal e óbito por outras causas, considerando-se que a ocorrência de um dos eventos exclui a possibilidade de ocorrência do outro.

Dos 2623 indivíduos, 45 (1,71%) experimentaram o evento de interesse (óbito por câncer gastrointestinal), 128 (4,87%) experimentaram o evento competitivo (óbito por outras causas).

Os modelos de Cox para o risco específico por causa e o modelo Fine e Gray para o risco da subdistribuição foram utilizados para avaliar o óbito por câncer gastrointestinal e outras causas.

O tempo foi medido em meses. O início da observação coincide com a realização da entrevista entre 2005 e 2007, enquanto o final da observação coincide com a data de morte por câncer gastrointestinal ou outra causa, por mudança de residência (perda de seguimento) ou por final do período de observação em 31 de dezembro de 2016.

Inicialmente, foi feito um ajuste em todos os modelos com uma única variável, sendo incluídas aquelas que apresentaram significância de 10%. Em seguida, foi aplicada uma análise clássica do modelo de Cox, para comparação dos resultados e verificação da não violação da suposição de riscos proporcionais para as covariáveis de interesse no estudo.

Posteriormente, foram aplicados modelos lineares generalizados Osella (2000), com distribuição de Poisson para avaliar a semelhança das estimativas com um modelo exponencial. Para se atingir esse objetivo, o tempo de observação de cada um dos sujeitos foi dividido em vários intervalos, assumindo-se que o risco é constante nos limites dos

intervalos, mas podendo variar entre eles.

Se o número de intervalos for definido pelo tempo em que ocorre cada um dos eventos durante a observação, o uso de um modelo de Poisson é matematicamente equivalente ao do modelo de Cox. Além disso, se o tempo de observação for dividido em pequenos intervalos, uma estimativa suavizada da função de risco basal pode ser obtida usando-se *splines* ou polinômios fracionários. A suavização da função basal pode então ser controlada com o número de nós, podendo o AIC e o BIC ser empregados para avaliar o processo.

Desse modo, o tempo de observação foi fracionado em períodos de um mês, para os quais foram aplicadas splines com 4 nós. Usando-se sucessivamente os comandos de pós-estimação Mozumder et al. (2017), a taxa de risco foi estimada para as duas categorias de exposição consideradas: morte por câncer gastrointestinal ou morte por outras causas, e para a taxa de risco basal com seu intervalo de confiança correspondente.

As taxas de risco basal foram estimadas para diferentes números de nós, sendo o mesmo procedimento aplicado ao logaritmo do tempo, em vez do tempo.

O estágio sucessivo consistiu na aplicação de polinômios fracionários, que constituem uma alternativa para modelar o risco basal, e são extensões dos polinômios-padrão, que oferecem grande flexibilidade para modelar relações não lineares. Para se avaliar a adaptação do modelo, foi utilizado um polinômio de potência 3. A avaliação foi realizada por meio da análise do desvio do modelo (Royston & Altman, 1994).

A última parte da análise consistiu na aplicação de modelos paramétricos flexíveis e modelos paramétricos flexíveis para riscos competitivos (Mozumder et al., 2017). Os modelos paramétricos flexíveis oferecem a possibilidade de se modelar a taxa de risco basal com uma função cúbica restrita, sem ter que dividir o tempo de observação. Além disso, o modelo adaptado é geralmente insensível ao número e às posições dos nós, sendo possível avaliar formalmente a forma, usando o AIC e o BIC. Para essa análise, foi usado o comando *stpm2* do *software* Stata 15.

Finalmente, os modelos flexíveis para riscos competitivos possibilitaram a avaliação simultânea de todas as distribuições acumuladas de incidência específica por causa, assim como a aplicação de modelos multiplicativos (razão de risco) e aditivos (Mozumder et al., 2017). Para essa análise, também foi utilizado o comando - *stpm2* do *software* Stata 15.

A mortalidade geral (censurada, mudança de endereço) foi utilizada para aplicar as

técnicas estatísticas dos modelos de Kaplan-Meier, Cox e modelos flexíveis de sobrevivência paramétricos. Um desfecho foi utilizado para a análise de riscos competitivos, consistindo em identificar todas as causas de morte por câncer gastrointestinal (CID10 – ICD10 C15 - C26) *versus* todas as outras causas de morte.

Ao término do estudo, foi realizada uma análise descritiva do valor de cada variável independente, com o objetivo de descrever e resumir as características da amostra.

4 RESULTADOS E DISCUSSÃO

A coorte de pacientes estudada reúne 2.623 indivíduos, dos quais 1.644 pertencem à coorte Micol e 979 pertencem à coorte Nutriep, sendo 1.439 homens (54,9%) e 1.184 mulheres (45,1%).

Dos 2623 indivíduos, 45 (1,71%) experimentaram o evento de interesse (óbito por câncer gastrointestinal), 128 (4,87%) experimentaram o evento competitivo (óbito por outras causas).

Verificou-se que a maioria dos indivíduos tinha idade entre 50 a 69 anos (40,7%), e que 28,4% tinham idade igual ou superior a 70 anos. Em relação à escolaridade, verificou-se a predominância do nível básico (43,8%). Aproximadamente 33% dos indivíduos são profissionais das áreas de vendas e artesanato, pensionistas e donas de casa, ou estão desempregados, e, 43,5% são obesos (Tabela 2).

Tabela 2: Estatísticas descritivas para dados demográficos das coortes Micol e Nutriep.

Dados relacionados as variáveis estudadas, evento de interesse e falha.

Variáveis		0 - Censura	1 - Óbito
		Frequência (%)	Frequência (%)
Projeto	Micol	1318 (80,2%)	326 (19,8%)
	Nutriep	885 (90,4%)	94 (9,6%)
Idade (anos)	50 - 59	1031 (96,5%)	37 (3,5%)
	59 - 69	726 (89,5%)	85 (10,5%)
	70 ou mais	446 (59,9%)	298 (40,1%)
Sexo	Masculino	1204 (83,7%)	235 (16,3%)
	Feminino	999 (84,4%)	185 (15,6%)
Educação	Primário	1149 (79,6%)	294 (20,4%)
	Secundário	474 (92,6%)	38 (7,4%)
	Ensino Médio	338 (92,3%)	28 (7,7%)
	Graduação	94 (91,3%)	9 (8,7%)
	Analfabeto	109 (71,2%)	44 (28,8%)
	Sem Informação	9 (60,0%)	6 (40,0%)
Trabalho	Gerentes e profissionais	80 (88,9%)	10 (11,1%)
	Artesanato, agricultura e vendas	585 (88,6%)	75 (11,4%)
	Ocupação elementar	667 (77,1%)	198 (22,9%)
	Dona de casa	358 (91,1%)	35 (8,9%)
	Aposentados	464 (83,3%)	93 (16,7%)
	Desempregados	15 (88,2%)	2 (11,8%)
	Sem Informação	3 (33,3%)	6 (66,7%)
Índice de Massa Corporal (IMC)	Normal	272 (78,4%)	75 (21,6%)
	Sobrepeso	753 (83,9%)	145 (16,1%)
	Obeso	970 (84,9%)	173 (15,1%)
Dieta Ocidental Padrão	< 33	781 (80,8%)	186 (19,2%)
	$33 \leq x < 66$	703 (85,1%)	123 (14,9%)
	≥ 66	719 (86,6%)	111 (13,4%)
Dieta Mediterrânea Padrão	< 33	633 (86,6%)	98 (13,4%)
	$33 \leq x < 66$	713 (82,3%)	153 (17,7%)
	≥ 66	857 (83,5%)	169 (16,5%)
Dieta Rica em Açúcar	< 33	651 (86,0%)	106 (14,0%)
	$33 \leq x < 66$	682 (82,8%)	142 (17,2%)
	≥ 66	870 (83,5%)	172 (16,5%)
Câncer Total	0 - Ausência	2198 (88,1%)	297 (11,9%)
	1 - Presença	5 (3,9%)	123 (96,1%)
Câncer Gastrointestinal	0 - Ausência	2202 (85,4%)	376 (14,6%)
	1 - Presença	1 (2,2%)	44 (97,8%)

Fonte: Elaborada pela autora.

