UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Eliane Ribeiro Carmes

ANÁLISE DE DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS

CURITIBA
2015

UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Eliane Ribeiro Carmes

ANÁLISE DE DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS

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

Orientadora: Profa. Dra. Suely Ruiz Giolo

CURITIBA
2015

Afinal, qual seria o valor da paixão pelo saber se ele resultasse apenas num certo conhecimento e não, de algum modo, num desgarramento de si mesmo por parte daquele que conhece? Há momentos na vida em que a questão de saber se se pode pensar de maneira diferente da que se pensa e perceber de maneira diferente da que se enxerga é absolutamente indispensável caso se pretenda, de fato, continuar a pensar e a refletir.

Michel Foucault

Ao Rafael,

a seus filhos,

aos filhos dos seus filhos,

aos filhos dos filhos dos seus filhos,

aos filhos dos ...

AGRADECIMENTOS

Acompanho, aqui e mais uma vez, Don Juan, personagem de Carlos Castaneda, dizendo que tudo é a escolha de um caminho, entre tantos.

A meus pais, Dely e Orvalino, *in memoriam*, que me iniciaram no caminho da vida.

Ao Luso, *in memoriam*, no caminho da maternidade.

À minha primeira professora, Sra. Terezinha Grill Bösel, que me iniciou no caminho das letras e números.

À Karina Brotto Rebuli pelo carinho, apoio e incentivo com os "caminhos" do R.

Aos professores e professoras do Curso de Estatística que nesses quase cinco anos compartilharam seu saber, seus pressupostos, seus métodos e muitas vezes um afeto carinhoso.

Às bibliotecárias e aos bibliotecários da Biblioteca do Setor de Ciência e Tecnologia pelo atendimento sempre eficiente e amigo, com que me honraram também a Sra. Arielza Cruz dos Santos e o Sr. Alcides Nepomuceno do Laboratório de Estatística.

E em especial,

à Professora Dra. Suely Ruiz Giolo que disponibilizou, sem nenhuma restrição, saber, tempo e generosidade e que com uma exigência amiga e gentil ajudou-me a vencer cada obstáculo transformando cada uma de nossas reuniões num momento mágico de ensino e aprendizado.

Relembro cada um de vocês e muitos dos momentos que compartilhamos.

RESUMO

Na pesquisa biomédica, particularmente em câncer e transplante de órgãos, a presença de possíveis múltiplos desfechos é quase rotina. Um contexto comum de riscos competitivos envolve recidiva de doença e morte em remissão, no qual a probabilidade de falha causa-específica, ou curva de incidência acumulada, sumariza corretamente a probabilidade de falha em um cenário de análise de dados com riscos competitivos. Entretanto, em diversas publicações ainda se verifica a aplicação do complemento da estimativa de Kaplan-Meier (1 - KM), para cada causa concorrente, como probabilidade de sobrevida livre de doença causa-específica. Esse procedimento não é adequado porque, em geral, superestima a incidência de uma causa em particular na presença de causas concorrentes. Essa limitação da abordagem clássica motivou esforços no sentido de se modelar diretamente as funções de incidência acumulada. Dentre os modelos propostos para estimar os efeitos das covariáveis na função de incidência acumulada estão o modelo de Fine-Gray, uma extensão do modelo de riscos proporcionais de Cox, e o modelo de Scheike-Zhang-Gerds, flexibilização do modelo de Fine-Gray que não pressupões a proporcionalidade dos riscos. Para ilustrar e comparar as metodologias citadas, analisou-se, neste trabalho, dados de um estudo que contém observações de pacientes portadores de linfoma folicular e que teve como objetivo estimar a probabilidade de recidiva da doença num cenário em que a morte sem recidiva se caracteriza como a causa concorrente de falha. Quanto aos métodos não paramétricos abordados para a análise exploratória dos dados, ficou evidenciado que o complemento da estimativa de Kaplan-Meier, em um contexto de riscos competitivos, tende a superestimar às estimativas de falha causa-específicas. Quanto aos modelos de Fine-Gray e Scheike-Zhang-Gerds, estes apresentaram resultados semelhantes quanto à avaliação do efeito das covariáveis na subdistribuição das falhas associadas à causa 1 e, apesar dos procedimentos utilizados para a avaliação da proporcionalidade dos riscos terem sido discordantes, ambos os modelos se mostraram válidos para a análise dos dados.

Palavras-chave: modelo de Fine-Gray; modelos de regressão; mortalidade; sobrevida.

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

1 INTRODUÇÃO

A análise de sobrevivência, por vezes denominada análise de sobrevida, é um conjunto de técnicas e de métodos estatísticos que possibilitam a análise de dados em que há interesse no tempo decorrido até que as unidades amostrais (indivíduos, equipamentos, etc.) apresentem um evento de interesse, também designado falha ou desfecho, na presença de censuras e de covariáveis.

A variável resposta *T*, contínua e não-negativa, representa o tempo até a falha, ou seja, o tempo decorrido até o evento em estudo. As covariáveis são características das unidades amostrais cujo efeito pode acelerar ou retardar a ocorrência da variável resposta (COLOSIMO; GIOLO, 2006).

A censura contempla a impossibilidade de se observar o tempo em que ocorre o evento de interesse para alguns elementos da amostra, seja por término do período de observação ou devido à ocorrência de algum outro evento - acidental ou controlado.

A censura define, então, a observação parcial da variável resposta devido à perda de seguimento por qualquer motivo que não o evento de interesse em estudo. Pressupõe independência entre os tempos de falha e de censura. Para os casos de censura à direita, a única informação que se tem disponível é que o tempo de falha das unidades amostrais classificadas como censuras é superior ao tempo de seguimento (KLEIN; MOESCHBERGER, 2003). Assim, as observações censuradas contribuem com a informação de que a falha ocorreu após o último tempo *t* observado. Os métodos de análise de sobrevivência se particularizam por incorporar essa informação na análise estatística (COLOSIMO; GIOLO, 2006).

A análise de sobrevivência tem aplicação em diversas áreas. Em ciências sociais para estudar o tempo de permanência no emprego. Na engenharia, de onde advém o uso do termo falha, é utilizada para modelar o tempo até a falha de um componente ou equipamento (CARVALHO, ANDREOZZI, CODEÇO, CAMPOS *et al*., 2011). Na área da saúde, foco deste estudo, tem aplicação relevante ao modelar os tempos até o adoecer, a cura, recaída, morte ou outro evento de interesse após a exposição a um fator de risco, o diagnóstico de doença ou após intervenções terapêuticas. Nessa área, destacam-se os estudos para estimar o tempo de sobrevida de pacientes submetidos a transplante de órgãos ou a quimioterapia.

Em estudos que envolvem dados de sobrevivência, é frequente que a causa que conduz à falha seja uma entre *k* possíveis causas. A causa de falha registrada para cada indivíduo será, então, aquela que ocorrer primeiro, dado que a ocorrência de uma delas impede que se observe a ocorrência de qualquer outra.

Estudos delineados para estimar a probabilidade total de falha, independentemente da causa, têm como objeto de interesse um desfecho combinado. Nesses casos, a análise do tempo até o desfecho, que certamente ocorrerá se o tempo de observação for suficientemente longo, pode ser realizada por meio de técnicas usuais de análise de sobrevivência. Contudo, nos estudos em que o objetivo é estimar a probabilidade de falha associada a uma das *k* possíveis causas, o fato de o evento associado a outros tipos de causa, que não a de interesse, não ser uma observação incompleta impede que este seja tratado como censura. Essas situações caracterizam dados de sobrevivência com estrutura de riscos competitivos, já considerados por David Bernoulli, em 1760, ao estudar o risco de morte associado à varíola e a outras causas (KLEIN; MOESCHBERGER, 2003; PINTILIE, 2006).

Na presença de riscos competitivos tem-se, então, que:

a) os indivíduos estão sob risco de falha por *k* causas diferentes e;
b) a ocorrência de uma das *k* causas impede que se observe a ocorrência de qualquer outra.

Os termos risco e causa se referem à mesma condição, embora difiram pela posição no tempo em relação ao desfecho de interesse. Antes do desfecho a condição é denominada risco. Será considerada causa quando for o motivo do desfecho de interesse em estudo (SANTOS; ORTIZ; YAZAKI, 1984).

O objeto de estudo de dados de sobrevivência na presença de riscos competitivos é a distribuição do tempo até a falha para uma causa específica *k (k* = 1, ..., *m)* na presença de todas as outras possíveis causas. Diversas abordagens foram propostas para a análise de dados dessa natureza.

A abordagem clássica, na presença de covariáveis, consiste em modelar a função taxa de falha causa-específica para as diferentes causas de falha sob a suposição de riscos proporcionais (LARSON, 1984; PRENTICE *et al.*, 1978). Entretanto, alguns pesquisadores (GRAY, 1988; PEPE, 1991) observaram que, para uma particular causa de falha, uma determinada covariável pode apresentar efeitos diferentes sobre a função risco ou função taxa de falha causa-específica e a sua correspondente função de incidência acumulada (FIA), também denominada subdistribuição, definida para a causa k por $F_k(t|\boldsymbol{x}) = P(T \leq t, K = k|\boldsymbol{x})$. Desse modo, concluíram ser impossível testar, sob a formulação de

funções taxa de falha causa-específica, o efeito de covariáveis sobre a função de incidência acumulada (FIA).

Essa limitação da abordagem clássica motivou esforços no sentido de se modelar diretamente as funções de incidência acumulada. Desses esforços resultou o modelo de regressão de Fine-Gray (FINE; GRAY, 1999) que, por semelhança com o modelo de regressão de Cox, também é flexível e apresenta muitas propriedades úteis. A popularidade do modelo de Fine-Gray se deve, em parte, por estar implementado nos *softwares* R, pacote cmprsk (R CORE TEAM, 2015), Stata (STATACORP) e SAS (SAS INSTITUTE) e por fornecer, na prática, predições úteis e interpretações relativamente simples.

Diversas extensões surgiram após a apresentação do modelo de Fine-Gray. Uma delas, proposta por Scheike, Zhang e Gerds (2008), acomoda situações em que não se faz necessário supor a proporcionalidade de riscos, o que implica ser possível acomodar no modelo covariáveis com efeito variando no tempo (efeitos tempo-dependentes). Têm-se, ainda, os modelos que permitem a inclusão de covariáveis tempo-dependentes (BEYERSMANN; SCHUMACHER, 2008) e os que permitem a inclusão de um efeito aleatório ou de um termo de fragilidade (KATSAHIAN *et al*, 2006, SCHEIKE; SUN; ZHANG; JENSEN 2010, KATSAHIAN; BOUDREAU, 2011, DIXON; DARLINGTON; DESMOND, 2011).

Este trabalho foi delineado com o objetivo de estudar e ilustrar com aplicações alguns modelos de regressão que, no contexto de dados de sobrevivência com riscos competitivos, modelam diretamente as funções de incidência acumulada. Em particular, o modelo de regressão proposto por Fine e Gray (1999) e uma de suas extensões proposta por Scheike, Zhang e Gerds (2008), foram alvos desse estudo. Os dois modelos foram aplicados a um conjunto de dados com informações sobre pacientes com linfoma folicular.

2 REVISÃO DE LITERATURA

2.1 Conceitos Básicos de Análise de Sobrevivência

A análise de sobrevivência compreende um conjunto de métodos utilizados para a análise de dados em que a variável dependente (ou variável resposta) é o tempo *T*, contado a partir de uma determinada situação bem definida, até a ocorrência de um evento ou mudança de estado, previamente especificado. Evento este também denominado falha ou desfecho (COLLETT, 2003; COLOSIMO; GIOLO, 2006). Além da variável resposta, os dados usualmente apresentam informações sobre as censuras, assim como sobre um conjunto de covariáveis. Um dos objetos de estudo da análise de sobrevivência é a distribuição do tempo até a falha, $F(t|\,\boldsymbol{x}) = P(T \leq t\,|\boldsymbol{x})$, que denota a probabilidade de um indivíduo sob risco, e com vetor de covariáveis $\boldsymbol{x}$, apresentar o desfecho até o tempo *t*.

Neste texto, os termos tempo de falha e tempo até o evento são utilizados, de acordo com o contexto, para descrever o tempo decorrido até a observação do desfecho de interesse, ou seja, da falha.

As metodologias estatísticas usualmente aplicadas para a análise de dados de sobrevivência se fundamentam nos estudos de Edward L. Kaplan e Paul Meier, publicado em 1958, de Sir David Roxbee Cox e de Wayne B. Nelson, publicados em 1972, e no de Odd Olai Aalen, em 1978 (KAPLAN; MEIER, 1958; COX, 1972; NELSON, 1972; AALEN, 1978; PÉREZ-MARÍN, 2008). Em todas elas, as informações provenientes das censuras são incorporadas nos procedimentos de análise.

