UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

Riscos competitivos: uma aplicação na sobrevida de pacientes com câncer

Natalia Elis Giordani

Orientador: Prof. Dra. Suzi Alves Camey
Co-orientador: Prof. Dra. Luciana Neves Nunes

Porto Alegre, setembro de 2015.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

Riscos competitivos: uma aplicação na sobrevida de pacientes com câncer

Natalia Elis Giordani

Orientador: Prof.Dra. Suzi Alves Camey

A apresentação desta dissertação é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Mestre.

Porto Alegre, Brasil.

2015

CIP - Catalogação na Publicação

Giordani, Natalia Elis
Riscos competitivos: uma aplicação na sobrevida de pacientes com câncer / Natalia Elis Giordani. -- 2015.
101 f.

Orientadora: Suzi Alves Camey.
Coorientadora: Luciana Neves Nunes.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Medicina, Programa de Pós-Graduação em Epidemiologia, Porto Alegre, BR-RS, 2015.

1. Câncer. 2. Mortalidade. 3. Análise de Sobrevivência. 4. Eventos Competitivos. I. Alves Camey, Suzi, orient. II. Neves Nunes, Luciana, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

BANCA EXAMINADORA

Prof. Dr. Álvaro Vigo, Programa de Pós-graduação em Epidemiologia, UFRGS.

Prof. Dra. Bárbara Niegia Garcia de Goulart, Programa de Pós-graduação em Epidemiologia, UFRGS.

Prof. Dra. Mônica Maria Celestina de Oliveira, Departamento de Saúde Coletiva, UFCSPA.

AGRADECIMENTOS

Agradeço aos meus pais, Volmir e Marta, irmãs, Geovanna e Samantha, e avó, Antônia, pelo apoio e orações em todos os momentos pelos quais passei desde a minha entrada no mestrado.

Aos meus sobrinhos, Ana e Guto, minhas paixões, pelo amor e brincadeiras.

Às minhas orientadoras, Suzi e Luciana, pela compreensão e auxílio durante esse percurso. Tenho imensa admiração profissional e pessoal por vocês!

Aos amigos que se mostraram presentes, muitas vezes estando longe, nos momentos que precisei de forças para seguir.

Hoje compreendo que nada acontece por acaso e sou muito grata por ter (e poder ter novamente) pessoas especiais em meu caminho, que me ajudam a superar obstáculos, persistir e acreditar que tudo vale a pena.

SUMÁRIO

ABREVIATURAS E SIGLAS

HCPA – Hospital de Clínicas de Porto Alegre

INCA – Instituto Nacional de Câncer

SIM – Sistema de Informações de Mortalidade

CID-10 – Classificação Internacional de Doenças e Problemas Relacionados à Saúde

RESUMO

A quantidade de novos casos de câncer, o número de mortes causadas por ele e a quantidade de pessoas convivendo com a doença (cinco anos após o diagnóstico) têm crescido em todo o mundo. Em função disso, analisar dados de pacientes com câncer torna-se uma ferramenta necessária para avaliar os programas de tratamento e monitorar o progresso das iniciativas de controle da doença. No que tange a análise, a mortalidade é um dos parâmetros utilizados para avaliar os resultados dessa área e as metodologias tradicionalmente utilizadas compreendem o método de Kaplan-Meier e o modelo de Cox. Ambos, porém, desprezam que um paciente com câncer pode vir a óbito por um câncer diferente do primeiro diagnosticado ou, até mesmo, por causas não relacionadas à doença. Portanto, propomos a utilização e entendimento de métodos de análise de sobrevivência que consideram eventos competitivos a fim avaliar incidências, letalidades e fatores associados ao óbito de pacientes com câncer primário atendidos no Hospital de Clínicas de Porto Alegre entre 2002 e 2009. Os resultados obtidos permitiram um melhor conhecimento dos tipos de cânceres com maiores incidências (pele (1.920 casos), próstata (1.080 casos), brônquios e pulmões (950 casos), mama (893 casos), sistema hematopoiético e reticuloendotelial (654 casos), cólon (573 casos), esôfago (497 casos), estômago (422 casos), neoplasia maligna secundária e não especificada dos gânglios linfáticos (360 casos) e colo do útero (328 casos)) e letalidades (pâncreas (145 óbitos; 57,1%), brônquios e pulmões (527 óbitos; 55,5%) e esôfago (262 óbitos, 52,7%)), considerando os eventos competitivos. Além disso, avaliou-se como *sexo* e *idade* contribuem para o risco de óbito de alguns tipos de câncer: homens têm risco menor de falecer por câncer de esôfago quando comparados às mulheres, já o aumento da idade aparece como fator de risco para o câncer de próstata. Esse estudo permitiu caracterizar o perfil dos casos de câncer atendidos pelo hospital considerando, para as estimativas, os eventos competitivos. Em função das vantagens do método, recomenda-se aos pesquisadores que não desprezem, em seus estudos, situações com eventos competitivos, uma vez que há *softwares* e diversos materiais disponíveis que auxiliam e facilitam sua aplicação.

Palavras-Chave: Câncer, Mortalidade, Análise de Sobrevivência, Eventos Competitivos

ABSTRACT

The amount of new cancer cases, the number of deaths caused by it, and the number of people living with the disease (five years after the diagnosis) have grown around the world. Due that, analyzing cancer patient's data becomes a necessary tool for evaluating treatment programs and monitor the progress of the disease control initiatives. Regarding the analysis, mortality is one of the parameters used to evaluate the results of this area and the methodologies traditionally used include the Kaplan-Meier and Cox model. However, these methodologies do not consider the fact that the death of a cancer patient can be caused by a different cancer diagnosed or even by causes unrelated to the disease. Therefore, we propose the use and understanding of survival analysis methods that consider competing events in order to assess incidence, lethality and factors associated with death in patients with primary cancer attended at Hospital de Clínicas de Porto Alegre from 2002 to 2009. The results allowed a better understanding of the types of cancers with higher incidence (skin (1,920 cases), prostate (1,080 cases), bronchi and lungs (950 cases), breast (893 cases), hematopoietic and reticuloendothelial system (654 cases), colon (573 cases), esophagus (497 cases), stomach (422 cases), second malignancy and not specified lymph nodes (360 cases) and cervix (328 cases)) and lethality (pancreas (145 deaths; 57.1%), bronchi and lungs (527 deaths; 55.1%) and esophagus (262 deaths; 52.7%)), considering the competing events. In addition, we also evaluated how *gender* and *age* contribute to the risk of death from some cancers: women has bigger risk of death for esophageal cancer, while age was associated with the risk of death for prostate cancer. This study allowed characterizing the profile of cancers attended by the hospital by considering the competing events into the estimates methods. Due the advantages of the method, we recommend to researchers do not despise, in their studies, situations with competing events, since there are many *softwares* and materials available to help and facilitate its implementation

Keywords: Neoplasms; Mortality; Survival Analysis; Competing Events

1. APRESENTAÇÃO

Este trabalho consiste na dissertação de mestrado intitulada **Riscos competitivos: uma aplicação na sobrevida de pacientes com câncer**, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 30 de setembro de 2015. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos;
2. Artigo;
3. Conclusões e Considerações Finais.

Documentos de apoio estão apresentados nos anexos.

2. INTRODUÇÃO

No ano de 2012, ocorreram 14,1 milhões de novos casos de câncer, 8,2 milhões de mortes causadas por ele e 32,6 milhões de pessoas estavam vivendo com a doença (cinco anos após o diagnóstico) no mundo (Ferlay et al., 2012). De acordo com o GLOBOCAN (Ferlay et al., 2012), 57% dos novos casos de câncer, 65% das mortes por câncer e 48% das pessoas que convivem com a doença pertencem a regiões menos desenvolvidas. Além disso, segundo dados divulgados pelo Instituto Nacional de Câncer (INCA, 2014), a quantidade de novos casos continuará aumentando nos países em desenvolvimento e crescerá ainda mais em países desenvolvidos se medidas preventivas não forem amplamente aplicadas.

Dessa forma, a análise de dados dos pacientes com câncer torna-se uma ferramenta necessária para avaliar os programas de tratamento, monitorar o progresso das iniciativas de controle da doença e permitir que gestores da saúde tomem decisões mais acertadas.

Em relação à análise de dados, a mortalidade é um dos parâmetros utilizados para avaliar os resultados da área oncológica, sendo que dentre as abordagens tradicionais utilizadas na área encontram-se o método de Kaplan-Meier, que fornece uma visão descritiva do tempo de sobrevivência dos pacientes avaliados, e o modelo de Cox, que permite estimar o efeito de covariáveis. Em ambas é possível considerar apenas um evento de interesse (Carvalho et al., 2011).

Ocorre, porém, que, ao longo do período de acompanhamento, um paciente pode experimentar um evento diferente do de interesse. Considere, por exemplo, um indivíduo com câncer de mama que morre por outro tipo de câncer ou, ainda, por alguma causa não relacionada à doença. Esses outros possíveis eventos são denominados *eventos competitivos*. A utilização das abordagens tradicionais nesses casos resulta em estimativas superestimadas da função de distribuição acumulada e, consequentemente, em interpretações questionáveis (Balakrishnan e Rao, 2004; Kleinbaum e Klein, 2005).

Assim, na presença de riscos competitivos, metodologias alternativas às usuais devem ser utilizadas, tais como a função de incidência acumulada e o modelo da subdistribuição do risco.

Portanto, o objetivo principal deste trabalho é aplicar a metodologia de riscos competitivos para estimar a letalidade e fatores associados ao óbito de pacientes com câncer primário atendidos no HCPA entre os anos de 2002 e 2009. Como objetivos específicos, deseja-se tanto descrever a metodologia utilizada em análise de sobrevivência na presença de

riscos competitivos quanto avaliar incidências, letalidades e fatores associados ao óbito dos pacientes em acompanhamento usando a abordagem de riscos competitivos.

3. REVISÃO DE LITERATURA

3.1. ANÁLISE DE SOBREVIVÊNCIA

Análise de sobrevivência (sobrevida) é a técnica estatística utilizada quando o interesse consiste em modelar o tempo a partir de um ponto bem definido até a ocorrência de determinado evento (desfecho) cuja taxa de ocorrência ao longo do tempo não é constante. Esse evento pode ser a incidência de uma doença ou a morte de um paciente. Já o tempo refere-se à quantidade de horas, dias, meses, ou anos desde o início do período de acompanhamento (observação) até a ocorrência do evento de interesse (Kleinbaum e Klein, 2005). Na área da saúde, o início do acompanhamento pode ser exemplificado pela data do início de seguimento de um paciente em um estudo observacional, como um estudo de coorte.

O tempo observado até o evento é denominado **tempo de sobrevivência**, uma vez que diz respeito ao tempo que um indivíduo sobreviveu (viveu com a doença) ao longo do período de observação. Esse evento é comumente chamado de falha, dado que, normalmente, o interesse dessa abordagem é relacionado a algum aspecto negativo, tal como a morte (Collet, 2003; Kleinbaum e Klein, 2005).

Um problema enfrentado nesse tipo de análise é que dificilmente todos os indivíduos são acompanhados até a ocorrência do evento. As causas mais comuns de perda de informação são: término do período do estudo; perda de contato com o paciente; ou recusa do indivíduo em continuar participando do estudo. Esses casos são denominados **censura** (Carvalho et al., 2011; Collet, 2003).

O modelo de sobrevivência, da mesma forma que qualquer modelo de regressão, é composto por uma variável resposta, covariáveis, função de ligação e estrutura de erro (Carvalho et al., 2011). Nesse caso, a variável resposta é expressa como a probabilidade de sobrevivência, a taxa de incidência (função de risco) ou, ainda, a taxa de incidência acumulada.

Normalmente, a análise de sobrevivência é utilizada em situações onde há um único evento de interesse e que ocorre uma única vez, como, por exemplo, o diagnóstico de AIDS. Esses casos são os chamados **eventos únicos**. Há situações, porém, em que cada indivíduo pode experimentar um mesmo evento mais do que uma única vez (como é o caso de gestações, internações, cáries) ou eventos diferentes decorrentes de um mesmo fator de risco (como efeitos adversos de medicamentos ou doenças oportunistas do câncer) (Carvalho et al.,

2011). Esses são denominados **eventos múltiplos**. Um caso particular de eventos múltiplos ocorre quando a ocorrência de um evento exclui a ocorrência dos demais (Carvalho et al., 2011). Esses são classificados como **eventos competitivos**. Apenas o caso de eventos competitivos será tratado neste trabalho.

Portanto, as próximas seções têm como objetivo caracterizar a metodologia adequada para dois cenários: um único evento de interesse e eventos competitivos. Serão apresentadas as principais abordagens utilizadas em problemas de sobrevivência, iniciando com definições básicas e evoluindo para o modelo que considera eventos únicos e o modelo para eventos competitivos.

3.1.1. Principais definições

Nesta seção, serão definidos conceitos básicos em análise de sobrevivência: tipos de eventos; tipos de censura; funções que descrevem a sobrevivência (função de sobrevivência, função de risco e função de risco acumulado); e relações matemáticas entre elas.

Na definição dos diferentes tipos de eventos, temos a abordagem tradicionalmente utilizada para o modelo de sobrevivência onde se assume que somente um evento é considerado de interesse – evento único. Essa metodologia é, portanto, adequada para situações em que o evento de interesse ocorre uma única vez durante o período de acompanhamento, por exemplo: diagnóstico de um câncer primário ou de hipertensão; óbito por uma causa específica; entre outros.

Há, também, situações em que é possível observar o evento de interesse mais de uma vez durante o acompanhamento – eventos múltiplos. Exemplos disso são: número de internações; cáries; gestações; entre outros.

E há, ainda, ocasiões em que diversos eventos podem ocorrer, mas a partir do momento em que um deles foi observado, não há como verificar a ocorrência dos demais. Esse é o caso dos chamados eventos competitivos, situação em que a cada momento olha-se para um evento como o de interesse e para os demais como competitivos. Como exemplo, podemos citar os possíveis eventos de um paciente com câncer, que pode ser óbito por câncer *versus* o óbito por causa não relacionada ao câncer. Situações com eventos competitivos são a motivação do presente trabalho.

Um problema comum em estudos de coorte, delineamento em que geralmente se utiliza análise de sobrevivência, é que nem todos os pacientes seguem no estudo até a

ocorrência do evento de interesse e, muitas vezes, não se conhece a data exata de início do acompanhamento da coorte (Carvalho et al., 2011). Dentre as causas de perda de acompanhamento, cita-se o óbito por causa não relacionada ao estudo, perda de contato, recusa do indivíduo em continuar participando do estudo e o término do período de seguimento. Esses casos são denominados **censura**.

De forma simples, censura diz respeito à falta de observação da data de ocorrência do evento de interesse, podendo ser classificada como: à direita; à esquerda; e intervalar. **Censura à direita** ocorre quando, ao término do período de acompanhamento, não foi obervado o evento de interesse. Nesse caso, sabe-se que o tempo entre o início do estudo e o evento é maior que o tempo de acompanhamento. O desprezo dessa informação acarreta na superestimação do risco, pois, apesar do tempo de sobrevivência ser desconhecido, sabe-se que até o final do período de acompanhamento o paciente estava vivo (Carvalho et al., 2011; Kleinbaum e Klein, 2005). Já a **censura à esquerda** ocorre quando o evento de interesse ocorreu antes do fim do período de acompanhamento, porém não há informação sobre o momento exato de sua ocorrência (Carvalho et al., 2011; Kleinbaum e Klein, 2005). Ou seja, o tempo registrado é maior que o tempo de ocorrência do evento de interesse (Dos Santos Junior, 2012). Por fim, a **censura intervalar** ocorre quando se sabe apenas que o evento de interesse ocorreu em um determinado intervalo de tempo (Dos Santos Junior, 2012).