Os indivíduos foram questionados em relação aos seus hábitos alimentares, medindo-se, assim, o grau de aderência à dieta mediterrânea. Do total de indivíduos (2.623), 27,86% pertenciam ao escore padronizado menor que 33 (aderência mínima à dieta), e 39,11% pertenciam ao escore padronizado maior ou igual a 66 (boa aderência à dieta). Em relação à dieta ocidental padrão, observou-se que dentre os vivos, a adesão à dieta é semelhante, enquanto dentre os mortos percebe-se uma pequena variação.

Em relação a outros tipos de câncer, observou-se que do total de indivíduos que apresentam algum tipo de câncer, 96,1% estão mortos e que dos que apresentam câncer gastrointestinal, 97,8% estão mortos.

No final do estudo, foi observado que o escore padronizado da dieta mediterrânea apresentou um comportamento semelhante nas três categorias em que foi dividido, com percentuais de adesão em torno de 80% dos indivíduos que não apresentaram o evento, e de 15% para os que faleceram.

A estimação do número de nós do modelo de Cox foi exposta na Tabela 3 onde a distribuição das taxas de risco basal estimadas para evidenciar a escolha do número de nós em diferentes graus de liberdade são apresentadas.

Tabela 3: Estimação do parâmetro de boa aderência a dieta mediterrânea ($IADM \geq 4$) do modelo de Cox com diferentes graus de liberdade no tempo e no log do tempo para identificar os efeitos no aumento do número de nós.

Variável	gl = 2	gl = 3	gl = 4	gl = 5	gl = 6	gl = 7	gl = 8
$MAI \geq 4$ /tempo	0,33*	0,32*	0,33*	0,32*	0,32*	0,32*	0,32*
$MAI \geq 4$ /log do tempo	0,33*	0,32*	0,33*	0,32*	0,32*	0,32*	0,32*

(*)valor – $p < 0,05$.

Fonte: Elaborada pela autora.

Observa-se que não existem efeitos muito diferentes quando se acrescentam mais do que 4 nós à função de risco basal no tempo ou no logaritmo do tempo. Na realidade, o que acontece é uma perda de significância, corroborando as afirmações feitas por Fine & Gray (1999), em relação à quantidade de nós utilizados.

As Figuras 2 e 3 apresentam os efeitos da variação do número de nós para os dados utilizados na pesquisa referentes às coortes Micol e Nutriep no tempo e no logaritmo do tempo entre 1 e 7 nós interiores (2-8 gls) para o tempo de seguimento.

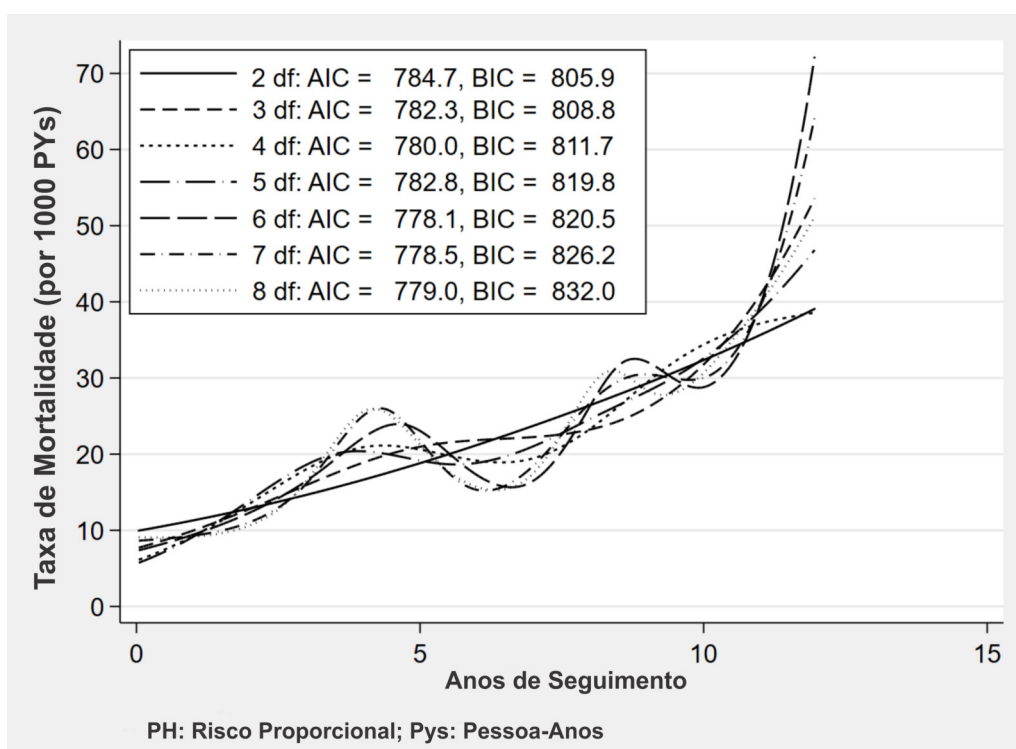


Figura 2: Efeito da variação do número de nós na função *spline* na estimação da função de risco basal

Fonte: Elaborada pela autora.

A curva ajustada para o modelo com grau de liberdade igual a 2 é visivelmente diferente da dos outros modelos. A forma da função de risco é semelhante para os outros modelos, mas percebe-se alguma fragmentação ao longo do tempo em que a função de risco atinge seu ponto mais alto. O AIC mais baixo refere-se ao modelo com 6gl. Assim, percebe-se que a variação do número de nós não apresenta ganho para o ajuste da taxa de risco, quando se colocam mais do que 4 nós.