Existem diferentes mecanismos de censura. Censura à esquerda ocorre quando o tempo de sobrevida é menor que o observado, ou seja, o evento de interesse aconteceu antes de o indivíduo entrar no estudo. A censura é dita intervalar quando a informação disponível é a de que o evento ocorreu em um intervalo de tempo conhecido. Na censura à direita, a mais frequentemente observada, o tempo de sobrevida é maior que o observado (COLLETT, 2003; COLOSIMO; GIOLO, 2006).

Em um estudo de sobrevivência, os dados observados para o *i*-ésimo indivíduo, para $i = 1, \ldots, n$, são representados pelo par (t_i, δ_i) ou pela tripla $(t_i, \delta_i, \mathbf{x}_i)$, em que t_i é o tempo de falha ou de censura associado ao indivíduo *i*; δ_i é a variável indicadora de falha ($\delta_i = 1$, se falha e $\delta_i = 0$, se censura) e; $\mathbf{x}_i$ é o vetor de covariáveis do indivíduo *i*.

Outro objeto dos estudos de sobrevivência, a probabilidade de sobreviver ao tempo *t*, isto é, o evento de interesse não ter sido observado até o instante *t*, pode ser estimada por meio da função de sobrevivência assim definida:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t), \quad t \geq 0,$$

em que T denota a variável aleatória contínua e não-negativa que representa o tempo decorrido até o evento de interesse; $S(t)$ é uma função monótona não crescente tal que $S(0) = 1$ e $\lim_{t\to\infty} S(t) = 0$. A representação gráfica de $S(t)$ é chamada de curva de sobrevivência.

A função de distribuição acumulada, $F(t)$, fornece a probabilidade de se observar o evento de interesse no intervalo [0, *t*] e é definida por:

$$F(t) = \int_0^t f(u)du, \quad t \geq 0.$$

Na expressão acima, $f(t)$ é a função densidade de probabilidade da variável *T*, que fornece a probabilidade incondicional de que o desfecho ocorra no tempo *t*, isto é,

$$f(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}, \quad t \geq 0.$$

Outras duas funções de interesse em análise de sobrevivência são: a função taxa de falha (ou função risco ou força de mortalidade), que expressa a taxa instantânea de falha no instante *t*, condicional à sobrevivência até esse instante *t*, ou seja,

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

e a função taxa de falha acumulada definida por

$$\Lambda(t) = \int_0^t \lambda(u)du, \quad t \geq 0,$$

função monótona e não decrescente que mede o risco ou taxa de ocorrência do evento de interesse até o tempo *t*, dado que $\lambda(t) \geq 0$ e $\int_0^\infty \lambda(u)du = \infty$.

Tem-se, ainda, que para tempos de vida contínuos com função densidade de probabilidade $f(t)$, a relação entre $S(t)$ e $\lambda(t)$ pode ser expressa por

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\partial \ln[S(t)]}{\partial t}.$$

Para estimar as funções mencionadas, métodos não paramétricos e modelos de regressão foram propostos. Dentre eles, o estimador de Kaplan-Meier (KAPLAN; MEIER,

1958), o de Nelson-Aalen (AALEN, 1978) e o modelo de Cox (COX, 1972), descritos brevemente a seguir.

2.1 Estimador de Kaplan-Meier

Apresentado à comunidade científica em 1958, o estimador não paramétrico proposto por Kaplan e Meier, estima a função de sobrevivência, $S(t)$, na presença de dados censurados. É uma versão atualizada das tábuas de vida clássicas da ciência atuarial desenvolvidas por John Graunt, em 1662, e pelo conhecido astrônomo Edmond Halley em 1693 (PINTILIE, 2006; SELVIN, 2004).

O estimador de Kaplan-Meier (KM), ou produto-limite, é um produto em que cada fator é obtido particionando-se o eixo do tempo em pequenos intervalos determinados pela ocorrência de uma falha e estimando a probabilidade de sobrevida em cada intervalo, condicional à proporção de sobreviventes. O produto dessas estimativas resulta na probabilidade incondicional de sobrevida ou sobrevida geral (KAPLAN; MEIER, 1958). Formalmente, o estimador de KM é dado por

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

em que $t_1 < t_2 < \cdots < t_m$ são os *m* tempos distintos e ordenados de falha; d_j é o número de falhas em $t_j,\ j = 1, 2, \ldots, m$ e n_j é o número de indivíduos sob risco, ou seja, livres de desfecho e não censurados até o tempo imediatamente anterior ao tempo t_j.

A partir do complemento do estimador de KM, isto é, $\hat{F}(t) = 1 - \hat{S}(t)$, tem-se a incidência acumulada de falha até o tempo *t*. $F(.)$ é uma função unidimensional do tempo, com domínio $t = [0, \infty)$ e limites entre 0 e 1 $(0 \leq F \leq 1)$ (COLE, HUDGENS, BROOKHART, WESTREICH, 2015).

Um intervalo de confiança para $S(t)$, com 100(1− α)% de confiança pode ser obtido por:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}\left[\hat{S}(t)\right]}$$

com a variância assintótica de $\hat{S}(t)$ estimada por meio da fórmula de Greenwood a seguir:

$$\widehat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)}.$$

Como o intervalo de confiança apresentado pode, em algumas situações, resultar em valores negativos ou maiores que 1, transformações foram propostas para $\hat{S}(t)$ a fim de o intervalo de confiança ficar restrito ao intervalo [0,1]. Dentre essas transformações tem-se:

$$\widehat{U}(t) = \log\left[-\log\left(\hat{S}(t)\right)\right],$$

de modo que o intervalo de confiança de 100(1− α)% para $S(t)$ resulta em:

$$\text{IC}(S(t)) = \hat{S}(t)^{\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}[\widehat{U}(t)]}\right\}}$$

sendo

$$\widehat{\text{Var}}\left[\widehat{U}(t)\right] = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}}{\left[\log\left(\widehat{S}(t)\right)\right]^2}.$$

O estimador de KM permite incorporar informações provenientes de censuras em seu procedimento de estimação, desde que haja independência entre os tempos de falha e de censura (KAPLAN; MEIER, 1958).

O estimador de KM é amplamente utilizado para estimar a função de sobrevida e a função de incidência acumulada, seu complemento. Pressupõe uma única causa de falha de interesse, que ocorre com probabilidade igual a 1 (um), se o tempo de seguimento for suficientemente longo, e trata como censura os tempos observados para aqueles que não apresentam o evento de interesse até o final do tempo de seguimento (KAPLAN; MEIER, 1958).

2.2 Estimador de Nelson-Aalen

Este estimador foi inicialmente proposto por Nelson (1972) e retomado por Aalen (1978), daí ser denominado estimador de Nelson-Aalen. Permite estimar a função taxa de falha ou de risco acumulado, $\Lambda(t)$, sendo o estimador para esta função dado por:

$$\widetilde{\Lambda}(t) = \sum_{j:t_j<t} \frac{d_j}{n_j}$$

com d_j e n_j definidos do mesmo modo que no estimador de KM.

Em consequência, o estimador para $S(t)$ resulta em:

$$\tilde{S}(t) = \exp\{-\widetilde{\Lambda}(t)\},$$

sendo a variância assintótica de $\tilde{S}(t)$ dada por:

$$\widehat{Var}\left[\tilde{S}(t)\right] = \left[\tilde{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{\left(n_j\right)^2}.$$

Da mesma maneira que para o estimador de KM, recomenda-se a utilização de transformações, tal como $\tilde{U}(t) = \log\left[-\log\left(\tilde{S}(t)\right)\right]$, a fim de se obter estimativas intervalares para $S(t)$ restritas ao intervalo [0,1].

2.3 Modelo de Regressão de Cox

O modelo de regressão de Cox foi proposto em 1972 (COX, 1972) e, desde então, vem sendo amplamente utilizado para analisar dados de sobrevivência na presença de um conjunto de covariáveis. Este modelo assume que o risco ou taxa de falha se relaciona com as covariáveis da seguinte forma:

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}),$$

sendo $\lambda_0(t)$ uma função taxa de risco basal não negativa a qual não é assumida nenhuma forma paramétrica; $\boldsymbol{\beta}$ um vetor de parâmetros de regressão desconhecidos e $\boldsymbol{x}$ um vetor de covariáveis.

O modelo de Cox assume que os efeitos das covariáveis não variam com o tempo e, devido a isso, é referenciado como sendo de riscos proporcionais. A estimação do vetor de parâmetros $\boldsymbol{\beta}$ é feita por meio da maximização da função de verossimilhança parcial (COX, 1975) e, a da função $\lambda_0(t)$, por meio do estimador não paramétrico de Breslow (COX, 1972; HANLEY, 2008). Diversos métodos gráficos, assim como testes, encontram-se disponíveis na literatura para verificar tanto a suposição de proporcionalidade dos riscos, quanto a qualidade de ajuste do modelo. Estes se baseiam, em geral, nos resíduos de Cox-Snell, Schoenfeld, *martingale* e *deviance* (COLOSIMO; GIOLO, 2006).

As metodologias descritas neste capítulo, foram propostas para a análise de dados de sobrevivência em que o evento ocorre devido a uma causa particular de falha. Contudo, existem diversas situações em que o evento pode ocorrer por uma dentre *k* causas (*k* = 1, ..., *m)*. Na área da saúde, no contexto particular de transplantes de medula óssea em pacientes com leucemia, as falhas de tratamento podem ser devidas a dois tipos de causas: recidiva da leucemia ou óbito livre de doença, este último, na maioria das vezes, associado a complicações do transplante.

Desse modo, há interesse em métodos e modelos que levem em conta na análise a presença de mais de uma causa de falha, uma vez que o objetivo é estimar a probabilidade de falha pela causa *k*, para $k = 1, ..., m$. Dados provenientes de situações como as citadas são referenciados na literatura como dados de sobrevivência com estrutura de riscos competitivos.

Dentre as metodologias estatísticas propostas para a análise de dados com a estrutura mencionada (em geral extensões das que foram apresentadas neste capítulo), encontram-se descritas no Capítulo 3 as que foram alvo de estudo neste trabalho.

3 CASUÍSTICA E MÉTODOS

3.1 Casuística

3.1.1 Conjunto de Dados

O conjunto de dados utilizado para ajustar os dois modelos foco deste trabalho, o de Fine-Gray e a extensão proposta por Scheike, Zhang e Gerds, encontra-se disponível em http://www.uhnres.utoronto.ca/labs/hill/datasets/Pintilie/datasets/follic.txt. O conjunto contém informações de 541 pacientes com linfoma de células foliculares (tipo I ou II) em estádio precoce, tratados com radioterapia ou radioterapia e quimioterapia. A idade dos pacientes (média = 57 anos e desvio padrão = 14) e os níveis de hemoglobina (média = 138 g/l e desvio padrão = 15) também estão disponíveis. O tempo mediano de seguimento dos pacientes foi de 5,46 anos (FINE; GRAY, 1999; SCHEIKE; ZHANG; GERDS, 2008).

No contexto de riscos competitivos, as duas causas de falha consideradas no estudo foram: (1) recidiva da doença ou ausência de resposta ao tratamento – causa de interesse e; (2) morte em remissão – causa concorrente. A variável resposta foi definida como o tempo, em anos, decorrido desde o início do tratamento até a falha devido à causa que ocorrer primeiro. Para os pacientes sem registro de falha (censura), o tempo considerado foi o decorrido desde o início do tratamento até a data final de seguimento de cada um deles (PINTILIE, 2006).

Dos 541 pacientes no estudo, foram observadas 348 falhas, sendo 272 associadas à causa 1 (24 em pacientes que não responderam ao tratamento e 248 em pacientes em recidiva de doença) e 76 à causa 2 (óbitos em pacientes livres de doença, isto é, em remissão). O Quadro 1 apresenta uma descrição geral das informações disponíveis para o conjunto de dados descrito.

3.1.2 Recursos Computacionais

O *software* R, versão 3.1.1 (R CORE TEAM, 2014) foi utilizado para ajustar os modelos aos dados descritos por meio dos pacotes `cmprsk` (GRAY, 2014) e `timereg` (SCHEIKE; MARTINUSSEN, 2006; SCHEIKE; ZHANG, 2011).

Quadro 1 – Descrição das variáveis disponíveis no conjunto de dados de pacientes com linfoma de células foliculares

VARIÁVEL	DESCRIÇÃO
Stnum	identificação do paciente
COVARIÁVEIS	
Age	idade em anos
Hgb	hemoglobina em g/l (gramas por litro)
clinstg	estádio clínico: 1 = estádio I 2 = estádio II
Ch	quimioterapia: Y = sim em branco = não
RT	radioterapia: Y = sim em branco = não
DESFECHOS	
Resp	resposta após tratamento: CR = remissão completa NR = sem resposta
Relsite	sítio da recidiva: L = local D = distante B = local e distante em branco = sem recidiva
Survtime	tempo decorrido desde o diagnóstico até o óbito ou até o último seguimento (em anos)
St	situação: 1 = óbito 0 = vivo
Dftime	tempo decorrido desde o diagnóstico até a falha que ocorrer primeiro: sem resposta, recidiva ou óbito
Dfcens	censura: 1 = falha 0 = censura

Fonte: Pintilie, M. Competing risks: a practical perspective. Chichester: John Wiley & Sons, Ltd, 2006, p. 17.