Pode-se, ainda, classificar censura como informativa ou não informativa. A primeira, **censura informativa**, ocorre se um paciente sai do estudo por uma razão relacionada ao evento estudado, por exemplo: abandono ao tratamento em decorrência de uma piora. A segunda, **censura não informativa**, é a suposição da maior parte dos métodos de análise de sobrevivência e considera que os tempos até o evento de interesse e até a censura sejam independentes. Ou seja, não há razão alguma para se suspeitar que o motivo da perda de informação esteja relacionado ao evento (Carvalho et al., 2011; Dos Santos Junior, 2012).

Análise de sobrevivência é a metodologia a ser empregada quando o interesse consistir em modelar a ocorrência de determinado evento ao longo do tempo, sendo que a taxa de ocorrência desse evento não é constante ao longo do tempo. Isso significa dizer que a média simples de eventos por unidade de pessoa-tempo não descreve de forma adequada o problema (Carvalho et al., 2011), uma vez que a probabilidade de um paciente vir a óbito por câncer de pulmão, por exemplo, é uma função que depende do tempo desde o diagnóstico.

Dessa forma, necessita-se das funções básicas de sobrevivência para responder a questões como: qual o tempo mediano de sobrevivência; qual o risco de um paciente

diagnosticado com câncer vir a óbito em até dois anos após diagnóstico; ou qual a probabilidade de que esse paciente sobreviva por mais de cinco anos após diagnóstico?

Antes da descrição das funções é necessário introduzir algumas notações e terminologias utilizadas nessa metodologia. Começaremos definindo $\boldsymbol{T}$ como uma variável aleatória que denota o tempo de sobrevivência de um indivíduo (Kleinbaum e Klein, 2005). Chamaremos de $\boldsymbol{t}$ um tempo específico, observável, da variável aleatória T. Por fim, denotaremos por $\boldsymbol{\delta}$ a variável aleatória binária que indica a ocorrência do evento, representada pelo valor 1, ou censura, representada pelo valor 0. Salienta-se que a suposição de censura não informativa deve ser satisfeita, ou seja, o valor de δ deverá ser igual a 0 somente se for verificada uma das seguintes condições: o indivíduo sobreviveu até o término do período de acompanhamento ou houve perda de acompanhamento do paciente por um motivo não relacionado ao evento de interesse (Kleinbaum e Klein, 2005).

Os próximos parágrafos descrevem as medidas de interesse em qualquer análise de sobrevivência: **função densidade de probabilidade**, $\boldsymbol{f(t)}$; **função de distribuição acumulada**, $\boldsymbol{F(t)}$; **função de sobrevivência**, $\boldsymbol{S(t)}$; **função de risco**, $\boldsymbol{h(t)}$; e **função de risco acumulado**, $\boldsymbol{H(t)}$.

Para definir a primeira função considera-se que o tempo de sobrevivência (T) é uma variável aleatória contínua e positiva cuja função densidade de probabilidade é definida por $f(t)$. Tal função representa a probabilidade instantânea de ocorrência do evento de interesse em um indivíduo (Carvalho et al., 2011). Por exemplo, podemos pensar na função densidade de probabilidade como a probabilidade de uma pessoa morrer no instante seguinte. Matematicamente, $f(t)$ é expressa por:

$$f(t) = \lim_{\varepsilon \to 0+} \frac{P\ (t \leq T < t + \varepsilon)}{\varepsilon}, \qquad (1)$$

onde ε representa um incremento de tempo infinitamente pequeno.

A segunda função básica é a função de distribuição acumulada, definida como a probabilidade de um evento ocorrer até o tempo t e matematicamente expressa por (Carvalho et al., 2011):

$$F(t) = P(T \leq t). \qquad (2)$$

A terceira função básica é a função de sobrevivência, através da qual se consegue responder a seguinte questão: qual a probabilidade de uma pessoa sobreviver por mais que um tempo específico t? De maneira formal, $S(t)$ fornece a probabilidade de que a variável aleatória T exceda um valor específico de tempo t, matematicamente representado por:

$$S(t) = P(T > t). \quad (3)$$

Através da equação 3 torna-se evidente que função de sobrevivência é simplesmente o complemento da função de distribuição acumulada (Carvalho et al., 2011). Além disso, teoricamente, como t assume qualquer valor maior ou igual a 0, a função de sobrevivência pode ser representada graficamente por uma curva suavizada com as seguintes propriedades (Carvalho et al., 2011; Kleinbaum e Klein, 2005): a) é não crescente, ou seja, decresce ou permanece constante conforme o aumento de t; b) em $t = 0$, o valor da função de sobrevivência é igual a 1. Isso significa que, ao início do estudo, quando nenhum evento de interesse foi observado, a probabilidade de sobreviver por mais que 0 unidades de tempo é igual a 1; c) em $t = \infty$, a probabilidade de sobrevivência é igual a 0, ou seja, à medida que o tempo passa, quando $t \to \infty$, a probabilidade de sobrevivência tende a ser nula.

Na prática, ao invés de curvas suavizadas, conforme mencionamos acima, é comum se obter gráficos com formato de função escada. Além disso, como o período de duração de um estudo nunca é infinito e pode haver riscos competitivos para as falhas, é possível que nem todas as pessoas observadas sofram o evento de interesse. Dessa forma, a função de sobrevivência estimada (expressa por $\hat{S}(t)$) pode não chegar a 0 ao término do estudo.

A quarta função básica, função de risco (*hazard function*), é matematicamente definida por (Kleinbaum e Klein, 2005):

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \quad (4)$$

e fornece o risco instantâneo de um indivíduo sofrer o evento de interesse entre o tempo t e $t + \Delta t$, dado que ele sobreviveu até o tempo t (Carvalho et al., 2011). Salienta-se que, diferentemente da função de sobrevivência, na função de risco o foco está nos casos em que o evento de interesse ocorreu (Kleinbaum e Klein, 2005).

Na equação 4, o numerador (probabilidade condicional) fornece a probabilidade de ocorrer o evento no intervalo de tempo entre t e $t + \Delta t$, dada sua não ocorrência no tempo anterior a t; o denominador representa um pequeno intervalo de tempo; e o limite dessa razão, com o intervalo de tempo tendendo a 0, fornece a probabilidade instantânea de falha no tempo t por unidade de tempo.

Por risco instantâneo entende-se o risco, em um dado momento, de ocorrer o evento em estudo sabendo que o indivíduo sobreviveu até o tempo t, conforme é expresso pelo numerador da equação 4. Por essa razão, a função de risco também é conhecida como força de mortalidade, taxa de mortalidade condicional ou taxa de mortalidade específica por idade (Kleinbaum e Klein, 2005; Lee, 2003).

Convém ressaltar que $h(t)$ expressa uma taxa e não uma probabilidade (Carvalho et al., 2011), pois o numerador dessa função é uma probabilidade condicional e o denominador representa um pequeno intervalo de tempo. Dessa razão obtém-se uma probabilidade por intervalo de tempo, que deixa de ser uma probabilidade (uma vez que pode assumir qualquer valor real maior do que 0, deixando de estar restrita ao intervalo $[0,1]$) e passa a representar uma taxa (Carvalho et al., 2011; Kleinbaum e Klein, 2005).

Da mesma forma que $S(t)$, a função de risco também pode ser representada graficamente, podendo iniciar em qualquer valor e subir ou descer em qualquer direção ao longo do tempo. Teoricamente, para um valor de t específico, $h(t)$ tem as seguintes características: é sempre não negativa e não tem um limite superior (Kleinbaum e Klein, 2005).

A quinta e última função básica de sobrevivência é a função de risco acumulado, $H(t)$, que mede o risco de ocorrência de certo evento de interesse até um determinado tempo t (Carvalho et al., 2011). Matematicamente essa função é definida por:

$$H(t) = \int_0^t h(u)d(u). \qquad (5)$$

Reforça-se que, da mesma forma que a função de risco, a função de risco acumulado representa uma taxa e não uma probabilidade (Carvalho et al., 2011).

Independente de qual função básica de sobrevivência seja do seu interesse, existe uma relação entre todas elas. Isso significa que ao conhecer a função de sobrevivência, por

exemplo, é possível estimar também as funções de risco e risco acumulado (Kleinbaum e Klein, 2005; Carvalho et al., 2011).

As relações entre as funções são dadas por (Carvalho et al., 2011; Collet, 2003):

$$S(t) = 1 - F(t); \quad (6)$$

$$S(t) = exp\big(-H(t)\big); \quad (7)$$

$$h(t) = -\frac{d\ ln\ (S(t))}{dt}; \quad (8)$$

$$h(t) = \frac{f(t)}{S(t)}; \quad (9)$$

$$H(t) = -\ ln\big(S(t)\big). \quad (10)$$

Como se pode ver a função de sobrevivência é obtida tanto através do complemento da função de distribuição acumulada quanto através da exponencial da função de risco acumulado. A função de risco é obtida através da derivada, em relação à t, do valor negativo do logaritmo natural da função sobrevivência ou pela razão entre as funções densidade de probabilidade e sobrevivência. E a função de risco acumulado é obtida através do valor negativo do logaritmo natural da função de sobrevivência.

3.1.2. Análise de sobrevivência considerando eventos únicos e eventos competitivos

No tópico anterior foram definidas as funções e conceitos necessários para avaliar o tempo de sobrevivência. Agora, serão abordadas as formas de estimação das funções básicas de sobrevivência na presença de observações censuradas considerando dois cenários: o primeiro considera apenas um evento de interesse (eventos únicos) e o segundo considera situações em que há eventos competitivos. Além disso, para cada um deles será considerada a inclusão de covariáveis na modelagem das funções básicas de sobrevivência.

3.1.2.1. Eventos únicos

Esta seção busca detalhar as duas abordagens comumente empregadas em estudos de avaliação do tempo de sobrevivência. Os métodos descritos a seguir diferem de acordo com a função básica de sobrevivência estimada e como eles incorporam covariáveis. Iniciar-se-á com a descrição do método de Kaplan-Meier, que é útil como uma análise descritiva do

tempo de sobrevivência da amostra, permitindo selecionar uma determinada característica (covariável) do paciente e testar a igualdade das curvas de sobrevivência entre os níveis dessa característica. Em seguida, será apresentado o modelo de Cox, que permite explorar o relacionamento de diversas características, denominadas **covariáveis**, com a sobrevivência através da estimação do efeito das covariáveis sobre a variável resposta (Carvalho et al., 2011; Collet, 2003).

3.1.2.1.1. Método de Kaplan-Meier

Anteriormente, foram citadas as funções básicas de sobrevivência e um problema comum em estudos desse tipo: a presença de observações censuradas. Agora, será abordado um método não paramétrico para estimar a função de sobrevivência, $S(t)$, que permite a presença de observações censuradas e a comparação entre níveis de uma única covariável. Trata-se do método de Kaplan-Meier, que é dito não paramétrico porque não supõe uma forma específica para $f(t)$.

Em uma situação em que não há censura (por exemplo: em uma coorte de pacientes com câncer de pulmão que é acompanhada até que se verifique o óbito de todos os indivíduos), a função de sobrevivência pode ser estimada, em cada tempo t em que ocorreu o evento de interesse, pela proporção de pessoas que sobreviveram além do tempo t. Matematicamente expressamos isso por:

$$\hat{S}(t) = \frac{R(t)}{N}, \qquad (11)$$

onde $R(t)$ representa o número de pessoas em risco no início do intervalo de tempo t e N é o número total de indivíduos no início da coorte (Carvalho et al., 2011).

Imagine agora uma situação onde determinado paciente tenha sido censurado no 36º dia de acompanhamento. Até a data de sua saída sabia-se que ele estava vivo e poderia, portanto, ser incluído tanto no numerador (pois era um sobrevivente nesse período) quanto no denominador (pois fazia parte do grupo em risco) da equação 11 para calcular a função de sobrevivência para $t < 36$. Porém, a partir do 36º dia não haveria mais qualquer garantia de que o indivíduo ainda estivesse vivo, dado que ele deixou o grupo de acompanhamento e, pelo mesmo motivo, se tornaria incorreto incluí-lo no grupo em risco (Carvalho et al., 2011).

Para contornar problemas como esse, foi proposto o **método de Kaplan-Meier**, também conhecido como **estimador produto-limite**, que permite estimar a função de sobrevivência na presença de censura incorporando as informações de todas as observações disponíveis, tanto as censuradas quanto as não censuradas, através da utilização dos conceitos de eventos independentes e probabilidade condicional (Carvalho et al., 2011; Hosmer e Lemeshow, 1999).

Para entender matematicamente esse método, considere uma amostra com n indivíduos independentes representada por (t_i, δ_i), $i = 1, 2, \dots, n$, sendo T a variável tempo de sobrevivência e δ, a variável indicadora de censura. Na prática, dizemos que t_i refere-se ao tempo de acompanhamento até a ocorrência, ou não, do evento de interesse do indivíduo i (de uma amostra que possui n pacientes) e que δ_i assume o valor 1 se o evento de interesse for observado para o paciente i, e 0 caso contrário. Assuma que entre os n indivíduos existam $m \leq n$ tempos distintos de ocorrência do evento em estudo. Representam-se esses tempos de sobrevivência ordenados de forma crescente por $t_1 < t_2 < \cdots < t_m$, o número de pessoas em risco de morte no tempo t_j por $R(t_j)$ e o número observado de mortes em t_j por $\Delta \boldsymbol{N}(\boldsymbol{t_j})$. Assim, o estimador de Kaplan-Meier para a função de sobrevivência no tempo t é obtido a partir da seguinte equação (Hosmer e Lemeshow, 1999):

$$\hat{S}(t_j) = \frac{R(t_1) - \Delta N(t_1)}{R(t_1)} \times \frac{R(t_2) - \Delta N(t_2)}{R(t_2)} \times \dots \times \frac{R(t_m) - \Delta N(t_m)}{R(t_m)}$$

$$\hat{S}(t_j) = \prod_{j:t_j \leq t} \frac{R(t_j) - \Delta N(t_j)}{R(t_j)} = \hat{S}(t_j - 1) \times \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}. \quad (12)$$

Dessa maneira, para os m tempos t_j em que ocorrer um evento, a probabilidade de sobrevivência será estimada pelo número de sobreviventes no tempo t_j dividido pelo número de pessoas que estavam em risco naquele momento (Carvalho et al., 2011; Hosmer e Lemeshow, 1999). Salienta-se que, uma vez estimada a função de sobrevivência, qualquer outra função básica pode ser estimada utilizando as relações matemáticas apresentadas na seção 1.1.