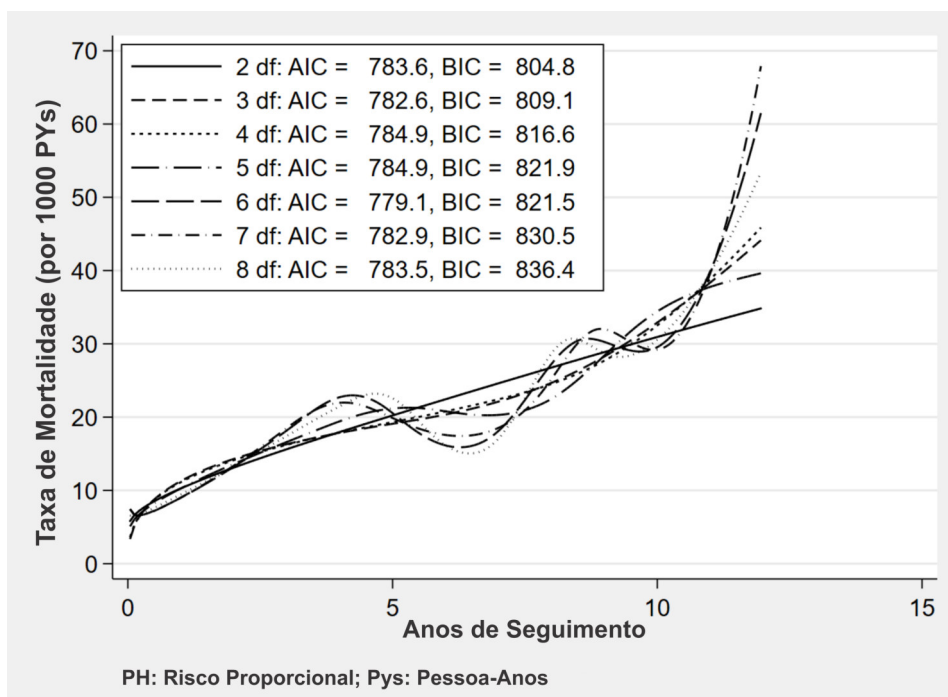


Figura 3: Efeito da variação do número de nós no log do tempo

Fonte: Elaborada pela autora.

Ao se estimar o efeito da variação em relação ao logaritmo do tempo, percebe-se um comportamento muito semelhante ao da Figura 2. Quando aplicados ao logaritmo do tempo, os valores estimados da taxa de risco são muito semelhantes entre os modelos, mesmo para o modelo com $gl = 2$.

Royston & Lambert (2011) e Royston & Parmar (2002) mostraram que quando uma variável é inclinada positivamente (como o tempo de sobrevivência geralmente o é), uma função *spline* na escala logarítmica geralmente proporcionará um melhor ajuste do que uma *spline* na escala não transformada para modelos com o mesmo número de graus de liberdade.

No presente estudo, não se perceberam diferenças no ajuste, quando estimado em relação ao tempo ou ao logaritmo do tempo.

Para se comparar os resultados e se identificar o modelo adequado aos dados do estudo, foi feita uma análise clássica do modelo de Cox (4).

Após completar-se a análise univariada, em que cada covariável foi tratada isoladamente, fez-se a seleção e inclusão das variáveis no modelo múltiplo, por meio do critério

de Akaike e do método *stepwise* de inclusão progressiva com eliminação regressiva. A análise inicial foi feita para o modelo nulo, isto é, o modelo que inclui apenas o parâmetro constante (β_0). Na sequência, foram ajustados aqueles modelos com a covariável IADM e todas as outras covariáveis, sendo escolhido o modelo que apresentou menor valor de AIC (6002,316).

Na análise do modelo completo e verificação do melhor ajuste pelos critérios de AIC e BIC, o modelo selecionado foi aquele em que constam as variáveis estudo, idade, gênero e IADM, como se pode verificar na Tabela 4.

Tabela 4: Estimativas obtidas pelo modelo clássico de Cox para inclusão de covariáveis. Razão de risco, erro padrão, valor de z, valor de p e IC.

Variável	Razão de Risco	Erro Padrão	z	$P > z $	$I.C._{.95\%}$	
Estudo						
Secundário	0,574	0,101	-3,17	0,002	0,407	0,809
Colegial	0,638	0,128	-2,24	0,025	0,430	0,946
Graduação	0,728	0,248	-0,93	0,351	0,374	1,418
Analfabetos	1,188	0,193	1,06	0,29	0,863	1,634
Informação	1,679	0,695	1,25	0,211	0,746	3,779
Idade	1,090	0,004	22,77	0	1,082	1,098
Sexo						
Feminino	0,887	0,089	-1,19	0,232	0,729	1,080
IADM						
Entre 4 e 5	0,785	0,109	-1,65	0,099	0,611	1,043

Fonte: Elaborada pela autora.

A partir do resultado da análise, em que cada variável foi considerada isoladamente, pode-se afirmar que a variável estudo nas classes secundário e ensino médio é a única que revela influência significativa no óbito por câncer gastrointestinal e a aderência à dieta mediterrânea na amostra das duas coortes. Apesar de as variáveis idade, gênero e IADM não apresentarem significância, elas foram mantidas no modelo, pois, por critérios epidemiológicos, são consideradas importantes e afetam o evento de interesse.

Um modelo de Poisson foi ajustado para função de risco de base com *splines* cúbicas. A Figura 4 apresenta o comportamento da taxa de mortalidade para os dois grupos, $IADM \leq 4$ e $IADM \geq 4$.

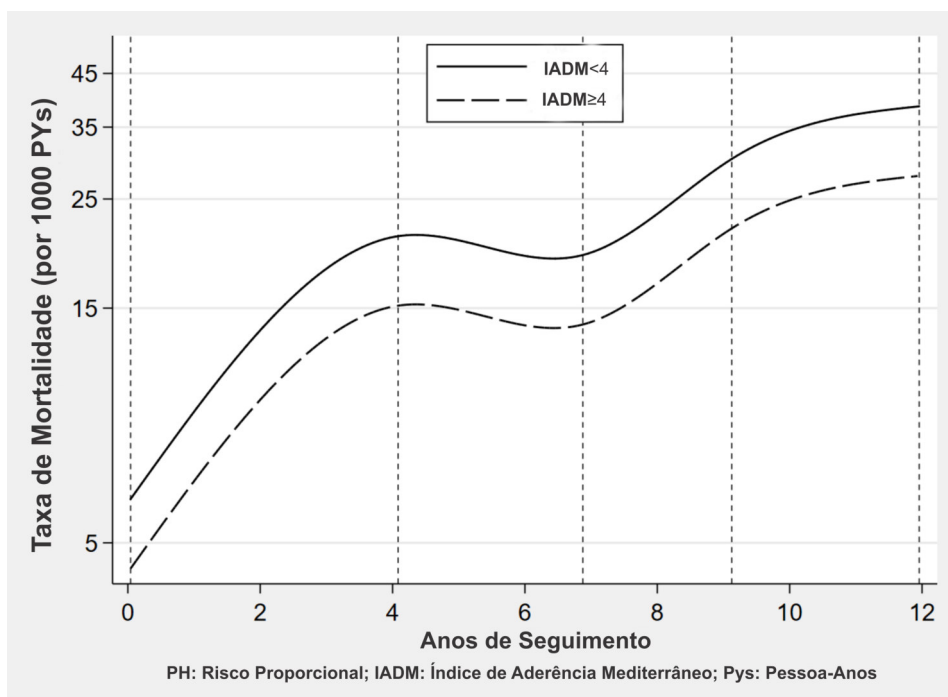


Figura 4: Modelo Poisson para o índice de aderência a dieta mediterrânea (IADM) com spline cúbica restrita para a função de risco basal.

Fonte: Elaborada pela autora.

Verifica-se que após a suavização da função de risco basal pelo modelo Poisson, a mortalidade se apresenta maior no grupo que tem baixa aderência à dieta mediterrânea. As localizações dos nós estão mais próximas umas das outras no final da escala de tempo, o que é indicativo do comportamento dos dados de sobrevivência.

Plotar a função de risco em ambos os grupos também demonstra a suposição do modelo de Poisson de riscos proporcionais em que a distância entre as duas linhas (na escala logarítmica) é a mesma em toda a escala de tempo. Esse fato foi mostrado também por meio do intervalo de confiança da função de risco basal (Figura 5).