3.2 Métodos

Em situações de análise de dados com riscos competitivos, há interesse em estimar e a probabilidade marginal de falha pela causa *k*. Nesse cenário, as funções propostas para essa finalidade revelam a falta de consenso quanto à nomenclatura. Função de incidência acumulada e subdistribuição são as denominações mais frequentes na literatura (PINTILIE, 2006). Outras são: força de mortalidade, função de risco, função de probabilidade marginal (FINE; GRAY, 1999) ou, ainda, força de transição considerando-se cada causa de falha como um processo markoviano absortivo.

Na presença de riscos competitivos, as funções identificáveis de maior relevância são: i) a função taxa instantânea de falha ou função risco causa-específica, $\lambda_k(t)$ e, ii) a função de incidência acumulada associada à causa *k*, $F_k(t), k = 1, \dots, m$.

A função risco causa-específica, $\lambda_k(t)$, fornece a probabilidade de falha associada à causa k no instante de tempo *t*, condicional a não ocorrência de qualquer outro tipo de falha entre os indivíduos sob risco até o tempo *t*, enquanto a função $F_k(t) = P\ (T \leq t, K = k)$, fornece, em cada tempo *t*, a probabilidade acumulada de falha devido à *k-ésima* causa, na presença de todas as demais.

Em um cenário de riscos competitivos se observa para cada indivíduo i, $i = 1, \dots, n.$, o tempo em estudo t_i, a variável indicadora de falha δ_i e, para aqueles que apresentaram o evento de interesse, ou seja $\delta_i = 1$, registram-se a causa de falha k.

Quando um conjunto de causas de falha, denotado por $k = \{1, 2, \dots, m\}$, age simultaneamente sobre um indivíduo, tem-se associado à cada causa: uma função risco causa-específica, $\lambda_k(t)$, uma função de incidência acumulada, $F_k(t)$, e uma função de sobrevivência causa-específica, $S_k(t) = P\ (T > t, K = k)$. Em sendo razoável a suposição de independência entre as k causas, tem-se que a soma das k funções de incidência acumulada resulta na função de incidência acumulada total (PRENTICE, 1979; COLE; HUDGENS; BROOKHART; WESTREICH, 2015), isto é,

$$F_1(t) + F_2(t) + \cdots + F_m(t) = \sum_{k=1}^{m} F_k(t) = F(t) = P(T \leq t),$$

de modo que $F_k(t)$ não necessariamente varia no intervalo [0,1].

As funções $F_k(t)$ podem, ainda, ser escritas em termos das funções taxa de falha causa-específica $\lambda_k(t)$ (SCRUCCA; SANTUCCI; AVERSA, 2007; COLE; HUDGENS; BROOKHART; WESTREICH, 2015), de modo que:

$$F_k(t) = \int_0^t \lambda_k(u)\, S(u)\, du, \quad k = 1, \dots, m,$$

em que $S(t)$, a função de sobrevivência geral, é explicitada por:

$$S(t) = \exp\left\{-\int_0^t \sum_{k=1}^{m} \lambda_k(u) du\right\}$$

$$= \exp\left\{-\int_0^t [\lambda_1(u) + \lambda_2(u) + \cdots + \lambda_m(u)] du\right\} = 1 - F(t).$$

Assim, observa- se que a subdistribuição de um evento devido à causa k depende das funções taxa de falha causa-específica associadas às causas concorrentes por meio da função de sobrevida geral.

3.2.1 Métodos Não Paramétricos no Contexto de Riscos Competitivos

3.2.1.1 Kaplan-Meier e suas Limitações neste Contexto

Considere um estudo de sobrevivência em que há duas possíveis causas de falha: recidiva e óbito. Uma abordagem natural para a análise dos dados desse estudo seria utilizar o estimador de Kaplan-Meier separadamente para os dois tipos de falha. Assim, quando o interesse estivesse no tempo até a recidiva, os óbitos ocorridos antes da recidiva contribuiriam como censura. De modo análogo, quando o interesse estivesse no tempo até o óbito, as recidivas contribuiriam como censura. Contudo, ao proceder desse modo, é comum, nesses casos, que se utilize erroneamente o complemento da função de KM, $1 - S_k(t)$, de interpretação clínica direta, para estimar a probabilidade de recidiva no tempo *t*, como se fosse impossível o óbito para os pacientes em remissão. Ou então, para estimar a mortalidade numa situação em que não existisse a possibilidade de recidiva da doença (KLEIN; MOESCHBERGER, 2003).

Essa abordagem assume que a distribuição da falha pela causa de interesse seja igualmente provável entre os indivíduos censurados e os que continuam em risco, ou seja, o desfecho de interesse é passível de ocorrer entre os censurados, mas não pode ser observado. Suposição essa que é violada na presença de outra(s) causa(s) que concorra(m) com o evento em análise.

Logo, ao considerar todos os riscos concorrentes como censura, as estimativas de KM se tornam viesadas na presença de várias possíveis causas competindo com o evento de interesse, o que motivou a proposição de estimadores alternativos (PINTILIE, 2006).

3.2.1.2 Método Alternativo ao de Kaplan-Meier

Tendo em vista a limitação do estimador de Kaplan-Meier, um estimador não paramétrico alternativo, o qual estima diretamente a FIA causa-específica, é dado por (KLEIN; MOESCHBERGER, 2003; PINTILIE, 2006; SCRUCCA; SANTUCCI; AVERSA, 2007):

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

em que d_{kj} é o número de falhas no tempo t_j associado à causa k, n_j é o número de indivíduos em risco no tempo t_j e $S(t)$ é a função de sobrevivência geral.

Intervalos de confiança para $F_k(t)$ podem ser obtidos fazendo-se uso da transformação log(-log) (SCRUCCA; SANTUCCI; AVERSA, 2007), de modo que o intervalo de confiança $(1-\alpha)100\%$ para $F_k(t)$ fica definido por:

$$\hat{F}_k(t)^{\exp\left\{\frac{\pm z\alpha/_2 \hat{\sigma}_k(t)}{\hat{F}_k(t)\, log[\hat{F}_k(t)]}\right\}}$$

em que $z\alpha/_2$ é o percentil superior $\alpha/_2$ da distribuição normal padrão e $\hat{\sigma}_k(t)$ é raiz quadrada da variância estimada para $\hat{F}_k(t)$ obtida de:

$$\widehat{Var}\left[\hat{F}_k(t)\right] = \sum_{j:t_j \le t} \left[\hat{F}_k(t) - \hat{F}_k(t_j)\right]^2 \frac{d_j}{n_j(n_j - d_j)}$$

$$+ \sum_{j:t_j \le t} \hat{S}(t_{j-1})^2 \frac{d_{kj}(n_j - d_{kj})}{n_j^3}$$

$$- 2 \sum_{j:t_j \le t} \left[\hat{F}_k(t) - \hat{F}_k(t_j)\right] \hat{S}(t_{j-1}) \frac{d_{kj}}{n_j^2}$$

sendo $d_j = \sum_{k=1}^{m} d_{kj}$.

3.2.1.2.1 Testes para a Comparação de Funções de Incidência Acumulada

Em 1988, Robert Gray propôs um teste, conhecido como teste de Gray, que permite comparar as funções de incidência causa-específica entre s categorias ($s \geq 2$) de uma dada covariável. Para a causa k, a hipótese nula associada a esse teste é dada por H_0: $F_{k1} = F_{k2} = \cdots = F_{kS}$, $s \geq 2$.

O teste faz uso da média ponderada do risco das funções de subdistribuição do evento de interesse (GRAY, 1988), especificado por:

$$\gamma(t) = \frac{f(t)}{1 - F(t)}.$$

O escore para o *s-ésimo* grupo (categoria da covariável) é definido por:

$$z_s = \int_0^{\tau} W_s(t) \{\gamma_i(t) - \gamma_0(t)\} dt,$$

em que τ é o tempo máximo observado nos grupos; $W_s(t)$ é a função de pesos; $\gamma_s(t)$ é o risco da subdistribuição para o *s-ésimo* grupo e; $\gamma_0(t)$ é o risco da subdistribuição para todos os grupos

Após algumas transformações (para detalhes consultar PINTILIE, 2006, pg. 75) o escore, considerando dois grupos ($s = 2$), pode ser assim reescrito:

$$Z_1 = \sum_{\text{all } t_j} R_{1j}\left(\frac{d_{1j}}{R_{1j}} - \frac{d_{1j} + d_{2j}}{R_{1j} + R_{2j}}\right),$$

sendo

$$R_{1j} = n_{1j}\frac{1 - \hat{F}(t_{j-1})}{\hat{S}(t_{j-1})},$$

em que d_{1j} é o número de eventos de interesse no grupo 1 no tempo t_j e n_{1j} é o número de eventos de interesse no grupo 1 no tempo t_j;

O escore Z_1 é somente o numerador da estatística do teste de Gray, que tem por denominador a variância de Z_1, de formulação extensa (para detalhes consultar o Apêndice A em PINTILIE, 2006 e GRAY, 1988). Para *s* grupos a referida estatística segue distribuição qui-quadrado com $s - 1$ graus de liberdade, denotada por χ^2_{s-1}. O teste de Gray está implementado no pacote `cmprsk` do R, função `cuminc` (R CORE TEAM, 2014).

Em 1993, Pepe e Mori também propuseram o teste de Pepe-Mori, que permite comparar diretamente duas funções de distribuição acumulada. O teste está implementado na função `compCIF`, disponível em http://www.uhnres.utoronto.ca/labs/hill/datasets/Pintilie/Rfunctions/compCIF.txt e em Pintilie (2006). Basicamente, o teste faz uma ponderação da área entre as duas curvas sendo comparadas (PEPE; MORI, 1993).

3.2.2 Modelos de Regressão no Contexto de Riscos Competitivos

Para analisar dados de sobrevivência com riscos competitivos, diversos modelos de regressão, que modelam diretamente as funções de incidência acumulada na presença de covariáveis, foram propostos na literatura. Dois deles são apresentados a seguir.

3.2.2.1 Modelo de Fine-Gray

O modelo de Fine-Gray (FINE; GRAY, 1999) avalia o efeito das covariáveis diretamente sobre a subdistribuição (ou função de incidência acumulada) de uma particular causa de falha *k* em um ambiente de riscos competitivos. Sob este modelo, a função de incidência acumulada para a causa *k* é modelada por:

$$F_k(t|\boldsymbol{x}) = P(T \leq t, K = k \mid \boldsymbol{x}) = 1 - \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}_k)\}, \qquad (1)$$

em que T denota a variável tempo, ϵ indica a causa de falha, $\boldsymbol{x} = (x_1, ..., x_p)$ corresponde a um vetor de covariáveis, $\boldsymbol{\beta}_k$ a um vetor de coeficientes de regressão associados à causa k e $\Lambda_0(t)$ a uma função não especificada e não decrescente tal que $\Lambda_0(0) = 0$.

Nota-se que $F_k(t \mid \boldsymbol{x})$ fornece a probabilidade de um indivíduo não sobreviver ao tempo *t* devido à *k*-ésima causa na presença de covariáveis, isto é, $P(T \leq t, K = k \mid \boldsymbol{x})$. Ainda, tem-se que a função que lineariza o modelo (1), denominada função de ligação, é a complemento log-log (cloglog), isto é,

$$\log\Big(-\log\big(1 - \mathrm{F_k(t|}\boldsymbol{x})\big)\Big) = \boldsymbol{x}^T\boldsymbol{\beta}_k + \log\big(\Lambda_0(t)\big).$$

Assim, denotando a função de ligação cloglog por $h(\cdot)$, tem-se o modelo de Fine-Gray expresso por:

$$\begin{aligned} h(1 - F_k(t|\boldsymbol{x})) &= \boldsymbol{x}^T\boldsymbol{\beta}_k + \log\big(\Lambda_0(t)\big) \\ &= \boldsymbol{x}^T\boldsymbol{\beta}_k + \eta\big(\Lambda_0(t)\big), \end{aligned}$$

com $\eta(\cdot)$ uma função de $\Lambda_0(t)$.