Além de explorar a função de sobrevivência de uma amostra, grande parte dos estudos se interessam em verificar se características de um paciente, tais como sexo ou faixa etária, afetam seu tempo de sobrevivência. Para exemplificar, imagine um estudo que esteja acompanhando uma coorte de pacientes com câncer de estômago e deseja comparar homens e

mulheres. Assim, há dois grupos a serem analisados: os pacientes do sexo feminino e os do sexo masculino. A curva de sobrevivência será estimada separadamente para cada um dos grupos utilizando o método de Kaplan-Meier descrito anteriormente. Como resultado, será possível ter uma ideia do comportamento semelhante, ou não, da função de sobrevivência entre os grupos comparados.

Essa ideia de comportamento semelhante utiliza testes de hipóteses para verificar se as curvas de sobrevivência são, de fato, estatisticamente significativas. Dentre esses testes pode-se citar log-rank, Wilcoxon, Tarone-Ware, Peto e Flemington-Harrington (Kleinbaum e Klein, 2005). Um dos mais conhecidos, porém, é o de Mantel-Haenzel, também chamado de **teste log-rank**, cuja hipótese nula é:

$$H_0: h_1(t) = h_2(t) = \cdots = h_k(t),$$

sendo que k representa o número de grupos comparados e $h_k(t)$ pode ser estimado a partir da relação entre as funções básicas de sobrevivência (Carvalho et al., 2011).

O teste utiliza a comparação entre os valores observados e os esperados para cada grupo supondo que o risco é o mesmo em todos (Carvalho et al., 2011; Kleinbaum e Klein, 2005). Rejeitar a hipótese nula, portanto, significa concluir que pelo menos uma das curvas é significativamente diferente das demais em algum momento do tempo. Já a não rejeição indica que a variável testada (sexo, considerando o exemplo referido anteriormente) não afeta a sobrevivência.

Assim, para realizar o teste deve-se: estimar o número de eventos esperados para cada estrato k, definido por $E_k(t)$, segundo a hipótese nula de incidência igual em todos os estratos; calcular a estatística de teste que, sob H_0, segue uma distribuição χ^2com $(k-1)$ graus de liberdade.

O número de eventos esperados para cada estrato k é matematicamente definido por (Carvalho et al., 2011):

$$E_k(t) = \Delta N(t)\frac{R_k(t)}{R(t)}, \qquad (13)$$

onde $\Delta N(t)$ representa o número total de eventos observados em t; $R_k(t)$ representa o número de indivíduos em risco no grupo k no tempo t; e $R(t)$ representa o número total de indivíduos em risco no tempo t.

Por fim, a estatística de teste é expressa da seguinte forma (Carvalho et al., 2011):

$$log - rank = \sum_{t} \frac{(O_k - E_k)^2}{Var(O_k - E_k)} \quad (14)$$

onde O_k representa o número total de eventos observados no grupo k e E_k o número total de eventos esperados para o grupo k. Além disso, considerando dois grupos (ou seja, $k = 2$), a variância é obtida da seguinte forma:

$$Var(O_k - E_k) = \sum_{t} \frac{R_1(t)R_2(t)\Delta N(t)[R(t) - \Delta N(t)]}{R(t)^2[R(t) - 1]}. \quad (15)$$

Para mais de dois grupos o cálculo da estatística de teste envolve variâncias e covariâncias, sendo que detalhes matemáticos são encontrados em Kleinbaum e Klein, 2005, por exemplo.

3.1.2.1.1.1. Praticando os conceitos abordados

Neste tópico, apresentaremos um passo a passo da estimação da função de sobrevivência tanto na ausência quanto na presença de observações censuradas e exemplificaremos a comparação de curvas de sobrevivência através do teste de log-rank.

Consideremos que os dados apresentados abaixo se referem ao tempo de sobrevivência ordenado, em anos, de 15 pacientes com câncer de pulmão, sendo que todos foram acompanhados até o óbito – ou seja, não há censura.

Tabela 1 - Tempos de sobrevivência, em anos, de 15 pacientes com câncer de pulmão

0,05	0,07	0,08	0,08	0,23	0,26	0,33	0,58	1,11	1,17	2,12	2,28	2,29	2,34	3,28

Visto que não há censura, utilizaremos a equação 11 para calcular a função de sobrevivência. Para tanto, iniciamos organizando a Tabela 1 com os intervalos de ocorrência

dos eventos e o número de pacientes em risco no início de cada intervalo de tempo x (Tabela 2).

Em seguida, basta aplicar a equação 11 para calcular a função de sobrevivência em cada intervalo. Assim, por exemplo, considerando o intervalo (0,26; 0,33], a sobrevivência é calculada como:

$$\hat{S}_{(0,26;0,33]}(t) = \frac{R_{(0,26;0,33]}(t)}{N} = \frac{9}{15} = 0{,}60.$$

Isso significa que a probabilidade de um paciente com câncer de pulmão sobreviver por mais do que 0,26 anos (3,16 meses) é 0,60. O mesmo raciocínio deve ser utilizado para os demais intervalos de tempo, cujas respectivas sobrevivências também encontram-se na Tabela 2.

Tabela 2 - Organização dos dados da Tabela 1

Intervalo de tempo de sobrevivência (anos)	t_j	$\Delta N(t)$	$\mathbf{R(t)}$	$\hat{\mathbf{S}}(\mathbf{t}) = \prod_{t_j \le t} \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}$
(0,00; 0,05]	0,05	1	15	1,00
(0,05; 0,07]	0,07	1	14	0,93
(0,07; 0,08]	0,08	2	13	0,87
(0,08; 0,23]	0,23	1	11	0,73
(0,23; 0,26]	0,26	1	10	0,67
(0,26; 0,33]	0,33	1	9	0,60
(0,33; 0,58]	0,58	1	8	0,53
(0,58; 1,11]	1,11	1	7	0,47
(1,11; 1,17]	1,17	1	6	0,40
(1,17; 2,12]	2,12	1	5	0,33
(2,12; 2,28]	2,28	1	4	0,27
(2,28; 2,29]	2,29	1	3	0,20
(2,29; 2,34]	2,34	1	2	0,13
(2,34; 3,28]	3,28	1	1	0,07

Agora, consideremos os dados de 30 pacientes também com câncer de pulmão, cujo diagnóstico foi realizado entre os anos de 2002 e 2009 e o acompanhamento realizado até 2013 (dados obtidos do presente trabalho). Dentre eles, nem todos faleceram até o término do período de acompanhamento, ou seja, há casos de censura. Os dados estão na Tabela 3 e apresentam, para cada paciente: uma variável que indica a qual paciente refere-se à informação; o sexo do indivíduo, sendo *Mas*, masculino e *Fem*, feminino; o tempo de

sobrevivência ordenado, em anos; e uma variável indicadora da ocorrência do evento, cujo valor é igual a 1 caso o óbito tenha ocorrido e 0, caso contrário.

Tabela 3 – Sexo e tempo de sobrevivência de 30 pacientes com câncer de pulmão

Paciente	Sexo	Tempo de sobrevivência (anos)	δ
1	Mas	0,02	1
2	Mas	0,10	1
3	Fem	0,12	1
4	Mas	0,14	1
5	Fem	0,15	1
6	Fem	0,17	1
7	Mas	0,18	1
8	Mas	0,19	1
9	Mas	0,25	1
10	Fem	0,39	1
11	Mas	0,51	1
12	Mas	0,60	1
13	Mas	0,67	1
14	Fem	0,72	1
15	Mas	0,73	1
16	Fem	0,87	1
17	Fem	0,96	1
18	Fem	1,11	1
19	Fem	1,14	1
20	Mas	1,26	1
21	Fem	4,03	0
22	Mas	5,05	0
23	Mas	5,23	0
24	Fem	5,48	0
25	Fem	6,68	0
26	Mas	6,73	0
27	Mas	7,97	0
28	Mas	8,51	0
29	Fem	11,37	0
30	Fem	11,71	0

Desses 30 pacientes, são 16 homens, 14 mulheres e 20 óbitos. Em função da presença de observações censuradas, a estimação da função de sobrevivência será realizada utilizando o método de Kaplan-Meier, conforme a equação 12.

Iniciaremos determinando: o número de pessoas em risco no tempo t_j, j: $1,\ldots,30$ e o número de eventos ocorridos exatamente em t_j. O número de pessoas em risco é definido

como o conjunto de pacientes que sobreviveram pelo menos até o tempo t_j, sendo que no tempo $t_j = 0$ (momento em que o indivíduo passou a ser acompanhado) o conjunto em risco é igual ao total de pacientes ($R(0) = 30$) e a probabilidade desses pacientes sobreviverem por mais do que 0 dia é igual a 1 ($S(0) = 1$). Essa situação permanece até que ocorra o primeiro óbito, pois é quando o grupo em risco tem seu valor alterado e $S(t)$ passa a ser calculada como a razão entre o número de sobreviventes e o tamanho do grupo que estava em risco antes do evento ocorrer. Esses passos para o cálculo são apresentados na Tabela 4.

O método de Kaplan-Meier gera uma função escada que salta em cada tempo onde ocorreu um evento, sendo que o tamanho desse salto depende do número de eventos observados nesse tempo e também do número de observações censuradas (representadas no gráfico por um "+") antes dele. O termo função escada deve-se ao fato de que o risco se mantém constante até a ocorrência do próximo evento de interesse, nesse caso, o óbito (Carvalho et al., 2011).

Utilizaremos o *software* estatístico R (R Development Core Team, 2012) para demonstrar como fica o gráfico da função de sobrevivência de acordo com os dados da Tabela 4. R é um programa gratuito e de código aberto que trabalha através de bibliotecas, sendo *survival* a necessária para proceder à análise de interesse (Terry Therneau (2012). *A Package for Survival Analysis in R*. R package version 2.36-14.). Tanto o *software* quanto a biblioteca estão disponíveis no endereço http://www.r-project.org/.

Informações acerca da instalação do R e *survival* estão disponíveis em http://cran.r-project.org/doc/manuals/r-release/R-admin.html; informações sobre a importação de banco de dados encontram-se em http://cran.r-project.org/doc/manuals/r-release/R-data.html; informações referentes à manipulação de banco de dados podem ser obtidas em http://cran.r-project.org/doc/manuals/r-release/R-intro.html; um tutorial em português encontra-se em http://leg.ufpr.br/~paulojus/.

Tabela 4 – Aplicação do método de Kaplan-Meier considerando os dados da Tabela 3

Intervalo de tempo de sobrevivência (anos)	$\mathbf{t_j}$	$\mathbf{R(t_j)}$	$\mathbf{\Delta N(t_j)}$	$\hat{\mathbf{S}}(t) = \prod_{j:t_j \leq t} \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}$
	0,00	30	0	1,000
(0,00; 0,02]	0,02	30	1	$1 \times \left(\frac{30-1}{30}\right) = 0{,}967$
(0,02; 0,10]	0,10	29	1	$0{,}967 \times \left(\frac{29-1}{29}\right) = 0{,}933$
(0,10; 0,12]	0,12	28	1	$0{,}933 \times \left(\frac{28-1}{28}\right) = 0{,}900$
(0,12; 0,14]	0,14	27	1	0,867
(0,14; 0,15]	0,15	26	1	0,833
(0,15; 0,17]	0,17	25	1	0,800
(0,17; 0,18]	0,18	24	1	0,767
(0,18; 0,19]	0,19	23	1	0,733
(0,19; 0,25]	0,25	22	1	0,700
(0,25; 0,39]	0,39	21	1	0,667
(0,39; 0,51]	0,51	20	1	0,633
(0,51; 0,60]	0,60	19	1	0,600
(0,60; 0,67]	0,67	18	1	0,567
(0,67; 0,72]	0,72	17	1	0,533
(0,72; 0,73]	0,73	16	1	0,500
(0,73; 0,87]	0,87	15	1	0,467
(0,87; 0,96]	0,96	14	1	0,433
(0,96; 1,11]	1,11	13	1	0,400
(1,11; 1,14]	1,14	12	1	0,367
(1,14; 1,26]	1,26	11	1	0,333
(1,26; 4,03]	4,03	10	1	0,300
(4,03; 5,05]	5,05	9	1	0,267
(5,05; 5,23]	5,23	8	1	0,233
(5,23; 5,48]	5,48	7	1	0,200
(5,48; 6,68]	6,68	6	1	0,167
(6,68; 6,73]	6,73	5	1	0,133
(6,73; 7,97]	7,97	4	1	0,100
(7,97; 8,51]	8,51	3	1	0,067
(8,51; 11,37]	11,37	2	1	0,033
(11,37; 11,71]	11,71	1	1	0,000

Para gerar a função de sobrevivência pelo método de Kaplan-Meier utiliza-se o comando:

```
survfit(formula = Surv(tempo, status) ~ 1, type = "kaplan-meier",
                  data = dados)
```

onde *tempo* refere-se à coluna com informações de tempo de sobrevivência; *status* refere-se à coluna com a variável binária que indica o evento sofrido pelo paciente (óbito ou censura); *type* indica o método utilizado para estimar a curva de sobrevivência (ver manual); *data* é o nome do banco de dados que possui as colunas *tempo* e *status*.

Como resultado, obtivemos:

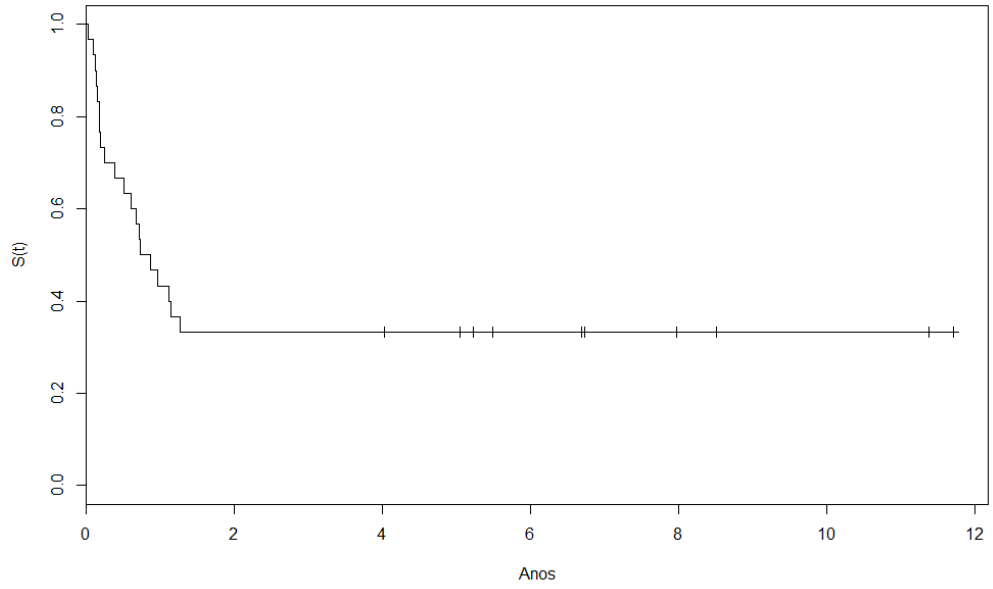


Figura 1 - Função de sobrevivência dos pacientes com câncer de pulmão (Tabelas 3 e 4) estimada via método de Kaplan-Meier

Continuemos considerando os dados da Tabela 3 para verificar se há diferença nas curvas de sobrevivência dos pacientes de acordo com o sexo. Testar se curvas de sobrevivência são iguais equivale a testar se a incidência de eventos é semelhante em cada grupo. Se for, a curva de sobrevivência será a mesma (Carvalho et al., 2011). Para tal, utilizaremos o teste de log-rank, cuja hipótese nula a ser testada é:

$$H_0: h_{Mas}(t) = h_{Fem}(t)$$

Para começar, olharemos de forma separada para os tempos de sobrevivência e ocorrência do óbito em cada sexo (valores extraídos da Tabela 3). As observações censuradas estão representadas pelo sinal "+" ao lado do tempo de sobrevivência.