Ao se analisar o intervalo de confiança, percebe-se que não há valores que apresentem comportamento atípico; ou seja, um grande aumento ou queda da taxa de risco provavelmente não se dará por acaso. É possível também verificar que no final do seguimento a taxa de risco apresenta um IC maior, num comportamento típico de dados de sobrevivência.

Uma questão importante é saber se as localizações dos nós impactam a forma da

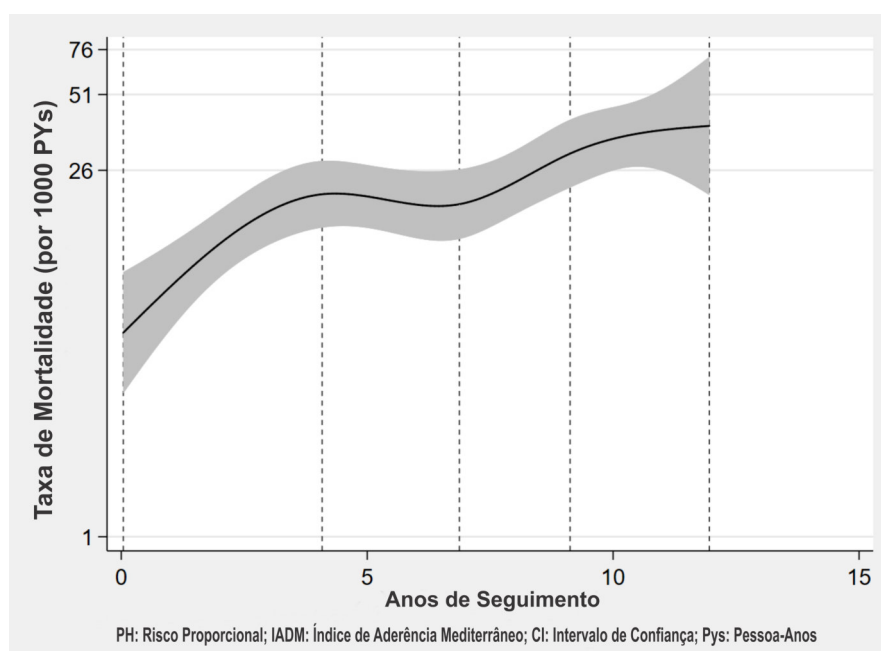


Figura 5: Modelo de Poisson para o índice de aderência a dieta mediterrânea (IADM) com *splines* cúbicas para a função de risco basal com 95% IC

Fonte: Elaborada pela autora.

função de risco estimada e as estimativas de interesse dos parâmetros. Para se investigar essa questão, foi ajustado um modelo com 6 nós (4 nós interiores e 2 nós-limite); cada nó foi selecionado aleatoriamente dentro de um intervalo de percentis definidos na Tabela 5.

Tabela 5: Localização de nós selecionados aleatoriamente para verificação da forma da função de risco estimada.

Nós	Rank*
1	$0^{th} - 5^{th}$
2	$6^{th} - 25^{th}$
3	$26^{th} - 50^{th}$
4	$51^{th} - 75^{th}$
5	$76^{th} - 95^{th}$
6	$96^{th} - 100^{th}$

* Intervalos referente aos percentis do log dos tempos dos eventos não censurados

Fonte: Elaborada pela autora.

Quando o ajuste de diferentes modelos com localização aleatória dos nós é efetivado, encontra-se uma pequena variação nos valores ajustados, e só é possível visualizar uma diferença no final do período de observação, o que é esperado, dado o pequeno número de eventos (morte por câncer gastrointestinal) no final do seguimento.

Para se modelar o risco basal e relações não lineares, fez-se uso de polinômios fracionários. Utilizou-se um polinômio de potência 3 para avaliar a adaptação do modelo, sendo a avaliação realizada por meio da análise do desvio do modelo (Figura 6).

Pode-se perceber que os comportamentos dos diferentes polinômios apresentam-se de forma muito semelhante no que tange à taxa de mortalidade e aos anos de seguimento; ou seja, a taxa de mortalidade cresce no decorrer do avanço do tempo. O FP3 se destaca, pois a taxa de mortalidade inicial é bem menor do que no FP1 e no FP2, sendo, portanto o que apresenta menor risco e melhor ajuste.

Para se identificar o melhor ajuste, comparou-se ainda o FP3 com o polinômio de potência 4, o FP4, percebendo-se que o FP3 continua apresentando as menores taxas de risco, sendo, portanto, o que melhor se ajusta à função (Figura 7).

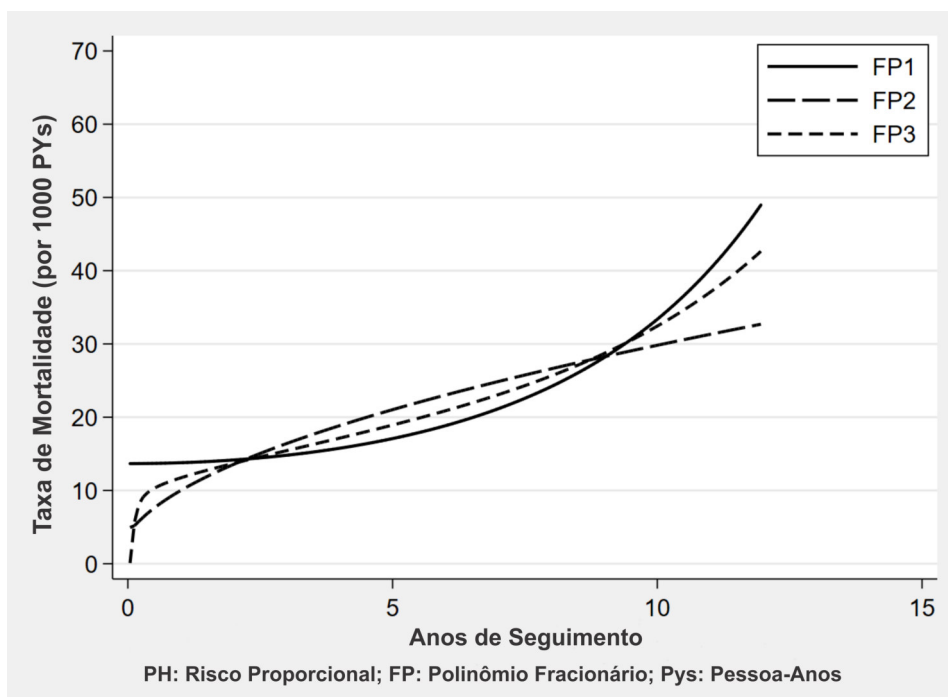


Figura 6: Estimação da função de risco de linha de base para um melhor ajuste FP1, FP2 e FP3

Fonte: Elaborada pela autora.