Se o principal interesse está em avaliar os efeitos das covariáveis sobre a função de incidência acumulada, outras funções de ligação conhecidas (por exemplo, a logito) podem ser consideradas a fim de tornar as interpretações mais simples. A função de incidência acumulada pode então, de modo geral, ser expressa como uma função $g(.)$ de $\Lambda_0(t)$, $\boldsymbol{\beta}_k$ e $\boldsymbol{x}$, isto é,

$$F_k(t \mid \boldsymbol{x}) = g(\Lambda_0(t), \boldsymbol{\beta}_k, \boldsymbol{x}),$$

de modo que:

$$h(1 - F_k(t \mid \boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}_k + \eta\big(\Lambda_0(t)\big),$$

com $h(\cdot)$ a função de ligação considerada e $\eta(\cdot)$ uma função de $\Lambda_0(t)$.

Na presença de um vetor de covariáveis $\boldsymbol{x}$, o modelo de Fine-Gray assume riscos proporcionais, ou seja, efeito constante das covariáveis ao longo do tempo. Para averiguar essa suposição, ou seja, testar a hipótese: $H_0: \beta_k(t) = \beta_k$, foi proposto a inclusão no modelo de interações entre as covariáveis e funções do tempo. Sendo significativos os efeitos de tais interações, a suposição estaria sendo violada, havendo a necessidade de considerar modelos alternativos ao de Fine-Gray para a análise dos dados como, por exemplo, o modelo de Scheike-Zhang-Gerds apresentado adiante.

3.2.1.1 Modelo de Scheike, Zhang e Gerds

Scheike, Zhang e Gerds (2008) propuseram uma extensão do modelo de Fine-Gray (1999), e a implementaram no pacote `timereg` do R (R CORE TEAM, 2014)

Tal modelo é mais flexível que o modelo (1), tendo em vista que para seu uso não é necessário assumir proporcionalidade dos riscos, ou seja, pode-se acomodar nesse modelo covariáveis com efeito tempo-dependentes. De modo geral, o modelo é dado por:

$$F_k(t \mid \boldsymbol{x}, \boldsymbol{z}) = 1 - \exp\left\{-\Lambda_0(t) \exp\left\{\boldsymbol{x}^T \boldsymbol{\beta}_k + \boldsymbol{z}^T \boldsymbol{\alpha_k}(t)\right\}\right\},$$

de modo que:

$$h\{1 - F_k(t \mid \boldsymbol{x}, \boldsymbol{z})\} = \boldsymbol{x}^T \boldsymbol{\beta}_k + \boldsymbol{z}^T \boldsymbol{\alpha_k}(t) + \eta\big(\Lambda_0(t)\big),$$

com $h(\cdot)$ uma função de ligação conhecida, $\boldsymbol{x}$ e $\boldsymbol{z}$ vetores de covariáveis, $\boldsymbol{\beta}_k$ coeficientes de regressão que não variam no tempo, $\boldsymbol{\alpha_k}(t)$ coeficientes de regressão que variam no tempo, ambos associados à causa k, e $\eta(\cdot)$ uma função de $\Lambda_0(t)$.

A estimação dos coeficientes de regressão associados aos modelos apresentados pode ser realizada por meio de uma abordagem baseada no ajuste direto de um modelo de regressão binomial (SCHEIKE; ZHANG, 2008).

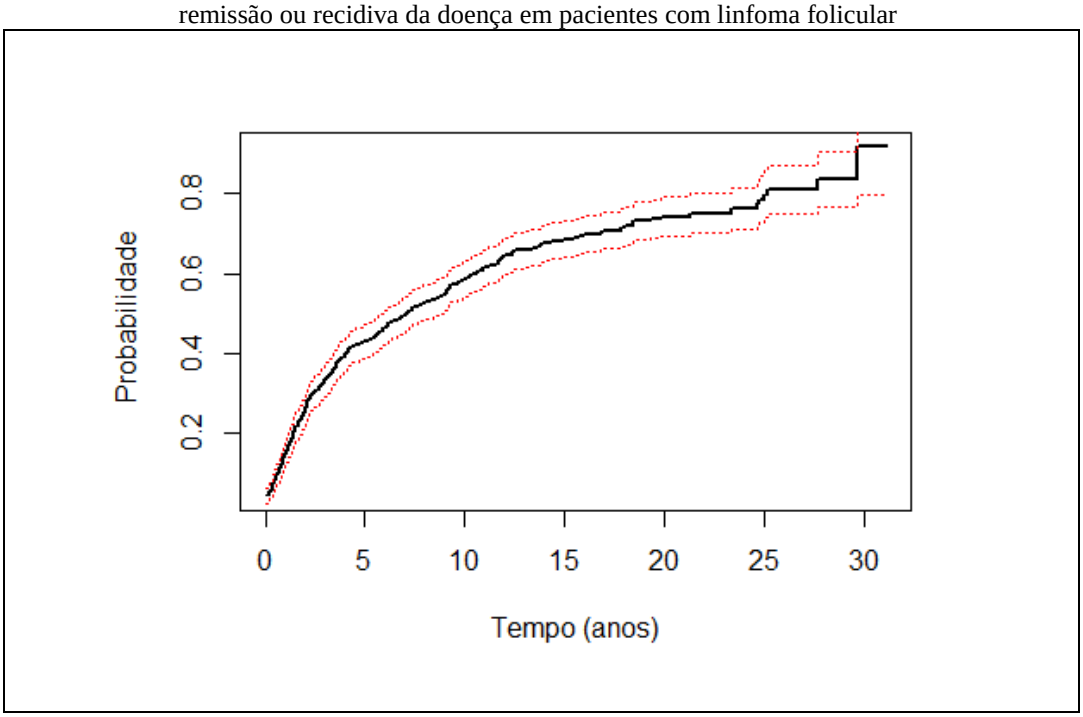
4 RESULTADOS E DISCUSSÃO

Para os dados descritos na Seção 3.1.1, o tempo mediano de seguimento para os 541 pacientes em estudo foi de 5,46 anos. Os dados observados para o *i-ésimo* indivíduo foram a quádrupla $(t_i, \delta_i, k_i, \mathbf{x}_i)$ em que $t \in [0;\ 31{,}1]$ representa o tempo observado, em anos, até a ocorrência de falha ou censura, $\delta = I$ (falha), $k = 1 \wedge 2$ indica a causa de falha e $\mathbf{x} = (x_1,\ x_2,\ x_3,\ x_4)$ é o vetor de covariáveis no qual $x_1 \in [17; 86]$ se refere à idade (em anos), $x_2 \in [40; 189]$ é a dosagem de hemoglobina sérica em g/l (gramas por litro), $x_3 = I$ (estádio clínico I) e $x_4 = I$ (quimioterapia), $i = 1, 2, \ldots, 541$.

4.1 Resultados dos Métodos não Paramétricos

Numa abordagem inicial, utilizou-se o complemento do estimador de Kaplan-Meier, $\hat{F}(t) = 1 - \hat{S}_{KM}(t)$, para estimar a probabilidade de falha considerando o desfecho composto: recidiva da doença ou óbito sem recidiva. Nessa situação, há independência entre os tempos de falha e de censura, sendo as estimativas obtidas para $F(t)$ não viesadas (Figura 1).

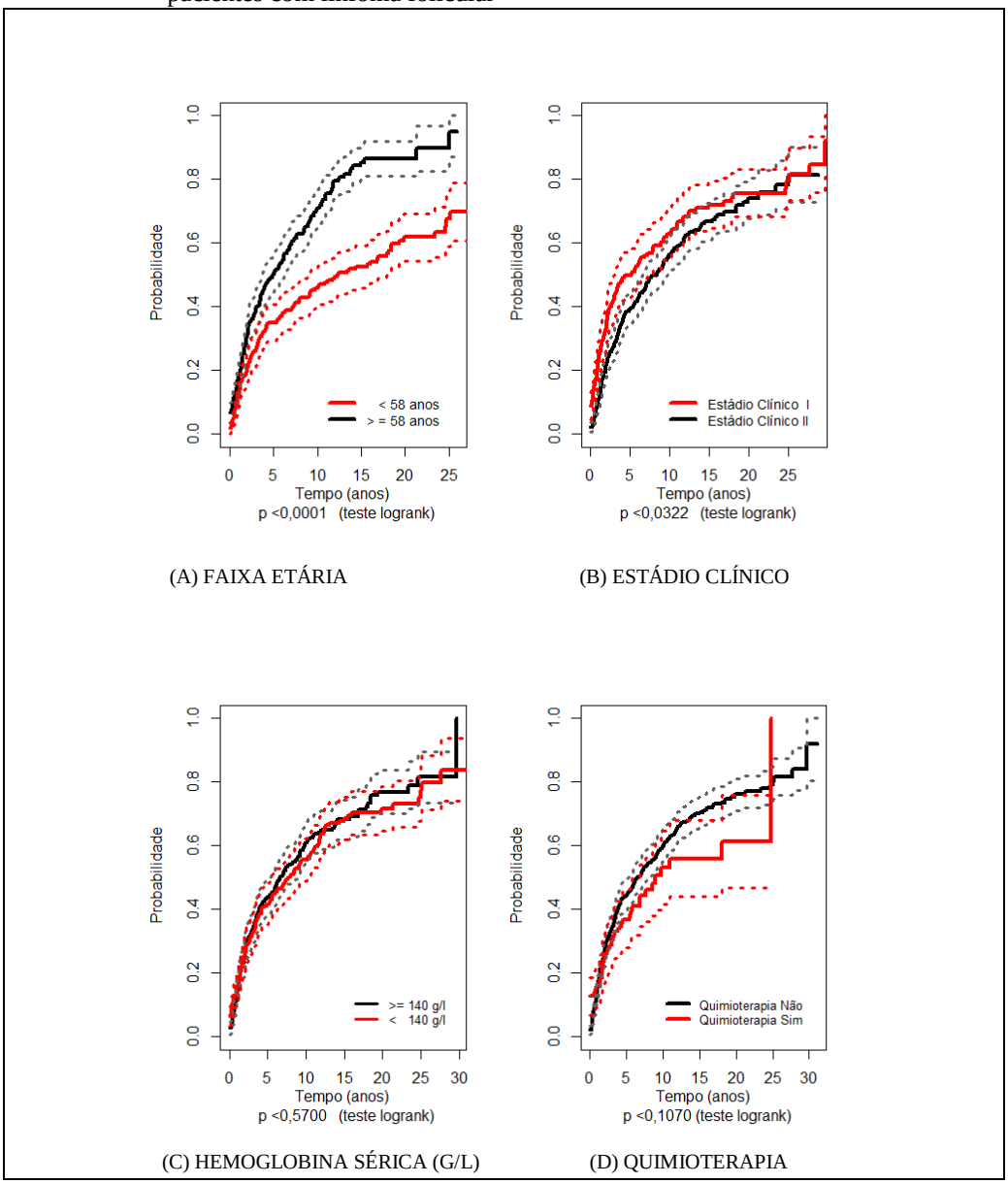
FIGURA 1 – Complemento da estimativa de Kaplan-Meier para desfecho composto: óbito em remissão ou recidiva da doença em pacientes com linfoma folicular



FONTE: O autor (2015).

A seguir, estimou-se as probabilidades de falha para o desfecho composto considerando-se, uma a uma, todas as covariáveis em estudo (Figura 2). Observa-se, sob esta abordagem e por meio do teste *logrank*, que as curvas de distribuição das falhas são semelhantes entre as categorias das covariáveis dosagem de hemoglobina (p = 0,5700) e quimioterapia (p = 0,1070), e diferentes entre as categorias das covariáveis idade (p < 0,001) e estádio clínico da doença (p = 0,0322) ao nível de significância de 5%.

FIGURA 2 – Complemento da estimativa de Kaplan-Meier conforme a faixa etária, estádio clínico da doença, hemoglobina sérica e a necessidade de quimioterapia em pacientes com linfoma folicular

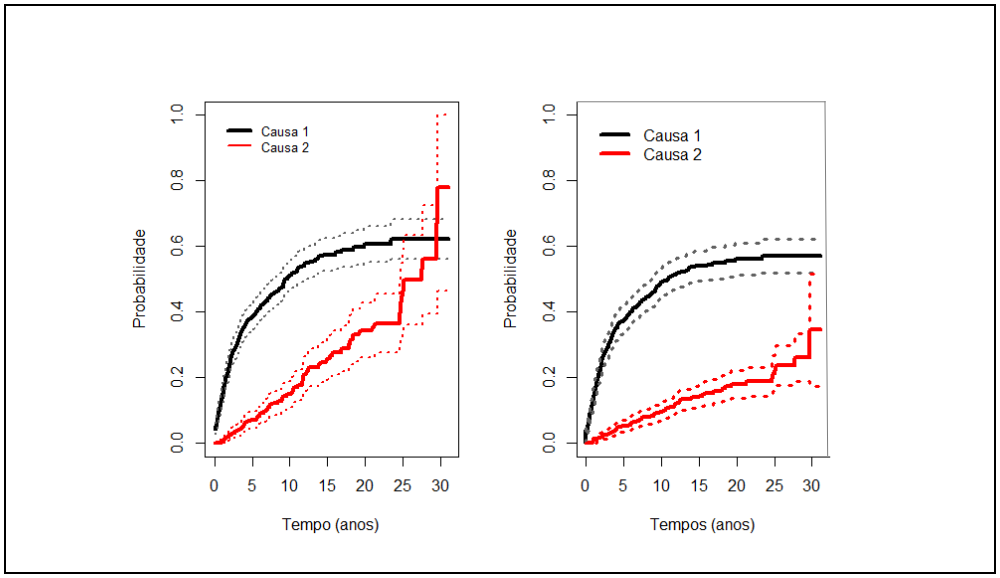


FONTE: O autor (2015).