Sexo masculino (Mas):

0,02	0,10	0,14	0,18	0,19	0,25	0,51	0,60
0,67	0,73	1,26	5,05+	5,23+	6,73+	7,97+	8,51+

Sexo feminino (Fem):

0,12	0,15	0,17	0,39	0,72	0,87	0,96	1,11
1,14	4,03+	5,48+	6,68+	11,37+	11,71+		

O próximo passo consiste em calcular o número de óbitos esperados para cada sexo. Para tal, deve-se definir: número total de óbitos observados em t ($\Delta N(t)$); número total de pessoas em risco em cada sexo no tempo t ($R_{Mas}(t)$ e $R_{Fem}(t)$); número total de pessoas em risco no tempo t ($R(t)$). A partir desses valores, calcula-se o número de óbitos esperado para cada sexo utilizando a equação 13, conforme mostra a Tabela 5.

Tabela 5 - Valores observados e esperados para número de óbitos considerando os sexos de 30 pacientes com câncer de pulmão

t_j	**Eventos observados**			**Grupo em risco**			**Eventos esperados** $E_k(t) = \Delta N(t)\frac{R_k(t)}{R(t)}$		
	$\Delta N_{Mas}(t)$	$\Delta N_{Fem}(t)$	$\Delta N(t)$	$R_{Mas}(t)$	$R_{Fem}(t)$	$R(t)$	$E_{Mas}(t)$	$E_{Fem}(t)$	$E(t)$
0,00	0	0	0	16	14	30	0	0	0
0,02	1	0	1	16	14	30	$1 \times \left(\frac{16}{30}\right)$=0,53	$1 \times \left(\frac{14}{30}\right)$=0,47	0,53 + 0,47 = 1
0,10	1	0	1	15	14	29	$1 \times \left(\frac{15}{29}\right)$=0,52	$1 \times \left(\frac{14}{29}\right)$=0,48	0,52 + 0,48 = 1
0,12	0	1	1	14	14	28	0,50	0,50	1
0,14	1	0	1	14	13	27	0,52	0,48	1
0,15	0	1	1	13	13	26	0,50	0,50	1
0,17	0	1	1	13	12	25	0,52	0,48	1
0,18	1	0	1	13	11	24	0,54	0,46	1
0,19	1	0	1	12	11	23	0,52	0,48	1
0,25	1	0	1	11	11	22	0,50	0,50	1
0,39	0	1	1	10	11	21	0,48	0,52	1
0,51	1	0	1	10	10	20	0,50	0,50	1
0,60	1	0	1	9	10	19	0,47	0,53	1
0,67	1	0	1	8	10	18	0,44	0,56	1
0,72	0	1	1	7	10	17	0,41	0,59	1
0,73	1	0	1	7	9	16	0,44	0,56	1
0,87	0	1	1	6	9	15	0,40	0,60	1
0,96	0	1	1	6	8	14	0,43	0,57	1
1,11	0	1	1	6	7	13	0,46	0,54	1
1,14	0	1	1	6	6	12	0,50	0,50	1
1,26	0	1	1	6	5	11	0,55	0,45	1
4,03	0	0	0	5	5	10	0,00	0,00	0
5,05	0	0	0	5	5	10	0,00	0,00	0
5,23	0	0	0	5	5	10	0,00	0,00	0
5,48	0	0	0	5	5	10	0,00	0,00	0
6,68	0	0	0	5	5	10	0,00	0,00	0
6,73	0	0	0	5	5	10	0,00	0,00	0
7,97	0	0	0	5	5	10	0,00	0,00	0
8,51	0	0	0	5	5	10	0,00	0,00	0
11,37	0	0	0	5	5	10	0,00	0,00	0
11,71	0	0	0	5	5	10	0,00	0,00	0
Total	**11**	**9**	**20**				**9,73**	**10,27**	**20**

A Tabela 5 fornece o numerador da estatística de teste de log-rank. Para obter o denominador, é necessário calcular a variância da diferença entre os valores observados e esperados para cada tempo t, conforme a equação 15. Os passos necessários para o cálculo são apresentados na Tabela 6.

Tabela 6 - Etapas para o cálculo do denominador do teste de log-rank

t_j	Eventos observados			Grupo em risco			$Var(O_k - E_k) = \sum_t \frac{R_{Mas}(t)R_{Fem}(t)\Delta N(t)[R(t) - \Delta N(t)]}{R(t)^2[R(t) - 1]}$
	$\Delta N_{Masc}(t)$	$\Delta N_{Fem}(t)$	$\Delta N(t)$	$R_{Mas}(t)$	$R_{Fem}(t)$	$R(t)$	
0,00	0	0	0	16	14	30	$\frac{16 \times 14 \times 0 \times [30 - 0]}{30^2 \times [30 - 1]} = 0$
0,02	1	0	1	16	14	30	$\frac{16 \times 14 \times 1 \times [30 - 1]}{30^2 \times [30 - 1]} = 0{,}25$
0,10	1	0	1	15	14	29	0,25
0,12	0	1	1	14	14	28	0,25
0,14	1	0	1	14	13	27	0,25
0,15	0	1	1	13	13	26	0,25
0,17	0	1	1	13	12	25	0,25
0,18	1	0	1	13	11	24	0,25
0,19	1	0	1	12	11	23	0,25
0,25	1	0	1	11	11	22	0,25
0,39	0	1	1	10	11	21	0,25
0,51	1	0	1	10	10	20	0,25
0,60	1	0	1	9	10	19	0,25
0,67	1	0	1	8	10	18	0,25
0,72	0	1	1	7	10	17	0,24
0,73	1	0	1	7	9	16	0,25
0,87	0	1	1	6	9	15	0,24
0,96	0	1	1	6	8	14	0,24
1,11	0	1	1	6	7	13	0,25
1,14	0	1	1	6	6	12	0,25
1,26	0	1	1	6	5	11	0,25
4,03	0	0	0	5	5	10	0,00
5,05	0	0	0	5	5	10	0,00
5,23	0	0	0	5	5	10	0,00
5,48	0	0	0	5	5	10	0,00
6,68	0	0	0	5	5	10	0,00
6,73	0	0	0	5	5	10	0,00
7,97	0	0	0	5	5	10	0,00
8,51	0	0	0	5	5	10	0,00
11,37	0	0	0	5	5	10	0,00
11,71	0	0	0	5	5	10	0,00
Total	**11**	**9**	**20**				**4,96**

Quando apenas dois estratos são comparados, como nesse caso, a estatística de log-rank é calculada utilizando os dados de apenas um dos estratos, já que, por simetria, o resultado equivale para o outro grupo. Agora, com todos os elementos calculados, podemos obter o valor da estatística de teste, conforme a equação 14, considerando o sexo feminino:

$$log - rank = \frac{(O_{Fem} - E_{Fem})^2}{Var(O_{Fem} - E_{Fem})} = \frac{(9 - 10{,}27)^2}{4{,}96} = 0{,}32.$$

Lembrando que a estatística de teste segue uma distribuição χ^2 com $k-1$ graus de liberdade, compararemos os valores calculado (obtido acima) e tabelado para obter uma conclusão em relação à hipótese nula.

Com auxílio de uma tabela da distribuição χ^2 obtemos que o valor tabelado para $\chi^2_{(1)}$, considerando 5% de nível de significância, é 3,84. Comparando com o valor calculado, decide-se não rejeitar a hipótese nula (valor de p = 0,569). Ou seja, com 5% de significância, conclui-se que não há diferença estatisticamente significativa entre os tempos de sobrevivência de pacientes com câncer de pulmão dos sexos feminino e masculino.

O teste de log-rank também pode ser feito utilizando o pacote *survival* do *software* R através do seguinte comando:

```
survdiff(Surv(tempo, status) ~ cov, data = dados)
```

onde *cov* refere-se a variável que se deseja testar a diferença entre as curvas. Considerando os dados desse exemplo, a saída do R (Figura 2) mostra: número de pessoas em risco em cada estrato no início do estudo (N); número total de eventos observados (*Observed*) e esperados (*Expected*) em cada estrato; uma estatística de teste aproximada ($(O-E)^2/E$); e a estatística de log-rank ($(O-E)^2/V$).

```
> survdiff(Surv(Sobr, Censura)~Sexo,data = dados)
Call:
survdiff(formula = Surv(Sobr, Censura) ~ Sexo, data = dados)

                N Observed Expected (O-E)^2/E (O-E)^2/V
Sexo=Feminino  14        9    10.27     0.157     0.324
Sexo=Masculino 16       11     9.73     0.165     0.324

 Chisq= 0.3  on 1 degrees of freedom, p= 0.569
```

Figura 2 - Saída do R para teste de log-rank considerando os dados da Tabela 3

O resultado do teste é apresentado na última linha e indica a não rejeição da hipótese nula (p = 0,569), considerando 5% de nível de significância.

3.1.2.1.2. Modelo de Cox

Frequentemente, o objetivo dos pesquisadores vai além da estimação da sobrevivência. O maior interesse consiste em explorar o relacionamento de diversas

características (**covariáveis**) com a sobrevivência (Carvalho et al., 2011; Collet, 2003). Isso é possível através de modelos de sobrevivência que permitem estimar o efeito das covariáveis sobre a variável resposta, sendo esse o foco do presente tópico.

As duas razões para modelar dados de sobrevivência são: determinar qual combinação de covariáveis afeta a forma da função de risco e obter um risco estimado para um indivíduo com dadas características.

Nesta seção, abordaremos o modelo de riscos proporcionais, proposto por Sir David Cox (1972), e, em função disso, chamado de **modelo de Cox**. Esse método não faz qualquer suposição sobre a distribuição de probabilidade do tempo de sobrevivência, sendo por isso denominado de modelo semiparamétrico (Collet, 2003). Assume-se apenas que as covariáveis agem multiplicativamente sobre o risco, sendo essa a parte paramétrica do modelo (Carvalho et al., 2011).

O modelo de riscos proporcionais de Cox modela a função de risco $h(t, \boldsymbol{X})$ da seguinte maneira (Collet, 2003; Kleinbaum e Klein, 2005):

$$h(t, \boldsymbol{X}) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i} \tag{16}$$

onde t representa um instante de tempo específico; $\boldsymbol{X} = \left(X_1, X_2, \dots, X_p\right)$ representa um conjunto de covariáveis; $h_0(t)$ indica a função de risco basal e β_i, o coeficiente associado à covariável X_i. Ou seja, a fórmula do modelo de Cox indica que o risco, no tempo t, é igual ao produto de duas quantidades: a função de risco basal e a exponencial da soma de $\beta_i X_i$, considerando p covariáveis.

O modelo de Cox tem a propriedade de que quando todos os valores das covariáveis forem nulos, a equação 16 se reduz para a função de risco basal, sendo essa a razão pela qual $h_0(t)$ é chamada de função basal. Outra propriedade desse modelo é que $h_0(t)$ é uma função não especificada, sendo essa a propriedade que faz do modelo de Cox um modelo semiparamétrico (Kleinbaum e Klein, 2005).

O termo riscos proporcionais, suposição desse modelo, significa que a razão entre o risco de ocorrência do evento para dois indivíduos, considerando as mesmas covariáveis $\boldsymbol{x}$, é constante no tempo. Na prática, isso quer dizer que, por exemplo, o risco de desenvolver câncer de pulmão dentre os pacientes que fumam é sempre o mesmo ao longo do tempo. Alguns desenvolverão a doença mais cedo, outros mais tarde, porém sempre na mesma

proporção que é estimada pela exponencial do coeficiente ($exp(\beta_{Fumo})$) (Carvalho et al., 2011).

Também se pode escrever o modelo de Cox em termos das funções de risco acumulado ou sobrevivência da seguinte forma (Carvalho et al., 2011):

$$H(t, \boldsymbol{X}) = H_0(t) e^{\sum_{i=1}^{p} \beta_i X_i} \tag{17}$$

$$S(t, \boldsymbol{X}) = S_0(t) e^{\sum_{i=1}^{p} \beta_i X_i} \tag{18}$$

onde a estimação das funções de risco acumulado basal e sobrevivência basal é feita utilizando a relação entre as funções apresentadas na seção 1.1.

O foco da discussão, a partir desse momento, será na forma como são estimados os parâmetros ($\boldsymbol{\beta}$) que pertencem à parte paramétrica desse modelo. Os valores estimados, representados por $\hat{\beta}_i$, são chamados de estimativas. Uma das maneiras de estimar um parâmetro é pelo método da verossimilhança. Isso significa dizer que os parâmetros são obtidos maximizando a **função de verossimilhança**, representada por L. A ideia central da função de verossimilhança é descrever a probabilidade de se obter os dados observados nos sujeitos em estudo como uma função dos parâmetros do modelo, ou seja, dos $\beta's$. Por isso, algumas vezes escreve-se L como $L(\boldsymbol{\beta})$ (Kleinbaum e Klein, 2005).

No modelo de Cox, o valor dos $\beta's$ é estimado a partir de uma verossimilhança parcial (Carvalho et al., 2011; Kleinbaum e Klein, 2005), denominada dessa forma por considerar apenas as probabilidades dos indivíduos que sofreram o evento de interesse (Kleinbaum e Klein, 2005).

A verossimilhança parcial pode ser escrita como o produto de diversas verossimilhanças, uma para cada m tempo de falha, ou seja:

$$L = L_1 \times \ldots \times L_m = \prod_{j=1}^{m} L_j = \prod_{j=1}^{m} \left(\frac{exp\{\beta_1 x_{1j} + \cdots + \beta_p x_{pj}\}}{\sum_{i \epsilon R_j} exp\{\beta_1 x_{1i} + \cdots + \beta_p x_{pi}\}} \right) \tag{19}$$

sendo que L_j representa a possibilidade de falha nesse momento, dada a sobrevivência até esse instante. Salienta-se que o conjunto de indivíduos em risco no j-ésimo tempo de falha é chamado de conjunto em risco, $R(t_{(j)})$, e esse conjunto mudará conforme o tempo de falha aumentar (Kleinbaum e Klein, 2005; Pintilie, 2006).

Assim, embora a verossimilhança parcial considere apenas os indivíduos que sofreram o evento de interesse, a informação sobre o tempo de sobrevivência antes da censura é usada para os indivíduos que são censurados. Ou seja, um indivíduo que é censurado depois do *j*-ésimo tempo de falha é parte do conjunto em risco usado para determinar o valor de L_j, mesmo que esse indivíduo seja censurado depois.

As estimativas dos coeficientes $\boldsymbol{\beta}$ são obtidas através da maximização do logaritmo natural da função de verossimilhança através de um processo iterativo (Kleinbaum e Klein, 2005).