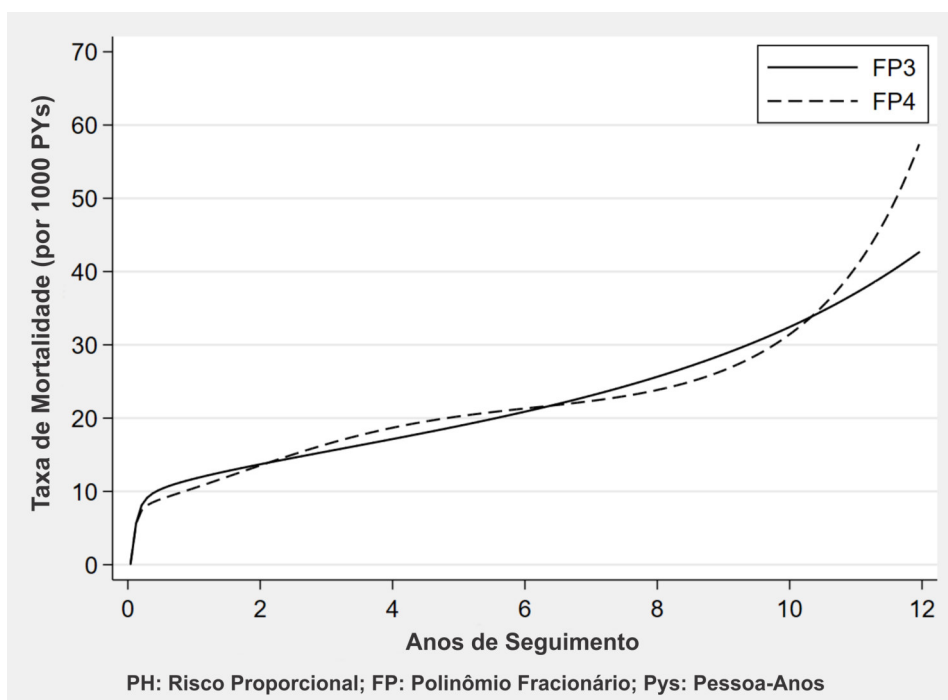


Figura 7: Comparação das funções com polinômios de grau 3 e 4 (FP3 e FP4) para a taxa de risco basal.

Fonte: Elaborada pela autora.

Os modelos paramétricos flexíveis foram empregados para modelar a taxa de risco basal com uma função cúbica restrita. A Tabela 6 mostra os efeitos da dieta mediterrânea nas coortes Micol e Nutriep.

Tabela 6: Modelo de regressão de Fine e Gray sobre o risco de subdistribuição para cada um dos desfechos: morte por câncer gastrointestinal e morte por outras causas – coortes Micol e Nutriep

	Câncer Gastrointestinal				Outras Causas			
	RR	EP	$IC_{95\%}$		HR	EP	$IC_{95\%}$	
IADM≥ 4	0,42	0,38*	0,2	0,89	0,71	0,14*	0,54	0,94
Idade (anos)	0,98	0,01	0,96	1,01	1,07	0,00**	1,07	1,08
Sexo	2	0,33*	1,05	3,83	0,95	0,11	0,77	1,18
Escore Ocidental	0,99	0,2	0,67	1,47	0,87	0,07*	0,76	0,99
Escore Rico em Açúcar	0,91	0,19	0,63	1,33	0,9	0,07	0,79	1,02

Legenda: * $p < 0.05$; ** $p < 0.01$; RR: Razão de Risco; EP: Erro Padrão;

IC: Intervalo de Confiança; IADM: Índice de Aderência Mediterrânea.

Fonte: Elaborada pela autora.

Observa-se que nos indivíduos com câncer gastrointestinal que apresentam boa aderência à dieta mediterrânea, o risco de morte é 0,42 vez menor do que naqueles que aderiram a outras dietas. Em comparação com as outras causas de morte, os indivíduos com boa aderência à dieta mediterrânea também apresentam riscos de morte menores do que os que aderiram a outras dietas.

Os modelos flexíveis para riscos competitivos foram usados para se avaliar as distribuições acumuladas de incidência específica por causa e a função de subdistribuição. A taxa de incidência acumulada para o $IADM \geq 4$ e para o $IADM \leq 4$ estão apresentadas na Figura 8 juntamente com a taxa específica por causa.

O gráfico (a) e (b) da Figura 8 mostram a função de risco de subdistribuição prevista para cada causa, enquanto o gráfico (c) e (d) ilustram a função de risco por causa específica por causa prevista por grau de aderência à dieta mediterrânea. O risco de subdistribuição mostra a associação do efeito do grau de aderência sobre o risco, e o risco de causa específica é a associação do efeito do grau de aderência sobre a taxa de risco (Royston & Lambert, 2011).

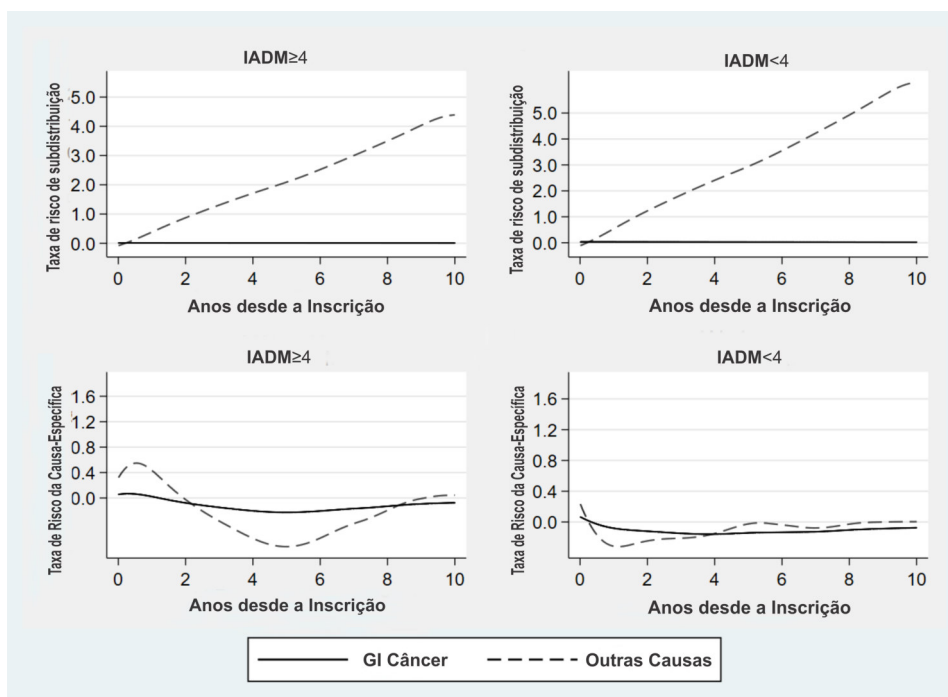


Figura 8: Função de risco de subdistribuição prevista para cada causa e taxa de risco específico por causa.

Fonte: Elaborada pela autora.

Pode-se observar através da função de subdistribuição que os indivíduos com câncer gastrointestinal que aderiram à dieta mediterrânea apresentam um risco constante e menor de morrer, especialmente na fase inicial e na fase final do seguimento, enquanto esse efeito é menos pronunciado em indivíduos que não aderiram à dieta mediterrânea.

Quando observamos a taxa de risco medido pela função específica por causa por grau de aderência, percebe-se um comportamento contrário, pois para os indivíduos com boa aderência a dieta o efeito é maior do que para os com baixa aderência a dieta mediterrânea. Mostrando assim o viés causado pelo uso do modelo de Kaplan-Meier.

A Figura 9 compara as funções de incidência acumulada específica por causa para morte por câncer gastrointestinal ou morte por outras causas preditas de maneira simultânea por um modelo de risco de subdistribuição acumulada e pelos modelos de Fine e Gray para a diferença entre os escores da dieta mediterrânea, considerando-se um intervalo de confiança de 95%.