Fez-se, a seguir, o processo de estimação numa estrutura de riscos competitivos no qual a causa 1, óbito em recidiva da doença, é a causa de interesse e a causa 2, óbito sem recidiva, é a causa concorrente.

Estimou-se, então, a probabilidade de falha causa-específica pelo método de KM e a função de incidência acumulada ou subdistribuição por meio do estimador não paramétrico descrito na Seção 3.2.1.2 (Figura 3) com o auxílio da função `cuminc` do pacote `cmprsk` do R (GRAY, 2014). Ao se comparar as curvas causa-específicas obtidas por este estimador com as obtidas pelo método de Kaplan-Meier, tem-se que este fornece estimativas superiores associadas à probabilidade de falha para a causa 2 (Figura 4). Para a causa 1, observa-se a partir da Tabela 1, que as estimativas intervalares de 1-KM e da subdistribuição se sobrepõem. Para a causa 2, a partir do décimo ano de seguimento, nessa amostra de tempos, a estimativa pontual da subdistribuição não é contemplada na estimativa intervalar de 1-KM.

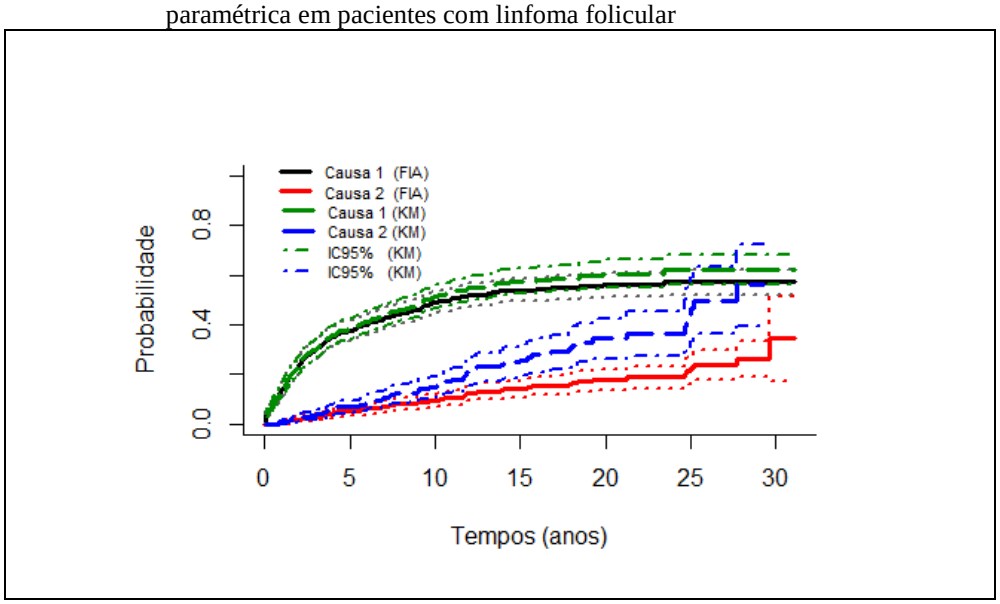
FIGURA 3 – Complemento da estimativa de Kaplan-Meier causa-específica e função de incidência acumulada para riscos competitivos em pacientes com linfoma folicular



FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença.
Causa 2 – óbito sem recidiva.

FIGURA 4 – Estimativa da incidência acumulada de falha: comparação entre abordagem de Kaplan-Meier causa-específica e função de incidência acumulada não paramétrica em pacientes com linfoma folicular



FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença.
Causa 2 – óbito sem recidiva.
FIA – função de incidência acumulada.

TABELA 1 – Comparação entre o método de Kaplan-Meier causa-específica e a estimação da função de incidência acumulada para recidiva da doença com óbito sem recidiva como causa concorrente em pacientes com linfoma folicular

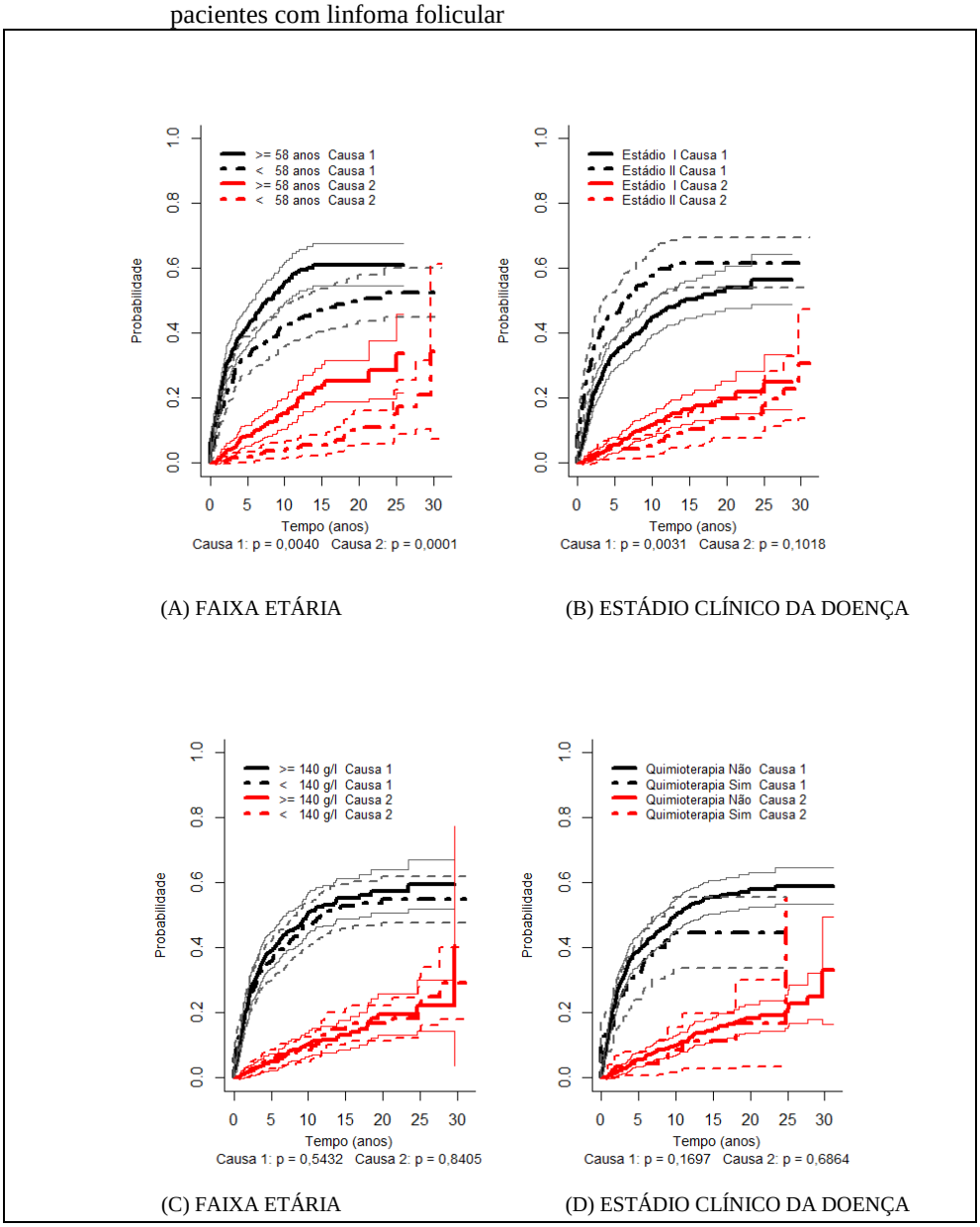
t_j	EM RISCO (N)	RECIDIVA (N)	ÓBITO (N)	1-KM	INTERVALO DE CONFIANÇA (95%, 1-KM)		FIA	INTERVALO DE CONFIANÇA (95%, FIA)	
					INFERIOR	SUPERIOR		INFERIOR	SUPERIOR
CAUSA 1									
0,60	488	1	0	0,100	0,075	0,125	0,100	0,076	0,127
5,13	283	1	0	0,388	0,346	0,429	0,379	0,338	0,420
10,18	147	1	0	0,517	0,470	0,563	0,494	0,448	0,537
14,32	87	1	0	0,575	0,526	0,625	0,537	0,489	0,582
19,82	42	1	0	0,607	0,552	0,662	0,562	0,512	0,609
23,39	25	1	0	0,622	0,561	0,683	0,562	0,512	0,609
CAUSA 2									
0,92	467	0	1	0,011	0,001	0,02	0,009	0,004	0,021
5,55	269	0	1	0,076	0,049	0,102	0,054	0,037	0,076
10,06	148	0	1	0,154	0,111	0,197	0,095	0,070	0,123
13,98	91	0	1	0,248	0,188	0,308	0,140	0,109	0,176
19,27	46	0	1	0,345	0,262	0,427	0,174	0,136	0,217
29,67	2	0	1	0,78	0,465	1	0,345	0,184	0,513

FONTE: O autor (2015).

NOTA: t_j - tempo de ocorrência dos eventos (em anos).
1 – KM – (1 – estimativa de Kaplan-Meier).
FIA – função de incidência acumulada.

Quando se testou o efeito isolado de cada uma das covariáveis em estudo sobre as curvas de incidência acumulada de falha causa-específica, foram observadas diferenças significativas, pelo teste *logrank* e pelo teste de Gray (GRAY, 1988), entre as categorias das covariáveis hemoglobina e estádio clínico (Figura 5; Tabela 2).

FIGURA 5 – Função de incidência acumulada conforme a faixa etária, o estádio clínico da doença, a hemoglobina sérica e a necessidade de quimioterapia em pacientes com linfoma folicular



(A) FAIXA ETÁRIA (B) ESTÁDIO CLÍNICO DA DOENÇA

(C) FAIXA ETÁRIA (D) ESTÁDIO CLÍNICO DA DOENÇA

FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença.
Causa 2 – óbito sem recidiva.
p – valor *p* associado ao teste de Gray

TABELA 2 – Testes de *logrank* e de Gray para avaliar a significância das covariáveis sobre a função de incidência acumulada para recidiva em pacientes com linfoma folicular

COVARIÁVEL	VALOR *p*	
	Teste *logrank*	Teste de Gray.
CAUSA 1		
Idade	0,0002	0,0040
Estádio Clínico	0,0045	0,0031
Hemoglobina	0,5370	0,5432
Quimioterapia	0,1590	0,1697
CAUSA 2		
Idade	< 0,0001	< 0,0001
Estádio Clínico	0,4290	0,1018
Hemoglobina	0,9610	0,8405
Quimioterapia	0,4320	0,6864

FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença.
Causa 2 – óbito sem recidiva.

4.2 Resultados do Modelo de Fine-Gray

O modelo de Fine-Gray, considerando as duas causas de falha e as quatro covariáveis disponíveis no estudo, apresentou os resultados mostrados na Tabela 3. Estes foram obtidos por meio da função `crr` do pacote `cmprsk` do R (GRAY, 2014). Na Tabela 3, tem-se as estimativas dos efeitos das covariáveis considerando a presença de uma covariável por vez no modelo, bem como com todas as covariáveis conjuntamente, sendo removidas, uma a uma, as não significativas ao nível de 5%. Permaneceram no modelo final as covariáveis idade e estádio clínico, ambas com valor $p < 0,001$.

Testou-se, ainda, uma possível modificação da magnitude, ao longo do tempo, do efeito das covariáveis que permaneceram no modelo final, isto é, a proporcionalidade dos riscos: $H_0: \beta_k(t) = \beta_k$. Para essa finalidade, foram testadas o efeito das interações de cada covariável no modelo (idade e estádio clínico) com funções do tempo. A partir da Tabela 4, observam-se efeitos não significativos associados às interações mencionadas, indicando que os efeitos das covariáveis não apresentam variações marcantes ao longo do tempo.