Para testar as covariáveis e comparar diferentes modelos de Cox utilizam-se as abordagens empregadas na classe de modelos lineares generalizados: **teste de Wald** e **teste da razão de verossimilhança** (Carvalho et al., 2011). Ambos verificam a significância de cada um dos coeficientes β_i, ou seja, matematicamente a hipótese nula a ser testada é representada por:

$$H_0: \beta_i = 0.$$

A forma como isso é feito é o que difere os dois testes.

No teste de Wald, a estatística de teste (z) é obtida através da razão entre a estimativa de máxima verossimilhança do parâmetro β_i ($\hat{\beta}_i$) e seu erro padrão estimado ($\widehat{ep(\hat{\beta}_i)}$), matematicamente definida como (Carvalho et al., 2011):

$$z = \frac{\hat{\beta}_i}{\widehat{ep(\hat{\beta}_i)}}. \qquad (20)$$

Sob a hipótese nula, a estatística de teste segue uma distribuição normal padrão, sendo seu valor calculado comparado com o valor tabelado aproximado de acordo com o nível de significância definido (Carvalho et al., 2011).

Já no teste da razão de verossimilhança, compara-se a diferença entre o logaritmo natural da função de verossimilhança do modelo completo (com tantas covariáveis quanto tiverem sido incluídas no modelo) e o logaritmo natural da função de verossimilhança do modelo sem a covariável em questão da seguinte forma (Carvalho et al., 2011):

$$RV = -ln\left[\frac{Verossimilhança\ do\ modelo\ sem\ a\ variável}{Verossimilhança\ do\ modelo\ com\ a\ variável}\right]. \qquad (21)$$

Sob a hipótese nula, a estatística de teste (*RV*) segue uma distribuição χ^2 com graus de liberdade igual à diferença no número de covariáveis dos modelos em questão (Carvalho et al., 2011). Além disso, esse teste é útil para determinar se o modelo completo é estatisticamente significativo e, desse modo, permite a comparação estatística entre diferentes modelos.

Nos casos em que a amostra é suficientemente grande, o teste da razão de verossimilhança é assintoticamente semelhante ao teste de Wald. Já em situações em que a amostra é pequena, aconselha-se optar pelo teste da razão de verossimilhança, já que esse é mais robusto (Carvalho et al., 2011).

3.1.2.1.2.1. Praticando os conceitos abordados

Utilizaremos os dados apresentados na Tabela 4 para exemplificar a obtenção dos coeficientes do modelo de regressão de Cox, e aplicar os testes de Wald e razão de verossimilhança apresentados na seção anterior.

O objetivo do exemplo será verificar o efeito da covariável sexo no risco de morte dos indivíduos em acompanhamento, bem como utilizar os teste de Wald e razão de verossimilhança para verificar a significância da covariável. Ou seja, desejamos estimar:

$$h(t, X) = h_0(t)e^{\beta X}$$

onde X, nesse caso, representa a covariável sexo e β, o valor do coeficiente associado à covariável.

Para isso, é necessário iniciar com a estimação de β que, conforme mencionado anteriormente, utiliza o conceito de verossimilhança parcial, processo que envolve a maximização do logaritmo natural da função de verossimilhança através de um processo iterativo. Portanto, não conseguiremos detalhar os passos do cálculo, mas apresentaremos os comandos necessários para obter tais valores no *software* R.

A fim de obter o modelo de Cox, é necessário utilizar a função *coxph*, do pacote *survival*, com os argumentos apresentados a seguir:

```
coxph(Surv(tempo, status)~ cov, data = dados, x = TRUE)
```

onde *cov*, refere-se à covariáveis e o argumento opcional *x = TRUE* é incluído na estimação do modelo de Cox quando for necessário salvar a matriz de covariáveis para outras análises. O resultado do ajuste do modelo poderá ser visto utilizando a função *summary()* do objeto em que foi armazenado o comando anterior (Carvalho et al., 2011).

No caso do exemplo apresentado obtivemos:

```
> m1 = coxph(Surv(Sobr, Censura)~ Sexo, data = dados, x = TRUE)
> summary(m1)
Call:
coxph(formula = Surv(Sobr, Censura) ~ Sexo, data = dados, x = TRUE)

  n= 30, number of events= 20

                coef exp(coef) se(coef)     z Pr(>|z|)
SexoMasculino 0.2562    1.2920   0.4511 0.568     0.57

              exp(coef) exp(-coef) lower .95 upper .95
SexoMasculino     1.292      0.774    0.5337     3.128

Concordance= 0.547  (se = 0.061 )
Rsquare= 0.011   (max possible= 0.981 )
Likelihood ratio test= 0.32  on 1 df,   p=0.569
Wald test            = 0.32  on 1 df,   p=0.5701
Score (logrank) test = 0.32  on 1 df,   p=0.569
```

Figura 3 - Saída do R para ajuste do modelo de Cox considerando os dados da Tabela 4

A coluna *coef* apresenta o valor de β estimado via verossimilhança parcial. Valor positivo indica que a covariável contribui para o aumento de risco de óbito, já valor negativo, que a covariável contribui para a redução do risco.

É apresentado também o valor de *exp(coef)*, interpretado como uma razão de riscos, onde: valor superior a 1 indica sobrerrisco e valor entre 0 e 1 indica proteção. Nesse caso, pacientes do sexo masculino, tem 1,292 vezes o risco de óbito de pacientes do sexo feminino. O intervalo de confiança da razão de riscos é apresentado nas colunas *lower .95* e *upper .95* ([0,53; 3,13]). O resultado também pode ser interpretado da seguinte forma, de acordo com a coluna *exp(-coef)*: indivíduos do sexo feminino tem risco 0,77 vezes (ou 33%) menor de falecer por câncer de pulmão do que os pacientes do sexo masculino, sendo o intervalo de confiança igual a [1/3,128; 1/0,5337] = [0,32; 1,87]. Vale lembrar que, uma vez que os intervalos de confiança incluem o valor 1, a estimativa de risco não é significativa.

O valor exibido em *se(coef)* refere-se ao erro padrão do coeficiente estimado; z apresenta o valor calculado para a estatística de Wald; e a última coluna, $Pr\ (>|z|)$, é o valor de p para o teste de Wald, conforme a hipótese nula apresentada anteriormente. Considerando 5% de nível de significância, decide-se não rejeitar a hipótese nula ($p = 0{,}57$), indicando que a variável sexo não contribui, de forma significativa, para a função de risco.

Por fim, no último bloco de informações apresentadas na Figura 3, há o valor do teste da razão de verossimilhança (*Likelihood ratio test*), cujo resultado ($p = 0{,}569$) também demonstra que a variável sexo não contribui, de forma significativa, para a função de risco.

3.1.2.2. Eventos competitivos

Nesta seção, serão discutidas situações onde cada indivíduo pode sofrer apenas um de D tipos de eventos ao longo do período de seguimento, sendo que, quando um dos eventos ocorrer, não há possibilidade de observar a ocorrência dos demais (Balakrishnan e Rao, 2004; Kleinbaum e Klein, 2005). Isso significa que para cada indivíduo em acompanhamento observa-se apenas um tempo e uma causa de falha, dentre diversas possibilidades. Tais situações são muito comuns na área da saúde. Exemplos disso são a análise de dados das diferentes causas de morte de pacientes diagnosticados com câncer (De Glass et al., 2015), os diferentes tipos de eventos em pacientes hemodialisados, como óbito ou transplante renal (Carvalho et al., 2011), ou ainda uma análise de mortalidade infantil nos primeiros dias de vida de uma criança (Ortiz, 1998). O objetivo dessas análises se concentrará em estimar a taxa de ocorrência dos riscos competitivos, comparar suas taxas entre grupos e modelar o efeito das covariáveis na taxa de ocorrência dos riscos competitivos (Balakrishnan e Rao, 2004).

Dessa forma, os tópicos a seguir apresentam a correta abordagem para análise de sobrevivência com riscos competitivos considerando tanto a estimação da taxa de ocorrência dos eventos competitivos quanto à modelagem das covariáveis.

Iniciar-se-á com a descrição da função de incidência acumulada e das curvas de probabilidade condicional que, análogo ao método de Kaplan-Meier, são úteis como uma análise descritiva do tempo de sobrevivência na presença de riscos competitivos permitindo selecionar uma característica (covariável) do paciente e testar a igualdade das funções estimadas entre os estratos dessa característica. Em seguida, será apresentada a modelagem da subdistribuição do risco, que, da mesma forma que o modelo de Cox, permite explorar o

relacionamento de diversas características com a taxa de ocorrência dos riscos competitivos através da estimação do efeito das covariáveis sobre a variável resposta.

3.1.2.2.1. Função de incidência acumulada

A função de distribuição acumulada de um evento utilizando o método de Kaplan-Meier na presença de riscos competitivos é, em geral, superestimada (Carvalho et al., 2011; Giordani, 2013). Isso ocorre porque na abordagem de Kaplan-Meier quando um paciente sofre um evento diferente do de interesse ele é censurado e, portanto, retirado do grupo em risco. Já na abordagem de riscos competitivos, como veremos a seguir, para o cálculo da sobrevivência geral, que leva em conta qualquer evento, se considera que esse indivíduo sofreu algum dos possíveis eventos. Assim, no momento em que esse paciente sofrer algum evento competitivo, a sobrevivência geral estimada é reduzida e, dessa forma, a incidência resultante do evento de interesse também é reduzida (Satagopan et al., 2004).

Como consequência dessa superestimação, tem-se que a relação entre as funções básicas de sobrevivência, apresentada na seção 3.1.1, é perdida na presença de riscos competitivos. Sendo assim, a exponencial da função de risco acumulado torna-se uma quantidade sem significado, não estando mais relacionada a qualquer função de sobrevivência (Balakrishnan e Rao, 2004; Kleinbaum e Klein, 2005), o que torna a interpretação da abordagem de Kaplan-Meier questionável se utilizada em tal situação.

Outra razão pela qual a utilização do método de Kaplan-Meier é inadequada na presença de riscos competitivos é que, além de considerar apenas um evento de interesse, essa abordagem requer a suposição de independência dos riscos competitivos, o que não pode ser verificada (Kleinbaum e Klein, 2005). Tais fatos motivaram o desenvolvimento de uma abordagem alternativa denominada **função de incidência acumulada**, também chamada de **subdistribuição**. Derivada da função de risco da causa específica, a função de incidência acumulada fornece uma estimativa da probabilidade marginal de um evento na presença de eventos competitivos e não requer a suposição de que os riscos competitivos sejam independentes (Kleinbaum e Klein, 2005).

Matematicamente, a função de incidência acumulada ($F_q(t)$) é definida como a probabilidade acumulada do evento competitivo q ocorrer na presença de outros eventos competitivos. Ou seja (Dignam e Kocherginsky, 2008):

$$F_q(t) = P(\, T \leq t, evento = q) = \int_0^t S(u) h_q(u) du, \qquad (22)$$

onde $h_q(u)$ representa o risco específico do evento q e $S(u)$ representa a probabilidade de ter sobrevivido a qualquer um dos possíveis eventos até o tempo t. Heuristicamente, para obter essa expressão necessita-se da estimação de duas quantidades: o risco de um evento específico e a probabilidade geral de sobrevivência no tempo t_{j-1} (onde t_{j-1} representa um instante de tempo anterior a t_j).

Para o cálculo da primeira quantidade, risco de um evento específico, considera-se que $0 < t_1 < t_2 < \cdots < t_m$ represente os tempos distintos ordenados de ocorrência de qualquer um dos q possíveis eventos $(q = 1, \ldots, Q)$ e que $d_q(t_j)$ seja o número de indivíduos que sofreram o evento q no tempo t_j. Assim, o **risco de um evento específico** ($\boldsymbol{\hat{h}_q(t_j)}$) é obtido simplesmente pela razão entre o número de eventos do tipo q que ocorreram em t_j e o número de indivíduos em risco no tempo t_j ($R(t_j)$), o que é matematicamente definido por (Carvalho et al., 2011; Kleinbaum e Klein, 2005):

$$\hat{h}_q(t_j) = \frac{d_q(t_j)}{R(t_j)}. \qquad (23)$$

A segunda quantidade necessária é a probabilidade geral de sobrevivência no tempo t_{j-1}, que se refere à probabilidade de que o paciente não tenha sofrido qualquer um dos q possíveis eventos antes do tempo t_j. Para chegar a esse valor define-se, primeiramente, o **número total de ocorrências considerando todos os eventos**, dado por $\boldsymbol{d(t_j)}$, obtido da seguinte forma (Kleinbaum e Klein, 2005):

$$d(t_j) = \sum_{q=1}^{Q} d_q(t_j). \qquad (24)$$

Com isso, é possível definir uma probabilidade geral de sobrevivência, análogo ao que foi visto no estimador de Kaplan-Meier, como (Carvalho et al., 2011):

$$\hat{S}(\mathrm{t}) = \prod_{j:t_j \leq t} \left(\frac{R(t_j) - d(t_j)}{R(t_j)} \right). \quad (25)$$

Tendo as duas quantidades, define-se que a probabilidade de falha, ou seja, a incidência para um evento do tipo q ($\hat{F}_q(t_j)$) é dada simplesmente pela probabilidade de sobreviver até o instante anterior à ocorrência do evento multiplicada pelo risco da causa específica para o evento q no tempo t_j, ou seja (Kleinbaum e Klein, 2005):

$$\hat{F}_q(t_j) = \hat{S}(t_{j-1}) \times \hat{h}_q(t_j). \quad (26)$$

Por fim, define-se a função de incidência acumulada ($\hat{F}(t)$) no tempo t_j como a soma acumulada até o tempo t_j ($j = 1, \ldots, m$) dos valores de incidência sobre todos os q tipos de eventos (Kleinbaum e Klein, 2005):

$$\hat{F}(t) = \sum_{\forall j, t_j \leq t} \hat{S}(t_{j-1}) \hat{h}_q(t_j), \quad (27)$$

sendo que, por definição, em $t = 0$, $S(t_0) = 1$.

O resultado da função de incidência acumulada fornece uma probabilidade marginal que não utiliza a formulação do produto limite e nem necessita da suposição de que os riscos competitivos sejam independentes. Porém, requer que o risco global seja igual à soma de todos os possíveis riscos, pressuposto que é satisfeito apenas em situações onde os riscos competitivos sejam mutuamente excludentes, e os eventos não recorrentes, ou seja, um, e somente um, evento pode ocorrer e apenas uma vez ao longo do tempo para cada indivíduo (Kleinbaum e Klein, 2005).

É possível, também, fazer comparação da função de incidência acumulada para dois ou mais grupos utilizando o **teste de Gray.** A hipótese nula, considerando o número de grupos (k) maior ou igual a dois, é (Balakrishnan e Rao, 2004; Gray, 1988):

$$H_0: F_{11}(t) = \cdots = F_{1k}(t)$$

onde $F_{1k}(t)$ é a função de incidência acumulada de um evento do tipo $q = 1$ no grupo k.