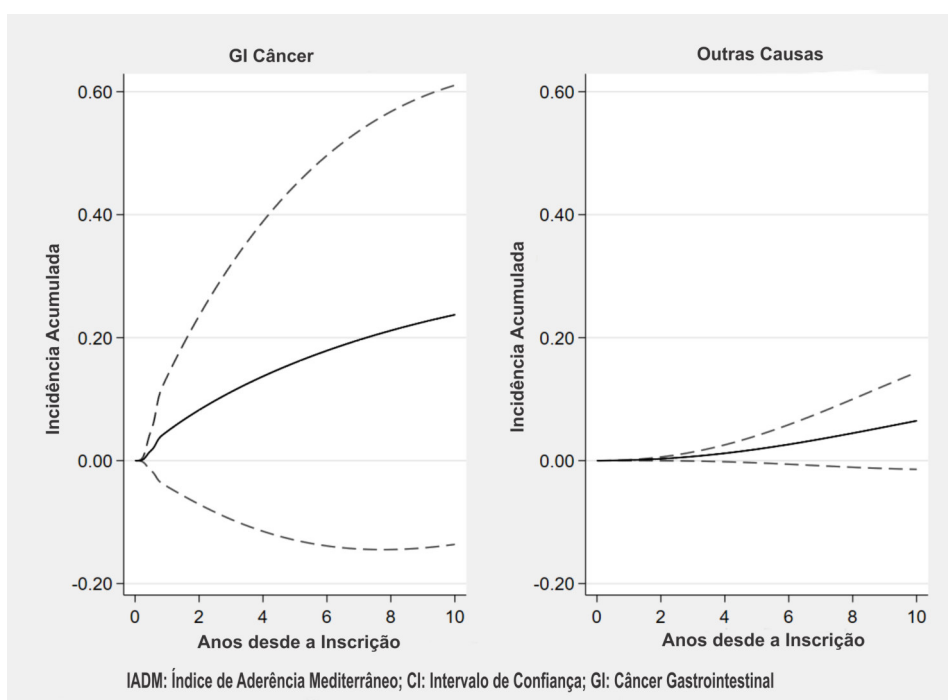


Figura 9: Previsão absoluta das diferenças ($IADM \leq 4 - IADM \geq 4$) nas funções de incidência acumulada específica da causa com 95% IC

Fonte: Elaborada pela autora.

Observa-se que a diferença de risco entre os sujeitos não aderentes e os sujeitos aderentes à dieta mediterrânea cresceu à medida que o tempo de seguimento aumentou (cerca de 20% no décimo ano de observação) entre os indivíduos que morreram de câncer gastrointestinal. A diferença de risco para sujeitos que morreram por outras causas, por outro lado, corresponde quase à metade (10%) no mesmo período de observação.

5 CONCLUSÃO

Este estudo faz uma abordagem dos métodos de análise de sobrevida nos aspectos paramétricos de modelos flexíveis para riscos competitivos.

Um dos seus objetivos consistia em aplicar o método estudado em uma coorte de dados de pacientes, para se avaliar os riscos de óbito para câncer gastrointestinal e outras causas, e perceber de que forma a aderência à dieta mediterrânea influencia o risco de óbito por um dos eventos analisados: morte por câncer gastrointestinal ou morte por outras causas.

Na aplicação dos métodos apresentados, analisou-se a sobrevivência de indivíduos portadores de câncer (gastrointestinal e outros) na presença de um modelo paramétrico flexível de risco competitivo, usando-se uma abordagem direta para a FIA e para a incidência acumulada específica por causa, incluindo variáveis sociodemográficas, clínicas e nutricionais.

Num primeiro momento, foi discutido o uso do método de Kaplan-Meier e do modelo de Cox para dados de sobrevivência na presença de riscos competitivos, verificando-se, então, ser inapropriado o uso do método tradicional de Kaplan-Meier para se calcular a probabilidade de um indivíduo sofrer um evento antes de determinado momento, bem como do modelo de Cox para se estimar o efeito das covariáveis para a ocorrência dos eventos, haja vista que, na presença de eventos múltiplos, estes fornecem uma estimativa enviesada da taxa de risco.

O método de Kaplan-Meier e o modelo de Cox consideram apenas falhas por evento de interesse, e tratam como censura não informativa todas as outras falhas causadas por eventos competitivos. Assim, ao se usar a função de distribuição acumulada de um evento na presença de riscos competitivos pelo método de Kaplan-Meier ou por meio do modelo de Cox, corre-se o risco de superestimar a função de sobrevida, assim como de não estimar adequadamente os efeitos das covariáveis.

Por ser de fácil interpretação e implementação, o modelo de Cox é bastante utilizado para análise de dados de sobrevivência. Entretanto, todo cuidado é pouco no momento de especificar a função de risco subjacente, pois ao se utilizar um modelo paramétrico adequado aos dados, isso levará à obtenção de estimativas mais precisas do risco e da

probabilidade de sobrevivência.

Em estudos recentes, alguns autores (Royston & Lambert (2011); Royston & Parmar (2002)) desenvolveram uma família de modelos paramétricos, com o objetivo não só de aumentar a flexibilidade dos modelos paramétricos padrão, como também de preservar as características de fácil interpretação e implementação do modelo de Cox.

De forma geral, nos modelos flexíveis, em nível univariado, a inclusão de nós internos melhora o processo de estimação das funções de sobrevivência. O aumento do número de nós internos presentes no modelo (até 3 no máximo) fornece uma melhoria, fazendo diminuir o valor de AIC associado a cada modelo (Godinho, 2016). Os modelos flexíveis com mais de 2 nós internos são os que chegam mais próximo das estimativas de Kaplan-Meier (Royston & Parmar, 2002).

Quando as regressões múltiplas são consideradas, os modelos flexíveis são os que apresentam o menor valor de AIC, tornando-se, então, mais adequados para descrever a função de risco e a função de sobrevivência. Portanto, apesar das vantagens do modelo de Cox, os modelos paramétricos flexíveis têm mais flexibilidade e fazem a correção do viés da superestimação na presença de riscos competitivos, representando, assim, de forma mais precisa a função de risco associada a cada grupo de indivíduos.

Nesta pesquisa, mostrou-se que a inclusão de nós internos no modelo flexível melhora a estimativa da taxa de risco, sendo o modelo com 3 nós o mais adequado e parcimonioso.

Em relação à função de risco de subdistribuição, a regressão sobre a função possibilitou a estimação dos efeitos das covariáveis para cada um dos tipos de desfecho considerados – morte por câncer gastrointestinal e morte por outras causas –, na presença dos eventos competitivos e aderência à dieta mediterrânea. Assim, foi possível observar que a aderência à dieta mediterrânea apresenta um risco permanente e menor de morte por câncer gastrointestinal do que a aderência a outras dietas nas duas coortes consideradas.

As funções de incidência acumulada específica por causa para morte por câncer gastrointestinal ou morte por outras causas preditas simultaneamente por um modelo de risco de subdistribuição acumulado e pelos modelos de Fine e Gray para o IC de 95% mostraram que a diferença de risco entre os sujeitos não aderentes e os aderentes da dieta mediterrânea cresceu à medida que o período de seguimento aumentou entre os indivíduos que morreram de câncer gastrointestinal.