TABELA 3 – Estimativa dos efeitos das covariáveis sobre a subdistribuição das falhas associadas à causa 1 em pacientes com linfoma folicular

COVARIÁVEL	ANÁLISE UNIVARIADA				ANÁLISE MÚLTIPLA			
	$\hat{\beta}$	INTERVALO DE CONFIANÇA $(95\%, \hat{\beta})$		VALOR *p*	$\hat{\beta}$	INTERVALO DE CONFIANÇA $(95\%, \hat{\beta})$		VALOR *p*
CAUSA 1								
MODELO *FULL*								
Idade	1,01	1,00	1,02	0,0026	1,017	1,008	1,03	0,0031
Estádio Clínico	1,45	1,13	1,87	0,0032	1,745	1,339	2,27	0,0004
Hemoglobina	1,00	0,99	1,01	0,9500	1,002	0,995	1,01	0,5600
Quimioterapia	0,79	0,57	1,10	0,1600	0,717	0,511	1,01	0,1921
MODELO 1								
Idade					1,017	1,008	1,03	0,0036
Estádio Clínico					1,730	1,328	2,25	< 0,0001
Quimioterapia					0,721	0,515	1,01	0,0580
MODELO FINAL								
Idade					1,02	1,01	1,03	0,0003
Estádio Clínico					1,63	1,26	2,12	0,0002

FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença.

TABELA 4 – Estimativa dos efeitos tempo-dependentes das covariáveis sobre a subdistribuição das falhas associadas à causa 1 em pacientes com linfoma folicular

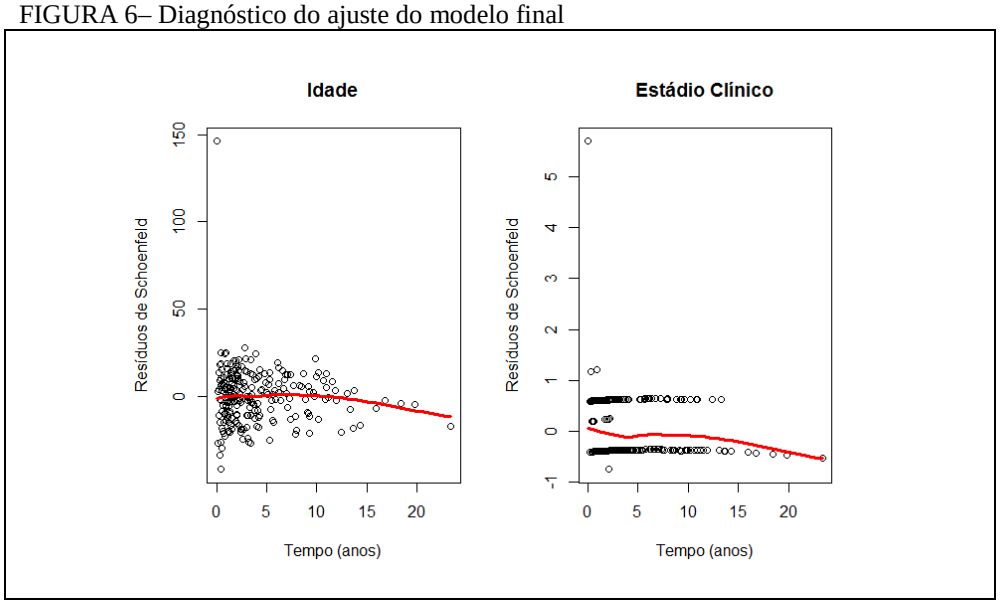
	$\hat{\beta}$	INTERVALO DE CONFIANÇA $(95\%, \hat{\beta})$		VALOR *p*
		INFERIOR	SUPERIOR	
MODELO COM INTERAÇÃO IDADE-TEMPO				
Idade	1,02	1,003	1,03	0,01700
Estádio Clínico	1,63	1,257	2,11	0,00023
Idade*tempo	1,00	0,997	1,01	0,62000
Idade*tempo^2	1,00	1,00	1,00	0,19000
MODELO COM INTERAÇÃO ESTÁDIO CLÍNICO-TEMPO				
Idade	1,017	1,007	1,03	0,00043
Estádio Clínico	2,129	1,418	3,20	0,00027
Estádio *tempo	0,935	0,801	1,09	0,40000
Estádio *tempo^2	0,998	0,989	1,01	0,71000

FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença.

A distribuição de resíduos associados ao modelo final, análogos aos resíduos de Schoenfeld, também evidenciou que a suposição de proporcionalidade, isto é, $\beta(t) = \beta$, não está sendo seriamente violada para as covariáveis que permaneceram no modelo final (Figura 6).

FIGURA 6– Diagnóstico do ajuste do modelo final



FONTE: O autor (2015).

4.3 Resultados do Modelo de Scheike, Zhang e Gerds

Ao utilizar-se o modelo proposto por Scheike, Zhang e Gerds (2008), que permite considerar efeitos fixos e efeitos variando no tempo na função de subdistribuição, obteve-se, quanto à significância dos efeitos fixos das covariáveis, resultados semelhantes ao do modelo de Fine e Gray (1999), isto é, efeito significativo das covariáveis idade e estádio clínico.

TABELA 5 – Estimativas dos efeitos das covariáveis sobre a subdistribuição das falhas associadas à causa 1 por meio do modelo de Scheike-Zhang-Gerds em pacientes com linfoma folicular

	EFEITO FIXO VALOR p (1)	EFEITO TEMPO-DEPENDENTE VALOR p (2)
COVARIÁVEL		
MODELO *FULL*		
Idade	0,0046	0,0006
Estádio Clínico	0,0008	0,0000
Hemoglobina	0,4250	0,8410
Quimioterapia	0,3130	0,1920
MODELO 1		
Idade	0,0030	0,0006
Estádio Clínico	0,0012	0,0000
Quimioterapia	0,3340	0,1810
MODELO FINAL		
Idade	0,004	0,0006
Estádio Clínico	< 0,0001	< 0,0001

FONTE: O autor (2015).

NOTA: Causa 1 – recidiva da doença

(1) Teste do Supremo descrito em Martinussen e Scheike (2006)

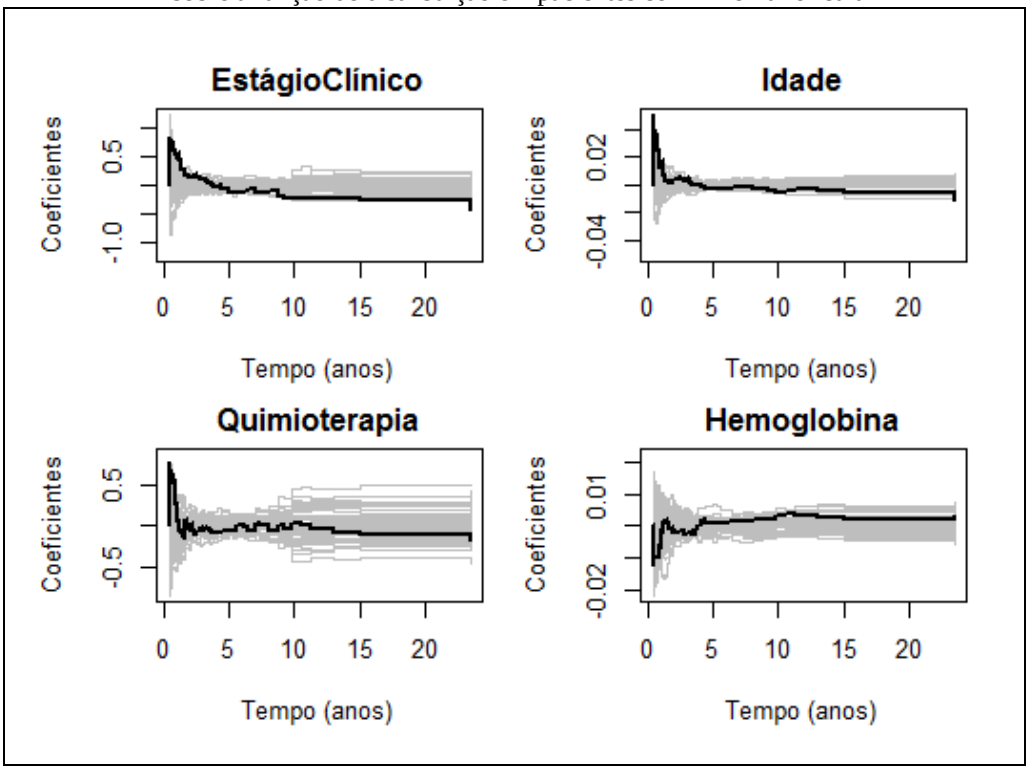
(2) Teste de Cramer von Misses descrito em Martinussen e Scheike (2006).

Já quanto à proporcionalidade dos efeitos das covariáveis, nota-se a partir dos resultados dos valores-p dos dois testes mostrados na Tabela 5, que a proporcionalidade dos efeitos ($\beta_k(t) = \beta_k$) foi rejeitada para ambas as covariáveis.

Os gráficos dispostos nas Figuras 7 e 8 (modelo completo) e na Figura 9 (modelo final), mostram as estimativas dos efeitos das covariáveis ao longo do tempo, sendo possível observar que estas apresentam variações, em particular no intervalo de tempo de 0 a 3 anos. A partir deste tempo, os efeitos podem ser considerados constantes.

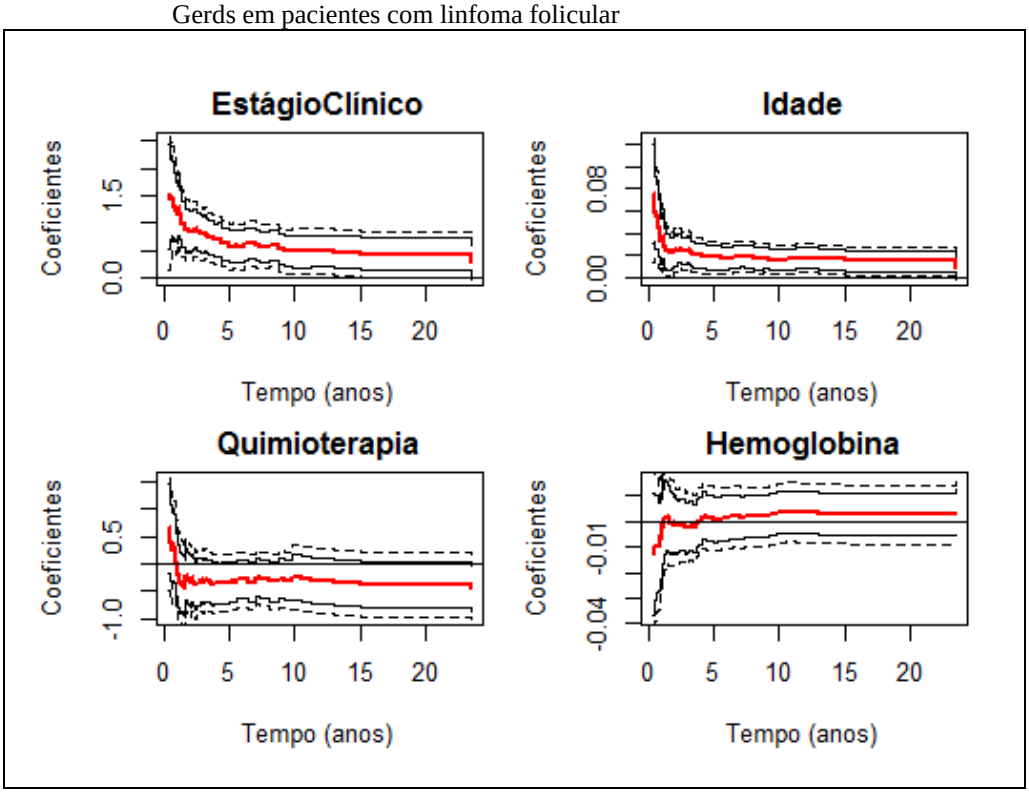
Em relação ao modelo de Fine-Gray, nota-se ter havido uma discordância com o de Sheike-Zhang-Gerds quanto à proporcionalidade dos efeitos das covariáveis. Contudo, os gráficos nas Figuras 7, 8 e 9 mostram que a variação dos efeitos das covariáveis não são tão marcantes e, sendo assim, o modelo de Fine-Gray pode ser considerado uma opção razoável para o ajuste dos dados do estudo.