A estatística de teste é baseada na comparação de médias ponderadas dos riscos da subdistribuição entre os grupos para o evento de interesse. Assim, considerando k grupos, a estatística de teste é definida por um escore da seguinte forma (Gray, 1988):

$$z_k(\tau) = \int_{t=0}^{\tau} W_k(t)\{h_k(t) - h_0(t)\}dt, \tag{28}$$

onde τ representa o maior tempo observado em ambos os grupos; $W_k(t)$ é uma função peso; $h_k(t)$ representa o risco da subdistribuição para o grupo k; e $h_0(t)$ é o risco da subdistribuição considerando todos os grupos.

A quantidade $h_k(t)$ é matematicamente definida por (Gray, 1988):

$$h_k(t) = \frac{f_k(t)}{1 - F_k(t)}, \tag{29}$$

onde $f_k(t)$ é a subdensidade para o evento de interesse e $F_k(t)$ a função de incidência acumulada no grupo k para o evento de interesse. O risco da subdistribuição para o grupo k pode ser interpretado como a probabilidade de observar o evento de interesse no próximo intervalo de tempo, dado que o evento de interesse não tenha ocorrido até então ou que o evento competitivo tenha sido observado (Pintilie, 2006).

Já a função peso tem, em geral, a forma $W_k(t) = L(t)R_k(t)$ para alguma função $L(t)$ e para

$$R_k(t) = n_k(t)\frac{1 - \hat{F}_k(t-)}{\hat{S}_k(t-)}, \tag{30}$$

onde $L(t)$ é um processo previsível que converge uniformemente em probabilidade para uma função limitada $L^0(t)$; $n_k(t)$ é o número de pacientes em risco no tempo t e grupo k; $F(t-)$ é o limite do lado esquerdo da função de incidência acumulada do evento de interesse e $S(t-)$ é o limite do lado esquerdo da probabilidade de estar livre de qualquer evento, como estimado pelo método de Kaplan-Meier. Dessa forma, R_k representa o número ajustado de pacientes em risco (Pintilie, 2006).

Assim, sob a hipótese nula, a estatística de teste tem distribuição aproximadamente χ^2 com $(k-1)$ graus de liberdade (Pintilie, 2006).

De forma geral, supondo o evento de interesse 1, esse teste é baseado na comparação de diferenças ponderadas entre as estimativas da função de risco $h_{1k}(t)$ e a estimativa de $h_1(t)$ a partir da amostra agrupada (Balakrishnan e Rao, 2004). Desse modo, rejeitar a hipótese nula significa concluir que pelo menos uma das funções de incidência acumulada é estatisticamente diferente em algum tempo. Como uma opção para responder quais grupos têm funções de incidência iguais e quais são diferentes pode-se repetir o teste usando somente dados de dois grupos por vez (Balakrishnan e Rao, 2004).

3.1.2.2.1.1. Praticando os conceitos abordados

Neste tópico, apresentaremos um passo a passo da estimação da função de incidência acumulada e exemplificaremos a comparação de curvas através do teste de Gray.

Consideremos que os dados apresentados na Tabela 7 referem-se ao tempo de sobrevivência, em anos, de 50 pacientes com câncer de estômago diagnosticado entre os anos de 2002 e 2009 (dados da população observada no presente trabalho). Esses pacientes foram acompanhados até o ano de 2013, quando se observou o sexo (Mas = masculino e Fem = feminino) e tipo de evento experimentado por cada indivíduo, sendo: censura, caso o paciente não tenha falecido até o término do período de acompanhamento; óbito por câncer, se a causa básica de morte tiver sido o câncer; ou óbito por outras causas, se a causa básica de morte não estiver relacionada ao câncer.

Uma vez que o indivíduo experimenta um dos eventos, os demais não podem ser observados para esse mesmo paciente. Sendo assim, trata-se claramente de um problema de eventos competitivos. Portanto, a partir desses dados, será calculada a função de incidência acumulada considerando como evento de interesse o óbito por câncer e considerando o óbito por outras causas como evento competitivo.

Iniciamos determinando o número de pessoas que sofreram cada um dos eventos em cada tempo t_j, sendo $\hat{d}_{Câncer}(t_j)$ o número de pacientes que tiveram óbito por câncer e $\hat{d}_{Outras}(t_j)$ o número de pacientes que faleceram por outras causas. Depois, usando a equação 24, calcularemos o número total de ocorrências considerando todos os eventos ($\hat{d}(t_j)$). Em seguida, aplicaremos a equação 23 para obter a função de risco para o evento de interesse conforme mostra a Tabela 8.

Tabela 7 - Dados de 50 pacientes diagnosticados com câncer de estômago

Paciente	Sexo	Tempo de sobrevivência (anos)	Evento	Paciente	Sexo	Tempo de sobrevivência (anos)	Evento
1	Fem	0,02	Óbito por câncer	26	Mas	2,07	Óbito por outras causas
2	Mas	0,05	Óbito por câncer	27	Fem	4,29	Óbito por câncer
3	Fem	0,06	Óbito por outras causas	28	Fem	4,76	Censura
4	Fem	0,09	Óbito por câncer	29	Mas	5,03	Censura
5	Mas	0,10	Óbito por câncer	30	Fem	5,10	Censura
6	Fem	0,12	Óbito por outras causas	31	Fem	5,13	Óbito por câncer
7	Mas	0,19	Óbito por câncer	32	Mas	5,76	Censura
8	Fem	0,22	Óbito por câncer	33	Fem	6,12	Censura
9	Fem	0,23	Óbito por câncer	34	Mas	6,88	Óbito por câncer
10	Fem	0,23	Óbito por outras causas	35	Mas	7,00	Censura
11	Fem	0,25	Óbito por câncer	36	Mas	7,29	Censura
12	Mas	0,26	Óbito por câncer	37	Mas	7,41	Óbito por câncer
13	Mas	0,34	Óbito por câncer	38	Mas	7,81	Óbito por câncer
14	Fem	0,37	Óbito por câncer	39	Fem	7,86	Censura
15	Fem	0,44	Óbito por câncer	40	Fem	8,30	Óbito por outras causas
16	Mas	0,50	Óbito por câncer	41	Fem	8,80	Óbito por outras causas
17	Mas	0,51	Óbito por câncer	42	Mas	9,64	Óbito por outras causas
18	Fem	0,54	Óbito por câncer	43	Fem	9,87	Censura
19	Mas	0,57	Óbito por outras causas	44	Fem	10,05	Censura
20	Fem	0,62	Óbito por câncer	45	Fem	10,53	Censura
21	Mas	0,67	Óbito por câncer	46	Fem	10,70	Óbito por outras causas
22	Mas	0,89	Óbito por outras causas	47	Mas	11,10	Censura
23	Fem	1,01	Óbito por câncer	48	Fem	11,47	Censura
24	Mas	1,04	Óbito por câncer	49	Fem	11,55	Censura
25	Mas	1,56	Óbito por câncer	50	Mas	11,81	Censura

Tabela 8 – Exemplo da estimação do risco para óbito por câncer considerando os 50 pacientes diagnosticados com câncer de estômago

t_j	$\widehat{d}_{Câncer}(t_j)$	$\widehat{d}_{Outras}(t_j)$	$\widehat{d}(t_j)$	$R(t_j)$	$\widehat{h}_{Câncer}(t_j)$
0,00	0	0	$0+0+0=0$	50	$\frac{0}{50}=0{,}0000$
0,02	1	0	$1+0+0=1$	50	$\frac{1}{50}=0{,}0200$
0,05	1	0	1	49	0,0204
0,06	0	1	1	48	0,0000
0,09	1	0	1	47	0,0213
0,10	1	0	1	46	0,0217
0,12	0	1	1	45	0,0000
0,19	1	0	1	44	0,0227
0,22	1	0	1	43	0,0223
0,23	1	1	2	42	0,0238
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
7,41	1	0	1	14	0,0714
7,81	1	0	1	13	0,0769
7,86	0	0	1	12	0,0000
8,30	0	1	1	11	0,0000
8,80	0	1	1	10	0,0000
9,64	0	1	1	9	0,0000
9,87	0	0	0	8	0,0000
10,05	0	0	0	7	0,0000
10,53	0	0	0	6	0,0000
10,70	0	1	1	5	0,0000
11,10	0	0	0	4	0,0000
11,47	0	0	0	3	0,0000
11,55	0	0	0	2	0,0000
11,81	0	0	0	1	0,0000

Com essas quantidades estimadas, consegue-se obter tanto a função de sobrevivência estimada para cada tempo t_j (equação 25) quanto às funções de incidência e incidência acumulada (equações 26 e 27), conforme apresentado na Tabela 9.

Tabela 9 – Exemplo de cálculo das funções de sobrevivência, incidência e incidência acumulada considerando o evento óbito por câncer

t_j	$\widehat{d}(t_j)$	$R(t_j)$	$\widehat{h}_{Câncer}(t_j)$	$\widehat{S}(t_j)$	$\widehat{F}_{Câncer}(t_j)$	$\widehat{F}_{Câncer}(t_j)$ **acumulada**
0,00	0	50	0,00	$\frac{50-0}{50} = 1{,}00$	0,00	0,00
0,02	1	50	0,02	$1 \times \frac{50-1}{50} = 0{,}98$	$1{,}0 \times 0{,}02 = 0{,}02$	$0 + 0{,}02 = 0{,}02$
0,05	1	49	0,02	$0{,}98 \times \frac{49-1}{49} = 0{,}96$	$0{,}98 \times 0{,}02 = 0{,}02$	$0{,}02 + 0{,}02 = 0{,}04$
0,06	1	48	0,00	0,94	$0{,}96 \times 0{,}00 = 0{,}00$	$0{,}04 + 0 = 0{,}040$
0,09	1	47	0,02	0,92	0,02	0,06
0,10	1	46	0,02	0,90	0,02	0,08
0,12	1	45	0,00	0,88	0,00	0,08
0,19	1	44	0,02	0,86	0,02	0,10
0,22	1	43	0,02	0,84	0,02	0,12
0,23	2	42	0,02	0,80	0,02	0,14
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
7,41	1	14	0,07	0,38	0,03	0,50
7,81	1	13	0,08	0,35	0,03	0,53
7,86	0	12	0,00	0,35	0,00	0,53
8,30	1	11	0,00	0,32	0,00	0,53
8,80	1	10	0,00	0,29	0,00	0,53
9,64	1	9	0,00	0,26	0,00	0,53
9,87	0	8	0,00	0,26	0,00	0, 53
10,05	0	7	0,00	0,26	0,00	0, 53
10,53	0	6	0,00	0,26	0,00	0,53
10,70	1	5	0,00	0,21	0,00	0, 53
11,10	0	4	0,00	0,21	0,00	0,53
11,47	0	3	0,00	0,21	0,00	0, 53
11,55	0	2	0,00	0,21	0,00	0, 53
11,81	0	1	0,00	0,21	0,00	0, 53

A partir da Tabela 9, podemos observar que o maior valor da função de incidência acumulada é 0,53, que ocorre em $t = 7{,}81$ anos, tempo em que ocorreu o último evento de interesse. Assim, a incidência acumulada, ou seja, a probabilidade marginal para morte por câncer em até 7,81 anos é de 0,53, considerando a presença do evento competitivo óbito por outras causas.

Da mesma forma que fizemos na seção em que abordamos a função de distribuição acumulada, podemos utilizar um *software* estatístico para visualizar graficamente a função de

incidência acumulada. Utilizando o R, precisamos do pacote *survival* e dos seguintes comandos:

```
incidencia_acumulada = survfit(Surv(tempo, event = status > 0)~ 1,
                               etype = motivo, data = dados)
plot(incidencia_acumulada)
```

onde *tempo* representa a coluna com informações de tempo de sobrevivência; *status* identifica a coluna que contém o tipo de evento (variável numérica); *etype* representa a coluna que identifica o tipo de evento (variável categórica), sendo esse o argumento que permite estimar a incidência acumulada por causa específica; *data* é o nome do banco de dados que possui as colunas *tempo*, *status* e *motivo*. A função *plot* permite obter o gráfico.

Como resultado, obtivemos:

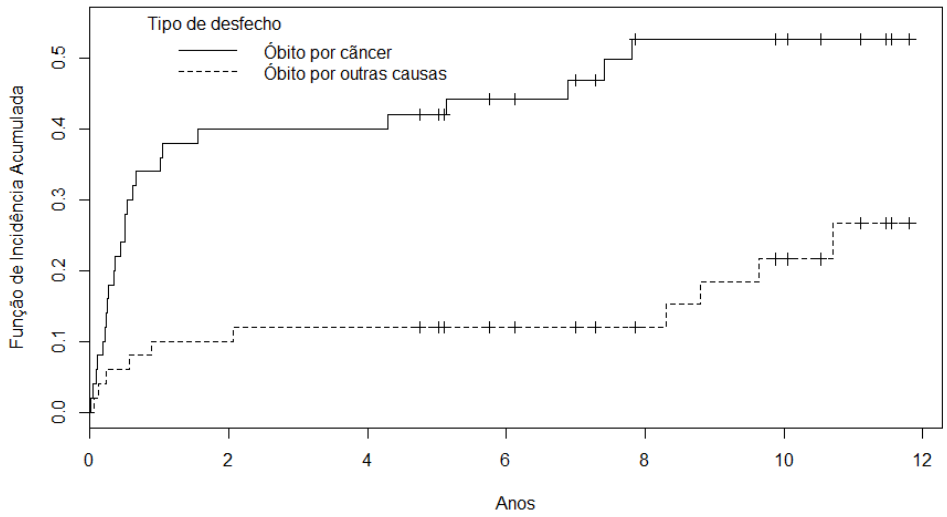


Figura 4 - Função de incidência acumulada para os 50 pacientes diagnosticados com câncer de estômago segundo os dois eventos estudados

Note que, da mesma forma que na abordagem de Kaplan-Meier, a função de incidência acumulada não considera covariáveis. Entretanto, é possível comparar as curvas para categorias de uma variável testando se há diferença entre as funções de incidência acumulada através do teste de Gray. Assim, seguiremos o exemplo utilizando esse teste para comparar as curvas de incidência acumulada de homens e mulheres.

No contexto de riscos competitivos, testar a igualdade de duas funções de incidência acumulada equivale a testar se a essa função é semelhante em cada estrato, considerando os

possíveis eventos. Para demonstração, consideraremos apenas o evento óbito por câncer e assim, a hipótese nula a ser testada é escrita como:

$$H_0: F_{Câncer\_Homens}(t) = F_{Câncer\_Mulheres}(t),$$

onde $F_{Câncer\_Homens}(t)$ é a função de incidência acumulada para o evento óbito por câncer no grupo de homens e $F_{Câncer\_Mulheres}(t)$ representa a função de incidência acumulada para o evento óbito por câncer no grupo de mulheres.

Como já mencionado, utilizaremos o teste de Gray para testar H_0, sendo que a estatística de teste é baseada na comparação de médias ponderadas dos riscos da subdistribuição entre os grupos para o evento de interesse.