Como trabalho futuro, seria interessante aplicar o presente estudo em outras morbidades, como, por exemplo, doenças cardiovasculares e diabetes, e analisar a associação à aderência à dieta mediterrânea e a outras dietas. Por fim, seria interessante também verificar se os modelos paramétricos flexíveis e o método da verossimilhança direta apresentam comportamento semelhante quando aplicados em diferentes doenças crônicas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, D. B. M.; MENEGUETTE-JR, M.; MONICO, J. F. Utilizando splines cúbicas naturais para atenuação de erros no posicionamento GPS. **Trends in Applied and Computational Mathematics**, v.6, n.2, p.197–205, 2005.

ALWAN, A.; ARMSTRONG, T.; BETTCHER, D.; BRANCA, F.; CHISHOLM, D.; EZ-ZATI, M.; GARFIELD, R.; MACLEAN, D.; MATHERS, C.; MENDIS, S.; POZNYAK, V.; RILEY, L.; TANG, K. C.; WILD, C. **Global status report on noncommunicable diseases 2010**. www.who.int/nmh/publications/ncd_report_full_en.pdf: World Health Organization, 2011.

ANDERSEN, P. K.; GESKUS, R. B.; WITTE, T.; PUTTER, H. Competing risks in epidemiology: possibilities and pitfalls. **International Journal of Epidemiology**, v.41, n.3, p.861–870, 2012.

AUSTIN, P. C.; LEE, D. S.; FINE, J. P. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. **Am J Epidemiol**, v.133, n.6, p.601–609, 2016.

BELOT, A.; ABRAHAMOWICZ, M.; REMONTET, L.; GIORGI, R. Flexible modeling of competitive risks in survival analysis. **Statistics in Medicine**, v.29, n.23, p.2453–2468, 2010.

BUCKLAND, G.; BACH-FAIG, A.; SERRA-MAJEM, L. Obesity and the Mediterranean diet: a systematic review of observational and intervention studies. **Obes Rev**, v.9, p.582–593, 2008.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de sobrevida: Teoria e aplicações em saúde**. 2. ed. Rio de Janeiro: Editora FIOCRUZ, 2011. 396p.

CHOODARI-OSKOOEI, B.; ROYSTON, P.; PARMAR, M. K. B. A simulation study of predictive ability measures in a survival model I: explained variation measures. **Statistics in medicine**, v.31, n.23, p.2627–2643, 2012.

COLLETT, D. **Modelling Survival Data in Medical Research**. 2. ed. Florida: Chapman and Hall/CRC, 2014. 548p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Blucher, 2006. 370p.

CONTRERAS, J. Antropología de la alimentación. **Ediciones de la Universidad Complutense**, p.9, 1993.

COVIELLO, V.; BOGGES, M. Cumulative incidence estimation in the presence of competing risks. **The Stata Journal**, v.4, n.2, p.103–112, 2004.

COX, D. R. Regression models and life-tables (with discussion). **Journal of the Royal Statistical Society**, v.34, p.187–220, 1972.

CROXFORD, S.; KITCHING, A.; DESAI, S.; KALL, M.; EDELSTEIN, M.; SKINGSLEY, A.; BURNS, F.; COPAS, A.; BROWN, A. E.; SULLIVAN, A. K.; DELPECH, V. Mortality and causes of death in people diagnosed with HIV in the era of highly active antiretroviral therapy compared with the general population: an analysis of a national observational cohort. **The Lancet Public Health**, v.2, n.1, p.35–46, 2017.

CUMMINGS, J. H.; BINGHAM, S. A. Diet and the prevention of cancer. **BMJ**, v.317, p.1636–1640, 1998.

FINE, J. P.; GRAY, R. J. A proportional hazards model for the subdistributions of a competing risk. **Journal of the American Statistical Association**, v.94, n.446, p.496–509, 1999.

GESKUS, R. B. **Data Analysis with Competing Risks and Intermediate States**. A Chapman and Hall/CRC, 2015. 245p.

GODINHO, A. R. S. Identificação de fatores que influenciam a persistência na terapêutica de uma doença cônica. Portugal, 2016. 83p. Dissertação (Mestrado) - Universidade de Lisboa - Faculdade de Ciências.

GOOLEY, T. A.; LEISENRING, W.; CROWLEY, J.; STORER, B. E. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. **Statistics in medicine**, v.18, n.6, p.695–706, 1999.

JEONG, J.-H.; FINE, J. Direct parametric inference for the cumulative incidence function. **J Royal Stat Soc**, v.55, p.187–200, 2006.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data**. 2. ed. New Jersey: John Wiley & Sons. Ltd, 2002. 439p. Wiley Series in Probability and Statistics.

KESSE-GUYOT, E.; AHLUWALIA, N.; LASSALE, C.; HERCBERG, S.; FEZEU, L.; LAIRON, D. Adherence to Mediterranean diet reduces the risk of metabolic syndrome: A 6-year prospective study. **Nutrition, Metabolism and Cardiovascular Diseases**, v.23, n.7, p.677–683, 2013.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis: A Self-Learning Text**. 3. ed. New York: Springer-Verlag, 2012. 700p. Statistics for Biology and Health.

KONTOU, N.; PSALTOPOULOU, T.; PANAGIOTAKOS, D.; DIMOPOULOS, M. A.; LINOS, A. The Mediterranean Diet in Cancer Prevention: A Review. **Journal of Medicinal Food**, v.14, n.10, p.1065–1078, 2011.

LAMBERT, P. C.; HINCHLIFFE, S. R. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. **BMC Medical Research Methodology**, v.13, p.1471–2288, 2013.

LAMBERT, P. C.; ROYSTON, P. Further development of flexible parametric models for survival analysis. **The Stata Journal**, v.9, n.2, p.265–290, 2009.

LAU, B.; COLE, S. R.; GANGE, S. J. Competing risk regression models for epidemiologic data. **Am J Epidemiol**, v.170, n.2, p.244–256, 2009.

LICHTENSTEIN, A. H.; APPEL, L. J.; BRANDS, M.; CARNETHON, M.; DANIELS, S.; FRANCH, H. A.; FRANKLIN, B.; KRIS-ETHERTON, P.; HARRIS, W. S.; HOWARD, B.; KARANJA, N.; LEFEVRE, M.; RUDEL, L.; SACKS, F.; VAN-HORN, L.; WINSTON,

M.; WYLIE-ROSETT, J. Diet and lifestyle recommendations revision 2006: a scientific statement from the American Heart Association Nutrition Committee. **Circulation**, v.114, n.1, p.82–96, 2006.

MARTÍNEZ-GONZÁLEZ, M. A.; FUENTE-ARRILLAGA, C.; NUNEZ-CORDOBA, J. M.; BASTERRA-GORTARI, F. J.; BEUNZA, J. J.; VAZQUEZ, Z.; BENITO, S.; TORTOSA, A.; BES-RASTROLLO, M. Adherence to Mediterranean diet and risk of developing diabetes: prospective cohort study. **BMJ**, v.336, n.7657, p.1348–1351, 2008.