FIGURA 7 – Análise da variação temporal do efeito das covariáveis do modelo completo sobre a função de distribuição em pacientes com linfoma folicular



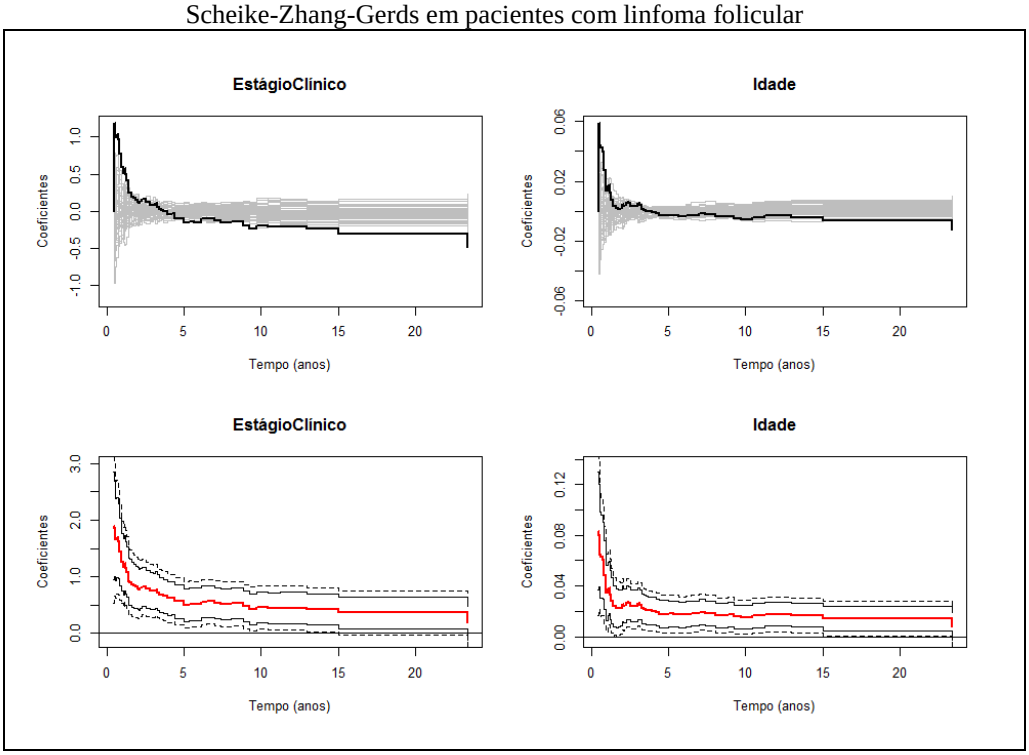
FONTE: O autor (2015).

FIGURA 8 – Estimativas do efeito tempo-dependente das covariáveis do modelo completo sobre a função de subdistribuição por meio do modelo de Scheike-Zhang-Gerds em pacientes com linfoma folicular



FONTE: O autor (2015).

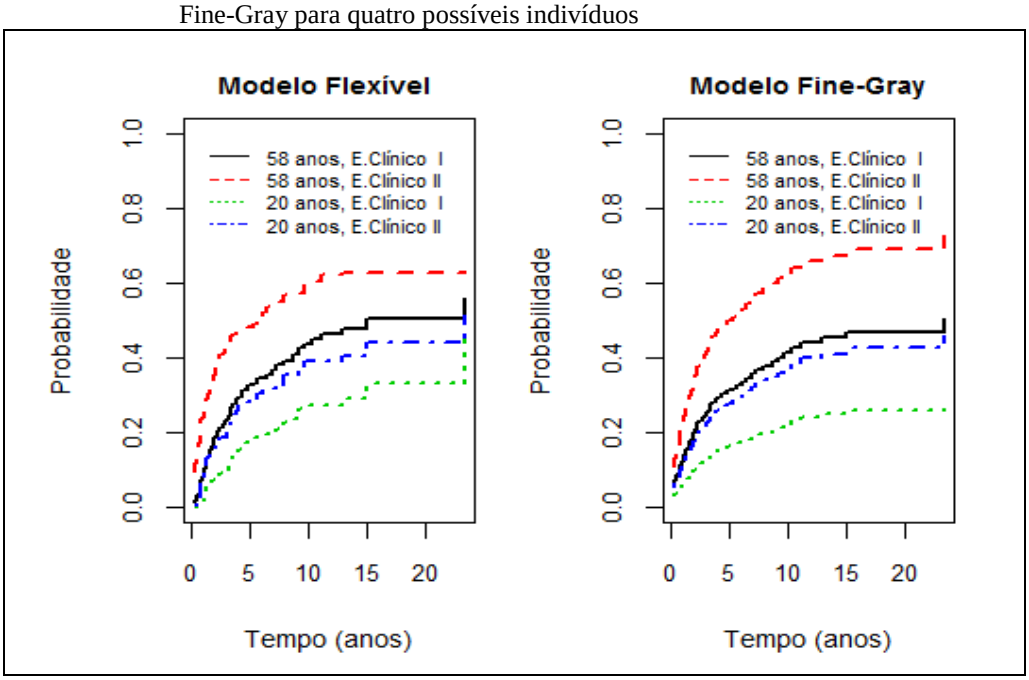
FIGURA 9 – Análise da variação temporal e estimativas do efeito das covariáveis do modelo final sobre a função de subdistribuição por meio do modelo de Scheike-Zhang-Gerds em pacientes com linfoma folicular



FONTE: O autor (2015).

A Figura 10 apresenta predições para as funções de subdistribuição de quatro possíveis indivíduos estimadas a partir de ambos os modelos, o de Sheike-Zhang-Gerds e o de Fine-Gray, sendo possível notar, como mencionado anteriormente, que ambos apresentam resultados similares.

FIGURA 10– Predições conforme estimativas dos modelos de Scheike-Zhang-Gerds e de Fine-Gray para quatro possíveis indivíduos



FONTE: O autor (2015).

5 CONCLUSÃO

Neste trabalho foram apresentados e comparados dois métodos não paramétricos e dois modelos de regressão frequentemente utilizados em estudos de Análise de Sobrevivência os quais têm por objetivo estimar a probabilidade de um particular desfecho numa estrutura de riscos competitivos.

Inicialmente, a abordagem não paramétrica de Kaplan-Meier causa-específica foi comparada com uma abordagem não paramétrica alternativa, a qual estima diretamente a função de incidência acumulada (FIA). Em seguida, os modelos de Fine-Gray e de Scheike-Zhang-Gerds foram ajustados a fim de considerar todas as covariáveis conjuntamente na análise e, então, testar o efeito dessas covariáveis na subdistribuição das falhas associadas à causa 1, causa de interesse.

Para o conjunto de dados utilizado, foi observado para a causa 1, em todo o período de observação, semelhanças entre as estimativas das probabilidades de falha obtidas pelo método de KM e FIA, na presença de causas concorrentes. Para a causa 2, as semelhanças ocorreram até o quinto ano de seguimento. Para a causa 1 verificou-se, ainda, que as estimativas pontuais fornecidas pelos dois métodos não paramétricos se apresentaram muito próximas até o décimo ano

Comparando-se os resultados dos testes, observou-se que o teste *logrank* e o teste de Gray foram concordantes quanto às diferenças nos riscos causa-específicas entre os diferentes níveis das covariáveis. Estes mostraram que idade igual ou superior a 58 anos está associada a maior probabilidade de falha, bem como que o estádio clínico II está associado a maior probabilidade de falha pela causa 1 e o estádio clínico I à causa 2.

Quanto aos modelos de Fine-Gray e Scheike-Zhang-Gerds, ambos evidenciaram efeito significativo das mesmas covariáveis na subdistribuição associadas às causas 1 e 2. Entretanto, a avaliação da suposição de proporcionalidade dos riscos resultou em conclusões discordantes. O modelo de Fine-Gray, quando estendido para testar efeitos tempo-dependentes, bem como a análise dos resíduos análogos aos de Schoenfeld do modelo final, não indicaram variação temporal dos efeitos das covariáveis. Por outro lado, o modelo de Scheike-Zhang-Gerds indicou efeito tempo-dependente para ambas as covariáveis que permaneceram no modelo final, ficando tal variação restrita, no entanto, aos primeiros três anos de seguimento. A partir do terceiro ano os efeitos se apresentaram constantes (Figura 9).

No geral, concluiu-se que ambos os modelos se apresentaram como opções razoáveis para a análise do conjunto de dados. Contudo, se o interesse se concentrar na subdistribuição das causas 1 e 2 nos primeiros anos de seguimento, o modelo de Scheike-Zhang-Gerds seria o mais indicado, tendo em vista que este considera de forma mais detalhada as oscilações dos efeitos das covariáveis nesse período.

Quanto aos métodos não paramétricos, a abordagem de Kaplan-Meier, apesar de útil enquanto técnica exploratória, mostrou limitações que sugerem cuidado quanto ao seu uso no contexto de riscos competitivos.

REFERÊNCIAS

AALEN, O.O. Nonparametric inference for a family of counting processes. **The Annals of Statistics**, Philadelphia ,v. 6, n. 4, p. 701-727, 1978.

BEYERSMANN, J.; SCHUMACHER, M. Time-dependent covariates in the proportional subdistribution hazards model for competing risks. **Biostatistics**, Oxford, v. 9, p. 765–776, 2008.

CARVALHO, M.S.; ANDREOZZI, V.L.;CODEÇO, C.T.; CAMPOS, D.P. *et al*. **Análise de sobrevivência: teoria e aplicações em saúde.** 2 ed. Rio de Janeiro. Editora Fiocruz, 2011. 432p.

COLE, S.R.;HUDGENS, M.G.;BROOKHART, M.A.;WESTREICH, D. Risk. **American Journal of Epidemiology**, Baltimore , v. 181, n. 4, 246-250, 2015. Disponível em: <http://aje.oxfordjournals.org/content/181/4/246.full>. Acesso em: 07/05/2015.

COLLETT, D. Modelling Survival Data in Medical Research. 2nd ed. London. Chapman and Hall/CRC, 2003. 410p.

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

COX, D.R. Partial Likelihood. **Biometrika**, London, v. 62, n. 2, p. 269-276, 1975. Disponível em: < http://www.jstor.org/stable/pdf/2335362>. Acesso em: 18/06/2015.

COX, D.R. Regression Models and Life-Tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, London, v. 34, n. 2, p. 187-220, 1972.

DIGNAM,J.J.; ZHANG, Q.; KOCHERGINSKY, M. The use and interpretation of competing risks regression models. **Clinical Cancer Research**, Denville, v. 18, n. 8, p. 2301-2308; 2012.

DIXON, S. N.; DARLINGTON, G. A.; DESMOND, A. F. A competing risks model for correlated data based on the subdistribution hazard, **Lifetime Data Analysis**, Boston, v. 17, p. 473–495, 2011.

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

GRAY, R. J. A class of K-sample tests for comparing the cumulative incidence of a competing risk. **Annals of Statistics**, San Francisco, v. 16, p. 1141-1154, 1988.

GRAY, R. J. **cmprsk: subdistribution analysis of competing risks**. R package version 2.2-7. 2014. Disponível em: <http://CRAN.R-project.org/package=cmprsk>. Acesso em: 15/10/2014.

HANLEY, J.A. The Breslow estimator of the nonparametric baseline survivor function in Cox's regression model: some heuristics. **Epidemiology**, Cambridge, v. 19, n. 1, p. 101-102, 2008.

KAPLAN, E.L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, New York, v. 53, n. 282, p. 457-481, 1958.

KATSAHIAN, S.; RESCHERIGON, M.; CHEVRET, S.; PORCHER, R. Analysing multicenter competing risks data with a mixed proportional hazards model for the subdistribution. **Statistics in Medicine**, Chichester, v. 25, p. 4267–4278, 2006.

KATSAHIAN, S.; BOUDREAU, C., Estimating and testing for center effects in competing risks, **Statistics in Medicine**, Chichester, v. 30, p. 1608–1617, 2011.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. 2nd ed. New York: Springer, 2003. 538p.

LARSON, M. G. Covariate analysis of competing risks models with log-linear models, **Biometrics**, Washington, v. 40, p. 459-469, 1984.

NELSON, W. Theory and Applications of Hazard Plotting for Censored Failure Data. **Technometrics**, Richmond, v. 14, n. 4, p. 945-966, 1972.

Pérez-Marín, A.M. Empirical comparison between the Nelson-Aalen estimator and the naive local constant estimator. **Statistics and Operations Research Transactions**, Barcelona, v. 32, n. 1, p.76-76, 2008.

PEPE, M. S. Inference for events with dependents risks in multiple endpoint studies. **Journal of the American Statistical Association**, New York, v. 86, p. 770-778, 1991.

PEPE, M. S.; MORI, M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? **Statistics in Medicine**, Chichester, v. 12, n. 8, p. 737–751, 1993.

PRENTICE, R. L.; KALBFLEISCH, J. D.; PETERSON, A. V.; FLOURNOY, N.; FAREWELL, V. T.; BRESLOW, N. E. The analysis of failure times in the presence of competing risks. **Biometrics**, Washington, v. 34, p. 541-554, 1978.

PINTILIE, M. **Competing Risks: a practical perspective**. Chichester: John Wiley & Sons, Ltd, 2006. 224 p.