Apenas para exemplificar, iremos **supor** que o tempo é discreto assumindo que $L(t) = 1$ e, desse modo, $W_{Mas}(t) = R_{Mas}(t)$. O número de óbitos por câncer no grupo do sexo masculino em t_j é dado por d_{Mas_j}; d_{Fem_j} representa o número de óbitos por câncer no grupo do sexo feminino em t_j; n_{Mas_j} refere-se ao número de pacientes do sexo masculino em risco no tempo t_j e n_{Fem_j} representa o número de pacientes do sexo feminino em risco em t_j. Dessa forma, podemos escrever o número ajustado de indivíduos em risco como:

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

O escore Z será dado por:

$$Z = \sum_{\forall\, t_j} W(t_j)\left(\frac{d_{Mas\_j}}{R_{Mas\_j}} - \frac{d_{Mas\_j} + d_{Fem\_j}}{R_{Mas\_j} + R_{Fem\_j}}\right)$$

$$Z = \sum_{\forall\, t_j} R_{Mas}\left(\frac{d_{Mas\_j}}{R_{Mas\_j}} - \frac{d_{Mas\_j} + d_{Fem\_j}}{R_{Mas\_j} + R_{Fem\_j}}\right)$$

Lembrando que a estatística de teste, z_k, tem forma quadrática e as etapas de seu cálculo necessitam tanto do cálculo de Z (numerador da estatística de teste) quanto da matriz de variância-covariância (denominador da estatística de teste). A fórmula da variância é um

tanto quanto complicada e, por esse motivo, não é apresentada nesse exemplo. Porém, uma descrição completa desse assunto pode ser encontrada em Gray (1988) e em Pintilie (2006).

A seguir, demonstraremos os passos para obtenção do numerador da estatística de teste de forma manual, **válido para o caso de tempo discreto** (mesmo não sendo o caso deste exemplo, apresentaremos como seria o cálculo): iniciamos pela determinação do número de pacientes em risco para qualquer evento (n_{Mas_j} e n_{Fem_j}) e do número de eventos sofridos (d_{Mas_j} e d_{Fem_j}) considerando os diferentes sexos; em seguida, determinamos o valor estimado para a sobrevivência ($\hat{S}_{Mas}(t_j)$ e $\hat{S}_{Fem}(t_j)$) através da equação 25. Para todos os cálculos, apenas os detalhes do sexo masculino serão apresentados, já que para o sexo feminino o raciocínio é análogo.

Tabela 10 – Primeira etapa para o cálculo do numerador da estatística do teste de Gray

t_j	Evento	Sexo	n_{Mas_j}	n_{Fem_j}	d_{Mas_j}	d_{Fem_j}	$\widehat{S}_{Mas}(t_j)$	$\widehat{S}_{Fem}(t_j)$
0,02	Óbito por câncer	Fem	23	27	0	1	$\frac{23-0}{23} = 1{,}00$	0,96
0,05	Óbito por câncer	Mas	23	26	1	0	$1 * \left(\frac{23-1}{23}\right) = 0{,}96$	0,96
0,06	Óbito por outras causas	Fem	22	26	0	1	0,96	0,93
0,09	Óbito por câncer	Fem	22	25	0	1	0,96	0,89
0,10	Óbito por câncer	Mas	22	24	1	0	0,91	0,89
0,12	Óbito por outras causas	Fem	21	24	0	1	0,91	0,85
0,19	Óbito por câncer	Mas	21	23	1	0	0,87	0,85
0,22	Óbito por câncer	Fem	20	23	0	1	0,87	0,81
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
9,87	Censura	Fem	2	6	0	0	0,15	0,33
10,05	Censura	Fem	2	5	0	0	0,15	0,33
10,53	Censura	Fem	2	4	0	0	0,15	0,33
10,70	Óbito por outras causas	Fem	2	3	0	1	0,15	0,22
11,10	Censura	Mas	2	2	0	0	0,15	0,22
11,47	Censura	Fem	1	2	0	0	0,15	0,22
11,55	Censura	Fem	1	1	0	0	0,15	0,22
11,81	Censura	Mas	1	0	0	0	0,15	0,22

Em seguida, calculamos o valor da função de incidência acumulada para cada um dos sexos ($\hat{F}_{Ca\_Mas}$ e $\hat{F}_{Ca\_Fem}$) utilizando a equação 27. Para isso, é necessário definir o número de eventos de interesse (óbito por câncer) em cada grupo comparado e em cada tempo (d_{Ca_Mas} e d_{Ca_Fem}).

Tabela 11 – Segunda etapa para o cálculo do numerador da estatística do teste de Gray

t_j	$d_{Mas\_j}(t_j)$	$d_{Fem\_j}(t_j)$	$\widehat{S}_{Mas}(t_j)$	$\widehat{S}_{Fem}(t_j)$	$d_{Ca\_Mas}(t_j)$	$d_{Ca\_Fem}(t_j)$	$F_{Ca\_Mas}(t_j)$	$F_{Ca\_Fem}(t_j)$
0,02	23	27	1,00	0,96	0	1	$1 * \left(\frac{0}{23}\right) = 0{,}00$	0,04
0,05	23	26	0,96	0,96	1	0	$0 + \left(1 * \left(\frac{1}{23}\right)\right) = 0{,}04$	0,04
0,06	22	26	0,96	0,93	0	0	$0{,}04 + \left(0{,}96 * \left(\frac{0}{22}\right)\right) = 0{,}04$	0,04
0,09	22	25	0,96	0,89	0	1	0,04	0,07
0,10	22	24	0,91	0,89	1	0	0,09	0,07
0,12	21	24	0,91	0,85	0	0	0,09	0,07
0,19	21	23	0,87	0,85	1	0	0,13	0,07
0,22	20	23	0,87	0,81	0	1	0,13	0,11
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
9,87	2	6	0,15	0,33	0	0	0,64	0,45
10,05	2	5	0,15	0,33	0	0	0,64	0,45
10,53	2	4	0,15	0,33	0	0	0,64	0,45
10,70	2	3	0,15	0,22	0	0	0,64	0,45
11,10	2	2	0,15	0,22	0	0	0,64	0,45
11,47	1	2	0,15	0,22	0	0	0,64	0,45
11,55	1	1	0,15	0,22	0	0	0,64	0,45
11,81	1	0	0,15	0,22	0	0	0,64	0,45

Com essas quantidades em mãos, determinamos o número modificado de pacientes em risco (R_{Mas} e R_{Fem}), conforme mostra a equação 30. E, por fim, utilizamos a equação 28 para cálculo do escore Z. Lembrando que como estamos comparando apenas dois grupos, basta calcular um valor de Z.

Tabela 12 – Última etapa para o cálculo do numerador da estatística do teste de Gray

t_j	n_{Mas_j}	$\widehat{S}_{Mas}(t_j)$	$F_{Ca\_Mas}(t_j)$	$R_{Mas}(t_j)$	$R_{Fem}(t_j)$	$Z_{Mas}(t_j)$
0,02	23	1,00	0,00	$23 * \left(\frac{1-0}{1}\right) = 23{,}00$	27,00	$23 * \left(\frac{0}{23} - \frac{0+1}{23+27}\right) = -0{,}46$
0,05	23	0,96	0,04	$23 * \left(\frac{1-0}{1}\right) = 23{,}00$	26,00	$23 * \left(\frac{1}{23} - \frac{1+0}{23+26}\right) = 0{,}53$
0,06	22	0,96	0,04	$22 * \left(\frac{1-0{,}04}{0{,}96}\right) = 22{,}00$	26,00	$22 * \left(\frac{0}{22} - \frac{0+0}{22+26}\right) = 0$
0,09	22	0,96	0,04	22,00	26,00	-0,46
0,10	22	0,91	0,09	22,00	25,00	0,53
0,12	21	0,91	0,09	21,00	25,00	0,00
0,19	21	0,87	0,13	21,00	25,00	0,54
0,22	20	0,87	0,13	20,00	25,00	-0,44
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
9,87	2	0,15	0,64	4,71	10,03	0,00
10,05	2	0,15	0,64	4,71	8,36	0,00
10,53	2	0,15	0,64	4,71	6,69	0,00
10,70	2	0,15	0,64	4,71	5,02	0,00
11,10	2	0,15	0,64	4,71	5,02	0,00
11,47	1	0,15	0,64	2,36	5,02	0,00
11,55	1	0,15	0,64	2,36	2,51	0,00
11,81	1	0,15	0,64	2,36	0,00	0,00
						$\boldsymbol{Z_{Mas} = -0{,}46 + \cdots + 0{,}00 = 1{,}71}$

Lembre-se: **esse cálculo seria válido no caso de tempos discretos**, o que não ocorre aqui. O objetivo foi apenas demonstrar os passos para obtenção de Z.

Agora, descreveremos os comandos necessários para realizar esse teste com o auxílio do *software* R, sendo necessária previamente à instalação do pacote *cmprsk* (Bob Gray (2013). *cmprsk: Subdistribution Analysis of Competing Risks*. R package version 2.2-6.).

```
testeGray = cuminc(ftime = tempo, fstatus = evento,
                   group = grupo_comparado, cencode = "Censura")
testeGray$Tests
```

onde *ftime* refere-se à coluna com informações de tempo de sobrevivência; *fstatus*, a coluna que identifica o tipo de evento; *group* indica a coluna com a variável de grupos a serem

comparados; *cencode* é o código que identifica as observações censuradas; e *testeGray$Tests* mostra o resultado do teste de Gray. Fazendo isso, obteremos:

```
> testeGray$Tests
                                    stat        pv df
Óbito por câncer          0.4694731 0.4932297  1
Óbito por outras causas 0.2049597 0.6507470  1
```

Figura 5 - Resultado do teste de Gray para os dados da Tabela 7

A Figura 5 apresenta: na primeira coluna, os dois eventos observados; na segunda coluna, os valores calculados da estatística de teste; na terceira coluna os valores de *p*; e, na quarta coluna, o número de graus de liberdade. Pelos valores de *p* apresentados (em ambos, $p > 0{,}05$), concluímos que não há diferença estatisticamente significativa entre as funções de incidência acumulada de homens e mulheres, considerando tanto o evento óbito por câncer quanto o óbito por causa não relacionada ao câncer.

Por fim, é possível visualizar o gráfico das funções de incidência acumulada para ambos os eventos considerando a estratificação por sexo através do seguinte comando:

```
plot(testeGray, xlab = "Anos", ylab = "Função de Incidência
      Acumulada")
```

cujo resultado é apresentado a seguir.

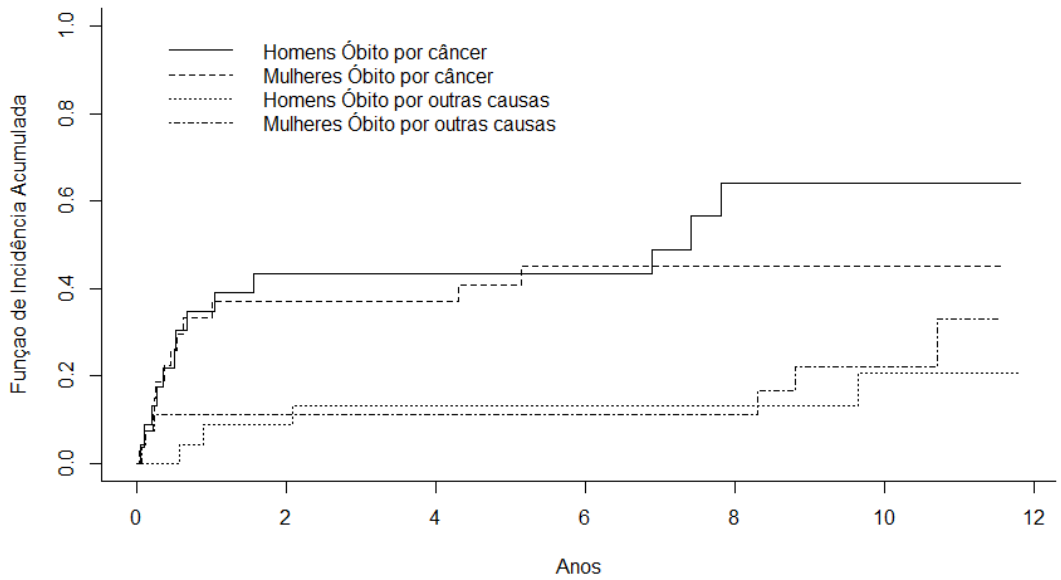


Figura 6 - Comparação da função de incidência acumulada para os eventos observados considerando a estratificação por sexo

3.1.2.2.2. Curvas de probabilidade condicional

Outra função utilizada no contexto de riscos competitivos é chamada de **curva de probabilidade condicional**. Por definição, essa curva fornece a probabilidade de um indivíduo experimentar o evento q no tempo t, dado que esse indivíduo não experimentou nenhum dos outros eventos competitivos até o tempo t, o que é matematicamente representado por:

$$CPC_q(t) = P\left(T_q \leq t \middle| T \geq t\right), \tag{31}$$

onde CPC_q diz respeito à curva de probabilidade condicional do evento q; T_q representa o tempo até a ocorrência do evento q; e T, o tempo até qualquer evento competitivo ocorrer (Kleinbaum e Klein, 2005). Da mesma forma que a função de incidência acumulada, a probabilidade condicional é uma função monótona crescente, com limites entre 0 e 1 (Pintilie, 2006).

Considerando o risco competitivo tipo q, há uma relação entre a função de incidência acumulada e a curva de probabilidade condicional dada por:

$$CPC_q(t) = \frac{F_q(t)}{1 - F_{q'}(t)} \tag{32}$$

sendo que $F_{q'}(t)$ representa a função de incidência acumulada para outros eventos que não q, ou seja, todos os outros possíveis eventos (Kleinbaum e Klein, 2005).

É possível gerar curvas de probabilidade condicional acumulada a partir dos gráficos da função de incidência acumulada, metodologia proposta por Pepe-Mori (1993) e Lunn (1998). O primeiro também propôs um teste que permite comparar duas curvas de probabilidade condicional acumulada diretamente, o que foi posteriormente estendido para k grupos pelo segundo autor (Kleinbaum e Klein, 2005).

A seguir, será apresentado o teste de Pepe-Mori, que considera apenas dois grupos e foi baseado em um trabalho anterior de Pepe (Pepe, 1991) onde o autor propõe um teste para comparar duas curvas de probabilidade condicional, sendo a hipótese nula (Pepe e Mori, 1993):

$$H_0: CPC_1(t) = CPC_2(t).$$

A estatística de teste é baseada na soma ponderada das diferenças entre as curvas considerando os dois grupos, sendo escrita como (Pintilie, 2006):

$$s = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \int_0^{\tau} W(t)\{CPC_1(t) - CPC_2(t)\}\, dt, \qquad (33)$$

onde N_1 e N_2 representam o número total de indivíduos nos grupos 1 e 2, respectivamente; τ é o tempo máximo observado, considerando os dois grupos; $W(t)$ é a função peso; e $CPC_1(t)$ e $CPC_2(t)$ representam as curvas de probabilidade condicional dos grupos 1 e 2.

Sob a hipótese nula, a estatística de teste tem uma distribuição aproximadamente χ^2 com $(k-1)$ graus de liberdade (Pintilie, 2006).