MISCIAGNA, G.; DÍAZ, M. D. P.; CARAMIA, D. V.; BONFIGLIO, C.; FRANCO, I.; NOVIELLO, M. R.; CHILOIRO, M.; ABBRESCIA, D. I.; TANZI, A. M. M.; CARUSO, M. G.; CORREALE, M.; REDDAVIDE, R.; CISTERNINO, R. I. A. M.; OSELLA, A. R. Effect of a Low Glycemic Index Mediterranean Diet on non-alcoholic fatty liver disease. A randomized controlled clinical trial. **The Journal of Nutrition, Health & Aging**, v.21, n.4, p.404–412, 2017.

MOZUMDER, S. I.; RUTHERFORD, M. J.; LAMBERT, P. C. A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function. **The Stata Journal**, v.17, p.462–489, 2017.

NELSON, C. P.; LAMBERT, C. P.; SQUIRE, I. B.; JONES, D. R. Flexible parametric models for relative survival, with application in coronary heart disease. **Statistic in Medicine**, v.26, n.30, p.5486–5498, 2007.

NOORDZIJ, M.; VAN STRALEN, K. J.; ZOCCALI, C.; DEKKER, F. W.; JAGER, K. J. When do we need competing risks methods for survival analysis in nephrology? **Nephrology Dialysis Transplantation**, v.28, p.2670–2677, 2013.

OSELLA, A. R. Epidemiology of hepatitis C virus infection in an area of Southern Italy. Hepatitis C virus (HCV) infection and liver-related mortality: a population-based cohort study in Southern Italy. The association for the study of liver disease in Puglia. **International Journal of Epidemiology**, v.29, n.5, p.922–927, 2000.

OSELLA, A. R.; MISCIAGNA, G.; LEONE, A.; DI-LEO, A.; FIORE, G. Epidemiology of hepatitis C virus infection in an area of Southern Italy. **Journal of Hepatology**, v.27, n.1, p.30–35, 1997.

RASHIDKHANI, B.; JESSRI, M.; HAJIZADEH, B.; J, B.; KREIGER, N.; BAJDIK, C. D. Adherence to dietary recommendations and risk of esophageal squamous cell carcinoma: A case-control study in Iran. **Ann Nutr Metab**, v.59, n.2-4, p.166–175, 2011.

ROYSTON, P.; ALTMAN, D. G. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). **Applied Statistics**, v.43, n.3, p.429–467, 1994.

ROYSTON, P.; ALTMAN, D. G.; SAUERBREI, W. Dichotomizing continuous predictors in multiple regression: A bad idea. **Statistics in Medicine**, v.25, p.127–141, 2006.

ROYSTON, P.; LAMBERT, P. C. **Flexible parametric survival analysis using Stata: beyond the Cox model**. Texas: Stata Press, 2011. 347p.

ROYSTON, P.; PARMAR, M. K. B. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. **Statistics in Medicine**, v.21, n.15, p.2175–2197, 2002.

ROYSTON, P.; SAUERBREI, W. **Multivariable Model - Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modeling Continuous Variables**. John Wiley & Sons. Ltd, 2008. 285p.

STATA. Corp LP (USA), 2017.

WOLBERS, M.; KOLLER, M. T.; STEL, V. S.; SCHAEER, B.; JAGER, K. J.; AAND G HEINZE, K. S. L. Competing risks analyses: objectives and approaches. **European Heart Journal**, v.35, n.42, p.2936–2941, 2014.

ZHANG, Z. Survival analysis in the presence of concurrent risks. **Annals of Translational Medicine**, v.5, n.3, p.47, 2017.

APÊNDICES

Código para estimar as funções de incidência acumulada, para ajustar o modelo de Cox para o risco específico por causa e para ajustar modelo de Fine e Gray para o risco de subdistribuição.

- Tempo denota o tempo até a ocorrência do primeiro evento
- Status é uma variável indicadora do tipo de evento:
- 1: óbito por câncer gastrointestinal
- 0: observação censurada: o indivíduo não morreu por câncer gastrointestinal.

Análise e estimação para o modelo de Cox

stset survival, failure(status)

Estimação modelo completo

```
putdocx begin
```

```
Embed Stata Output
```

```
putdocx paragraph
```

```
putdocx text ("Cox's Model")
```

```
stcox i.trabalho i.estudo i.idade i.BMI_C i.gênero i.IADM i.High_Sugar_Score_50 i.W
```

```
putdocx table mytable0=etable
```

```
putdocx paragraph
```

```
putdocx text ("Cox's Model Statistics")
```

```
estat ic
```

```
matrix A0=r(S)
```

```
putdocx table MatrixA0=matrix(A0)
```

Comando para cálculo da razão de risco

```

stpm2 i.IADM i.sexo High_Sugar_Score Western_Score, scale(hazard) df(2) eform
estimates store rs_hazard2

* Modelo 5 estatisticamente significativo para IADM e escolhido pelo AIC
stpm2 i.IADM i.idade i.sexo i.High_Sugar_Score i.Western_Score, scale(hazard) df(2) eform
estimates store rs_hazard2

estat ic

stpm2 i.IADM i.idade i.gender High_Sugar_Score Western_Score, scale(hazard) df(3) eform
estimates store rs_hazard2

estat ic

stpm2 i.IADM i.idade i.gender High_Sugar_Score Western_Score, scale(hazard) df(4) eform
estimates store rs_hazard2

estat ic

stpm2 i.IADM i.idade i.gender High_Sugar_Score Western_Score, scale(hazard) df(5) eform
estimates store rs_hazard2

estat ic

```

Modelo com a covariável IADM e diferentes df

```

stpm2 IADM i.sexo, scale(hazard) df(2) eform
estat ic

stpm2 IADM i.sexo, scale(hazard) df(3) eform
estat ic

stpm2 IADM i.sexo, scale(hazard) df(4) eform
estat ic

stpm2 IADM i.sexo, scale(hazard) df(5) eform
estat ic

```

Comando para modelo Weibull com AIC

```

streg IADM, distribution(weibull) nolog nohr noheader noshow
estat ic

streg i.sexor, distribution(weibull) nolog nohr noheader noshow

```

```
estat ic
streg idade, distribution(weibull) nolog nohr noheader noshow
estat ic
```

Comandos para razão com intervalos entre as variáveis

```
stpm2 agegroup2-agegroup5, scale(hazard) df(2) bhazard(rate)
tvc(agegroup2-agegroup5) dftvc(3)
estimates store rs_hazard2
stpm2 agegroup2-agegroup5 , scale(hazard) df(7) bhazard(rate)
tvc(agegroup2-agegroup5) dftvc(4)
stpm2 BP, scale(hazard) df(5) bhazard(rate) eform
estimates store rs_hazard5
stpm2 agegroup2-agegroup5 , scale(hazard) df(6) bhazard(rate)
tvc(agegroup2-agegroup5) dftvc(5)
estimates store rs_hazard6
stpm2 agegroup2-agegroup5 , scale(hazard) df(7) eform bhazard(rate)
tvc(agegroup2-agegroup5) dftvc(5)
estimates store rs_hazard7
save "C:\Users\ReinaldaSouza\Dropbox\Reinalda\Workflow Sobrevida\Sobrevida\Posted-Env
```

program setup

Comando para eventos competitivos

```
stset survival, failure(GI_cancer_comp==1,2) scale(12) exit(time 120)
stpm2cr [gi_cancer: IADM Idade sexo Western_Score High_Sugar_Score , scale(hazard)
[other: IADM idade sexo Western_Score High_Sugar_Score , scale(hazard) df(4)
events(GI_cancer_comp) cause (1 2)
cens(0) eform
```