R CORE TEAM. *R:* **A Language and Environment for Statistical Computing**. Vienna, Austria, 2014. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

SANTOS, J.L.F; ORTIZ, L.P.; YAZAKI, L.M. **Aplicação da técnica de riscos competitivos a dados brasileiros**. Disponível em: < http://www.abep.nepo.unicamp.br/docs/anais/pdf/1984/T84V02A21.pdf>. Acesso em: 20/04/2015.

SAS INSTITUTE INC. **SAS**. North Carolina, USA. Disponível em: <http://www.sas.com/pt_br/home.html/>. Acesso em: 20/06/2015.

SCHEIKE, T. H.; MARTINUSSEN, T. **Dynamic regression models for survival data**. New-York: Springer, 2006. 470 p.

SCHEIKE, T. H.; ZHANG, M. J. A flexible competing risks regression modeling and goodness-of-fit. **Lifetime Data Analysis**, Boston, **v. 14, p. 464-483, 2008.**

SCHEIKE, T. H.; ZHANG, M. J.; GERDS, T. **A.** Predicting cumulative incidence probability by direct binomial regression. **Biometrika**, London, v. 95, p. 205–220, 2008.

SCHEIKE, T. H.; SUN, Y.; ZHANG, M. J.; JENSEN, T. K. A semiparametric random effects model for multivariate competing risks data. **Biometrika**, London, v. 97, p. 133–145, 2010.

SCHEIKE, T. H.; ZHANG, M. J. Analyzing competing risk data using the R timereg package. **Journal of Statistical Software**, Los Angeles, v. 38, n. 2, p. 1–15, 2011.

SCRUCCA, L.; SANTUCCI, A.; AVERSA, F. Competing risk analysis using R: an easy guide for clinicians. **Bone Marrow Transplantation**, London, v. 40, n. 4, p. 381-387, 2007.

SELVIN, S. **Statistical analysis of epidemiologic data**. 3rd ed. New-York: Oxford University Press, 2004. 492 p.

STATACORP LP. **Data Analysis and Statistical Software**. College Station. TX. USA. Disponível em: < http://www. http://www.stata.com/>. Acesso em: 29/05/2015.

APÊNDICE

```
# Complemento das estimativas de Kaplan-Meier para causa 1 e causa 2 - F(t)
require(survival)
ekm.1<- survfit(Surv(dftime,evcens)~1,conf.type="plain")
ekm.1
summary(ekm.1)
ekm.2<- survfit(Surv(dftime,crcens)~1,conf.type="plain")
ekm.2
plot(ekm.1$time, 1-ekm.1$surv, xlab = 'Tempo (anos)',lwd=4,
     main='',cex.main=1,
     ylab='Probabilidade',
     col='black',type="s",ylim=c(0,1))
lines(ekm.1$time, 1-ekm.1$lower, type="s", lty=3, col='gray38', lwd=2)
lines(ekm.1$time, 1-ekm.1$upper, type="s", lty=3, col='gray38', lwd=2)

lines(ekm.2$time, 1-ekm.2$surv, col='red', lwd=4)
lines(ekm.2$time, 1-ekm.2$lower, type="s", lty=3, lwd=2, col='red1')
lines(ekm.2$time, 1-ekm.2$upper, type="s", lty=3, lwd=2, col='red')
legend(0.2,1,lty=c(1,1,3), col=c('black', 'red', 'red'),lwd=c(4,2,2),
       c("Causa 1", "Causa 2"),cex=.8, bty='n')

###  Idade  ###
# geral #
ekm.age<- survfit(Surv(dftime,cens)~age1,conf.type="plain")
ekm.age
summary(ekm.age)
names(ekm.age)
str(ekm.age)
plot(ekm.age[1]$time, 1-ekm.age[1]$surv, xlab = 'Tempo (anos)\n  p <0,0001   (teste
logrank)', ylab=' Probabilidade', type="s", ylim=c(0,1),
     main='',
     col=c('black'),lwd=4, cex.main=.9)
lines(ekm.age[2]$time, 1-ekm.age[2]$surv,  type="s", col = 'red',lty=1, lwd=4)

lines(ekm.age[1]$time,1- ekm.age[1]$lower, lty=3, col='gray38', type='s', lwd=3)
lines(ekm.age[1]$time,1- ekm.age[1]$upper, lty=3, col='gray38', type='s',lwd=3)
lines(ekm.age[2]$time,1- ekm.age[2]$lower, lty=3, col='red1', type='s', lwd=3)
lines(ekm.age[2]$time,1- ekm.age[2]$upper, lty=3, col='red1', type='s', lwd=3)
legend(10,.15,lty=c(1,1), col=c('red', 'black'), lwd=c(4,4),
       c("   < 58 anos", "> = 58 anos"),cex=.9, bty='n')

## teste log-rank
fit2.ekm.age=survdiff(Surv(dftime,cens)~age1, data=da,rho=0); fit2.ekm.age
fit2.ekm.age
fit2.ekm.age$var

fit2.ekm.age1=survdiff(Surv(dftime,evcens)~age1, data=da,rho=0); fit2.ekm.age1
fit2.ekm.age2=survdiff(Surv(dftime,crcens)~age1, data=da,rho=0); fit2.ekm.age2

## Estimativa KM da incidência de falha causa-específica por faixa etária
##  KM por faixa etária para cada Causa
ekm.age.1<- survfit(Surv(dftime,evcens)~age1,conf.type="plain")
summary(ekm.age.1)
ekm.age.2<- survfit(Surv(dftime,crcens)~age1,conf.type="plain")
ekm.age.1
ekm.age.2
summary(ekm.age.1)
summary(ekm.age.2)
fit2.ekm.age1=survdiff(Surv(dftime,evcens)~age1, data=da,rho=0); fit2.ekm.age1
fit2.ekm.age2=survdiff(Surv(dftime,crcens)~age1, data=da,rho=0); fit2.ekm.age2


# F(t)
plot(ekm.age.1[1]$time, 1-ekm.age.1[1]$surv, xlab = 'Tempo (anos)', ylab='
Probabilidade', type="s", ylim=c(0,1),
     main='', col=c('black'),lwd=2, cex.main=.9)
lines(ekm.age.1[2]$time, 1-ekm.age.1[2]$surv,  type="s", col = 'black',lty=4, lwd=2)

lines(ekm.age.1[1]$time,1- ekm.age.1[1]$lower, lty=3, col='gray38', type='s')
lines(ekm.age.1[1]$time,1- ekm.age.1[1]$upper, lty=3, col='gray38', type='s')
lines(ekm.age.1[2]$time,1- ekm.age.1[2]$lower, lty=3, col='gray38', type='s')
lines(ekm.age.1[2]$time,1- ekm.age.1[2]$upper, lty=3, col='gray38', type='s')

lines(ekm.age.2[1]$time, 1- ekm.age.2[1]$surv , lty=1, col='red', lwd=2, type='s')
lines(ekm.age.2[2]$time, 1- ekm.age.2[2]$surv,  lty=4, col='red', lwd= 2, type='s')
lines(ekm.age.2[1]$time, 1- ekm.age.2[1]$upper, lty=3, col='red1', type='s')
lines(ekm.age.2[1]$time, 1- ekm.age.2[1]$lower, lty=3, col='red1', type='s')
lines(ekm.age.2[2]$time, 1- ekm.age.2[2]$upper, lty=3, col='red1', type='s')
lines(ekm.age.2[2]$time, 1- ekm.age.2[2]$lower, lty=3, col='red1', type='s')
```

```
legend(-1,1.05,lty=c(1,1,4,4), col=c('red','black', 'red', 'black'),
       c("<   58 anos  Causa 1",">= 58 anos  Causa 1","<   58 anos  Causa 2", ">= 58 anos
Causa 2"),cex=.8, bty="n")

# função cuminc - geral

layout(1)
ci1<-cuminc(dftime, cause, rho=0, cencode=0)
print(ci1,digits=4)
plot(ci1, lty=1,lwd=4, col=1:2, xlab="Tempos (anos)", ylab="Probabilidade",ylim=c(0,1),
     main='', cex.main=1,
     curvlab = c("Causa 1", "Causa 2"))

## intervalo de confiança 95%
lines(ci1[[1]][1]$time,ci1[[1]][2]$est+1.96*sqrt(ci1[[1]][3]$var), lty=3, col='gray38',
type='s', lwd=3)
lines(ci1[[1]][1]$time,ci1[[1]][2]$est-1.96*sqrt(ci1[[1]][3]$var), lty=3, col='gray38',
type='s', lwd=3)
lines(ci1[[2]][1]$time,ci1[[2]][2]$est+1.96*sqrt(ci1[[2]][3]$var), lty=3, col='red1',
type='s', lwd=3)
lines(ci1[[2]][1]$time,ci1[[2]][2]$est-1.96*sqrt(ci1[[2]][3]$var), lty=3, col='red1',
type='s', lwd=3)
# função cuminc - faixa etária
op=par(mar=c(4,4,3,2)+.1)
par(mfrow=c(1,1))
ciage<-cuminc(dftime, cause,age1, rho=0, cencode=0)
print(ciage, digits=4)
plot( ciage,col=c(1,1,2,2),  xlab="Tempo (anos)\nCausa 1: p = 0,0040   Causa 2: p =
0,0001",
     main='', cex.lab=.9,
     cex=.8, lty=c(1,4,1,4),lwd=c(4,4,4,4), ylab="Probabilidade",
     curvlab = c(">= 58 anos  Causa 1","<   58 anos  Causa 1",">= 58 anos  Causa 2", "<
58 anos  Causa 2"))

lines(ciage[[1]][1]$time,ciage[[1]][2]$est+1.96*sqrt(ciage[[1]][3]$var), lty=1,
col='gray38', type='s', lwd=1)
lines(ciage[[1]][1]$time,ciage[[1]][2]$est-1.96*sqrt(ciage[[1]][3]$var), lty=1,
col='gray38', type='s', lwd=1)
lines(ciage[[2]][1]$time,ciage[[2]][2]$est+1.96*sqrt(ciage[[2]][3]$var), lty=2,
col='gray38', type='s', lwd=2)
lines(ciage[[2]][1]$time,ciage[[2]][2]$est-1.96*sqrt(ciage[[2]][3]$var), lty=2,
col='gray38', type='s', lwd=2)

lines(ciage[[3]][1]$time,ciage[[3]][2]$est+1.96*sqrt(ciage[[3]][3]$var), lty=1,
col='red1', type='s', lwd=1)
lines(ciage[[3]][1]$time,ciage[[3]][2]$est-1.96*sqrt(ciage[[3]][3]$var), lty=1,
col='red1', type='s', lwd=1)
lines(ciage[[4]][1]$time,ciage[[4]][2]$est+1.96*sqrt(ciage[[4]][3]$var), lty=2,
col='red1', type='s', lwd=2)
lines(ciage[[4]][1]$time,ciage[[4]][2]$est-1.96*sqrt(ciage[[4]][3]$var), lty=2,
col='red1', type='s', lwd=2)


## função crr
# matriz do delineamento
x=cbind(age, hgb, clinstg, Quimioterapia)

# first regression model for relapse

## mod1 = full model
mod1 = crr(dftime, cause, x)
summary(mod1)
names(mod1)

mod3 = crr(dftime, cause, x[,c(1,3)])   # idade + estádio clínico
summary(mod3)

# diagnóstico do modelo
mod3$res
names(mod3$coef)=c("Idade", "Estádio Clínico")
par(mfrow = c(1,2))  #, mar = c(4.5,4,2,1))
for(j in 1:ncol(mod3$res))
  scatter.smooth(mod3$uft, mod3$res[,j],lpars=list(col='red', lwd=3),
                 main = names(mod3$coef)[j],
                 xlab = 'Tempo (anos)',
                 ylab = 'Resíduos de Schoenfeld')

# variáveis tempo-dependentes
mod8 = crr(dftime, cause, cov1 = x[,c(1,3)],
           cov2 = cbind(age,age), tf = function(t) cbind(t,t^2))
summary(mod8)

# utilizando o timereg
```

```
out1 <- comp.risk(Event(dftime, cause ) ~ + 1, data = da,
                  cause= 1, n.sim = 5000,  model="additive")
summary(out1)
pout1 <- predict(out1, X = 1)

## modelos de regressão - coeficientes

outf3 <- comp.risk(Event(dftime, cause) ~ EstádioClínico +Idade,
                   data = da, cause = 1, n.sim = 5000,
                   model = "prop", cens.model = "cox")
summary(outf3)
plot(outf3, sim.ci = 2, col=c(2,2,2), lty=c(1,3,3), lwd=c(2,1,1), specific.comps =
c(2,3),
     cex.main=.9,ylab='Coeficientes', xlab='Tempo (anos)')
plot(outf3, score = 1, specific.comps = c(2,3),ylab='Coeficientes', xlab='Tempo (anos)')
```