Supondo o caso discreto, onde $t_1, t_2, \ldots, t_n$ referem-se aos tempos únicos ordenados para todas as observações (eventos ou não) em ambos os grupos, a fórmula 33 pode ser reescrita como (Pintilie, 2006):

$$s = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \sum_{\forall\, t_j} \left\{ W(t_j) \left[\widehat{CPC}_1(t_j) - \widehat{CPC}_2(t_j) \right] (t_{j+1} - t_j) \right\}. \qquad (34)$$

E a função peso é dada por (Pintilie, 2006):

$$W(t_j) = \frac{(N_1 + N_2)\hat{C}_1(t_{j-1})\hat{C}_2(t_{j-1})}{N_1 \hat{C}_1(t_{j-1}) + N_2 \hat{C}_2(t_{j-1})}, \qquad (35)$$

sendo que $1 - \hat{C}(t)$ é o estimador de Kaplan-Meir para a distribuição de censura. Note que $\hat{C}(t)$ é o estimador de Kaplan-Meier para a função de sobrevivência onde os eventos são definidos como observações que são censuras ou eventos com riscos competitivos. Isso significa que a fórmula 34 muda toda vez que um evento de interesse ocorre, devido à mudança em $F(t)$, e também quando, no período de tempo anterior a observação, ocorreu uma censura ou um evento competitivo, devido à mudança em $C(t)$. Essa função peso

decresce com o tempo, portanto seu efeito em s é dar menos peso para a diferença das funções de incidência acumulada conforme o tempo aumenta (Pintilie, 2006).

A variância de s é estimada como uma média ponderada das variâncias dos dois grupos. Ela é, matematicamente, definida por (Pintilie, 2006):

$$\hat{\sigma}^2 = \frac{N_1 N_2 (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}{N_1 + N_2}. \tag{36}$$

Uma vez que o cálculo da variância é o mesmo em cada um dos dois grupos, apresenta-se o estimador da variância somente para um deles e considera-se que: n_j representa o número de pessoas em risco no tempo t_j no grupo k; d_j, o número de eventos (de interesse ou competitivo) no tempo t_j no grupo k; d_{Intj}, o número de eventos de interesse no tempo t_j no grupo k; $\hat{F}_q$, a função de incidência acumulada para o evento de interesse no grupo k; $\hat{F}_{q'}$, a função de incidência acumulada para o evento competitivo no grupo k; e σ, o desvio padrão para o grupo k. Com isso, o estimador da variância pode ser calculado como (Pintilie, 2006):

$$\hat{\sigma}_k^2 = \sum_{\forall\, t_j} [v(t_j)]^2 \frac{[1 - \hat{F}_{q'}(t_j)]^2 d_{Intj} + [\hat{F}_q(t_j)]^2 (d_j - d_{Intj})}{n_j\,(n_j - 1)} \tag{37}$$

sendo que:

$$v(t_j) = \sum_{t_k \geq t_j} \frac{W(t_k)\hat{S}(t_k)}{\left(1 - \hat{F}_q(t_k)\right)^2}. \tag{38}$$

Lunn (1998) estendeu o teste de Pepe-Mori considerando k grupos através da seguinte estatística de teste (Pintilie, 2006):

$$z_k = \sqrt{n_k} \int W(t)\{\mathrm{F_k(t)} - \mathrm{F_0(t)}\}\, dt, \tag{39}$$

onde W é a função peso, F_k é a função de incidência acumulada no grupo k e F_0 é a função de incidência global, considerando todos os grupos. Sob a hipótese nula, $\boldsymbol{Z}^t = (z_1, z_2, \ldots, z_k)$ segue uma distribuição χ^2 com $(k-1)$ graus de liberdade.

3.1.2.2.2.1. Praticando os conceitos abordados

Neste tópico, apresentaremos um passo a passo da estimação da curva de probabilidade condicional e exemplificaremos a comparação de curvas através do teste de Pepe-Mori. Para isso, continuaremos utilizando os dados apresentados na Tabela 7.

Conforme visto na equação 31, há uma relação entre a função de incidência acumulada e a curva de probabilidade condicional. O primeiro passo para estimar a curva de probabilidade condicional para cada evento será calcular a função de incidência acumulada para o evento óbito por outras causas, cujo raciocínio é análogo ao apresentado na seção anterior. Em seguida, aplicaremos a equação 32 a fim de obter os valores da probabilidade condicional para cada evento, em cada tempo. Os resultados são apresentados a seguir.

Tabela 13 – Cálculo das curvas de probabilidade condicional para os eventos óbito por câncer ($\boldsymbol{CPC_{Câncer}(t_j)}$) e óbito por outras causas ($\boldsymbol{CPC_{Outras}(t_j)}$)

$\boldsymbol{t_j}$	$\boldsymbol{\widehat{F}_{Câncer}(t_j)}$	$\boldsymbol{\widehat{F}_{Outras}(t_j)}$	$\boldsymbol{CPC_{Câncer}(t_j)}$	$\boldsymbol{CPC_{Outras}(t_j)}$
0,02	0,02	0,00	$\frac{0,02}{1-0} = 0,02$	$\frac{0}{1-0,02} = 0,00$
0,05	0,04	0,00	$\frac{0,04}{1-0} = 0,04$	$\frac{0}{1-0,04} = 0,00$
0,06	0,04	0,02	0,04	0,02
0,09	0,06	0,02	0,06	0,02
0,10	0,08	0,02	0,08	0,02
0,12	0,08	0,04	0,08	0,04
0,19	0,10	0,04	0,10	0,04
0,22	0,12	0,04	0,13	0,05
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
9,87	0,53	0,22	0,67	0,46
10,05	0,53	0,22	0,67	0,46
10,53	0,53	0,22	0,67	0,46
10,70	0,53	0,27	0,72	0,57
11,10	0,53	0,27	0,72	0,57
11,47	0,53	0,27	0,72	0,57
11,55	0,53	0,27	0,72	0,57
11,81	0,53	0,27	0,72	0,57

Podemos, também, visualizar as curvas de probabilidade condicional no *software* R. Porém, como no pacote utilizado ainda é possível fazer isso diretamente, é necessário,

primeiramente, criar a função que gera as curvas através dos seguintes comandos (Pintilie, 2006):

```
CPvar=function(time,cens,group=rep(1,length(time)))
{
#####################################################
# this function calculates                          #
#     the CIF for the event of interest             #
#     the CIF for competing risks (fcr)             #
#     the conditional probability                   #
#     the variance for the conditional probability  #
#           based on Pepe and Mori-s article in     #
#           Statistics in Medicine, 1993.           #
#####################################################
lg=labels(table(group))$group
ng=length(lg)
dd=table(time,cens,group)
tt=sort(unique(time))
for (i in 1:ng)
{
dd1=dd[,,i] ## table of censor, ev, cr
dd2=apply(dd1,1,sum) ## sum for each time point
nrisk=sum(dd2)-cumsum(dd2)
nrisk=c(sum(dd2),nrisk[1:(length(nrisk)-1)])
dev=dd1[,2]
dcr=dd1[,3]
dall=dev+dcr
si=(nrisk-dall)/nrisk
s=cumprod(si)
sminus=c(1,s[1:(length(s)-1)])
fi=dev/nrisk*sminus
f=cumsum(fi)
fcri=dcr/nrisk*sminus
fcr=cumsum(fcri)
fcrminus=c(0,fcr[1:(length(fcr)-1)])
cp=f/(1-fcr)

t1i=dev*(1-fcr)^2
t2i=f^2*(dall-dev)
v1i=(t1i+t2i)/(nrisk*(nrisk-1))
v1i[nrisk==1]=0
v1=cumsum(v1i)
v=sminus^2*v1/((1-fcr)^4)

res=data.frame(time=tt,cif=f,fcr,cp,varCP=v,
group=rep(lg[i],length(v)))
res1=res[dall>0,]
if (i==1) cpvar=res1
else cpvar=list(cpvar,res1)

}

return(cpvar)
}
```

Depois disso, basta chamar a função *CPvar* especificando a coluna onde estão os tempos de sobrevivência e a coluna que indica a ocorrência de evento ou censura:

```
cpc = CPvar(time = tempo., cens = eventos).
```

Para conseguir visualizar o gráfico para o evento de interesse (óbito por câncer), executamos os seguintes comandos (Pintilie, 2006):

```
plot.cp=function(ab,reverse=F,xlab="Time to event",
ylab="Conditional probability")
{
#######################################
## plots the conditional probability  #
## aa is an object created with Cpvar #
#######################################
aa=ab
if (is.data.frame(ab)) aa=list(ab)
ng=length(aa)
m=0
for (i in 1:ng)
{
m=max(m,aa[[i]]$time)
}

for (i in 1:ng)
{
doubletime=sort(c(aa[[i]]$time,aa[[i]]$time))
doublecp=sort(c(aa[[i]]$cp,aa[[i]]$cp))
doubletime=c(0,doubletime)
doublecp=c(0,0,doublecp[1:(length(doublecp)-1)])
if (reverse) doublecp=1-doublecp
if (i==1)
{plot(doubletime,doublecp,xlim=c(0,m),ylim=c(0,1),
xlab=xlab,ylab=ylab,type="l",lty=1)}
else {lines(doubletime,doublecp,lty=i)}
}
}

plot.cp(cpc, xlab = "Tempo", ylab = "Cruva de probabilidade
        condicional")
```

Como resultado obtivemos:

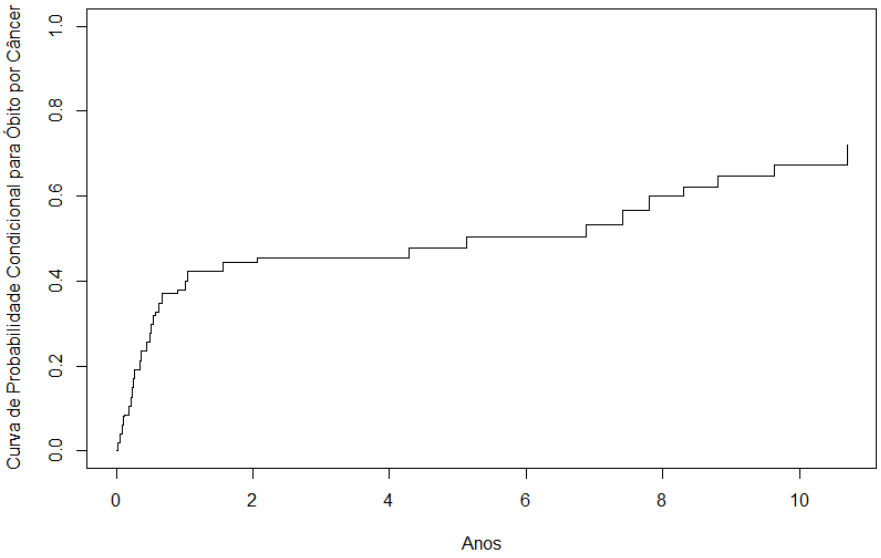


Figura 7 - Curva de probabilidade condicional para o evento óbito por câncer

Por fim, também é possível comparar as curvas entre grupos. Como temos apenas dois (masculino e feminino), utilizaremos o teste de Pepe-Mori.

Iniciaremos pelo cálculo de W, conforme apresentado na equação 35, onde no numerador é necessário estimar $C(t)$ para cada um dos grupos. Para obter esse valor, basta considerar como evento de interesse os casos de censura (representados por c_{k_j}) e proceder normalmente com a estimação da função de sobrevivência. Lembrando que 23 pacientes são do sexo masculino ($N_{Mas} = 23$) e 27 do sexo feminino ($N_{Fem} = 27$). Os passos são detalhados a seguir:

Tabela 14 – Cálculo de $\boldsymbol{W}$ para o cálculo da estatística do teste de Pepe-Mori

t_j	c_{Mas_j}	c_{Fem_j}	$\widehat{C}_{Mas}(t_j)$	$\widehat{C}_{Fem}(t_j)$	$W(t_j)$
0,02	0	0	1,00	1,00	$\frac{(23+27)*1*1}{(23*1)+(27*1)} = 1{,}00$
0,05	0	0	1,00	1,00	$\frac{(23+27)*1*1}{(23*1)+(27*1)} = 1{,}00$
0,06	0	0	1,00	1,00	1,00
0,09	0	0	1,00	1,00	1,00
0,10	0	0	1,00	1,00	1,00
0,12	0	0	1,00	1,00	1,00
0,19	0	0	1,00	1,00	1,00
0,22	0	0	1,00	1,00	1,00
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
9,64	0	0	0,15	0,33	0,61
9,87	0	1	0,15	0,33	0,61
10,05	0	1	0,15	0,33	0,57
10,53	0	1	0,15	0,33	0,51
10,70	0	0	0,15	0,22	0,43
11,10	1	0	0,15	0,22	0,43
11,47	0	1	0,15	0,22	0,31
11,55	0	1	0,15	0,22	0,22
11,81	1	0	0,15	0,00	0,00

Uma vez obtido o valor de $W(t_j)$, os próximos passos consistem em obter o valor da probabilidade condicional para cada tempo e grupo ($\widehat{CPC}_{Mas}$ e $\widehat{CPC}_{Fem}$) e resolver a equação 33, dado que os tempos de observação do exemplo não são discretos. Quanto à probabilidade condicional, o raciocínio para seu cálculo em cada grupo é análogo ao exemplo anterior e, por isso, não detalhado. Pode-se também utilizar a função a *CPvar* para visualizar os resultados de cada grupo. Já em relação à equação 33, iremos novamente **supor o caso de tempo discreto** apenas para apresentar o raciocínio de obtenção da estatística de teste. O resultado correto para o caso do exemplo será apresentado ao final e necessita da implementação de uma função, já que o pacote utilizado ainda não contempla isso.

Tabela 15 – Cálculo da estatística do teste de Pepe-Mori **supondo caso de tempo discreto**

t_j	$W(t_j)$	$\widehat{CPC}_{Mas}(t_j)$	$\widehat{CPC}_{Fem}(t_j)$	$s(t_j)$
0,02	1,00	0,00	0,04	$1{,}00 \times (0{,}00 - 0{,}04) \times (0{,}05 - 0{,}02) = -0{,}0011$
0,05	1,00	0,04	0,04	$1{,}00 \times (0{,}04 - 0{,}04) \times (0{,}06 - 0{,}05) = 0{,}0000$
0,06	1,00	0,04	0,04	0,0001
0,09	1,00	0,04	0,08	-0,0004
0,10	1,00	0,09	0,08	0,0002
0,12	1,00	0,09	0,08	0,0005
0,19	1,00	0,13	0,08	0,0016
0,22	1,00	0,13	0,12	0,0000
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
9,64	0,61	0,80	0,58	0,0325
9,87	0,61	0,80	0,58	0,0251
10,05	0,57	0,80	0,58	0,0624
10,53	0,51	0,80	0,58	0,0201
10,70	0,43	0,80	0,67	0,0232
11,10	0,43	0,80	0,67	0,0216
11,47	0,31	0,80	0,67	0,0032
11,55	0,22	0,80	0,67	0,0076
11,81	0,00	0,80	-	-
				Total $= -0{,}0011 + \ldots + 0{,}0076 = \mathbf{0{,}708}$
				$\boldsymbol{s} = 3{,}524 \times 0{,}708 = \mathbf{2{,}495}$

Demonstrados os passos necessários para obtenção da estatística de teste supondo o caso discreto (lembre-se que esse valor não é válido uma vez que nesse exemplo o tempo não é discreto), mostraremos como obter o valor da equação 33 com ajuda computacional. Para isso, é necessário criar a função que compara as curvas de probabilidade através dos seguintes comandos (Pintilie, 2006):

```
compCP=function(time,cens,group=rep(1,length(time)))
{
#########################################################
# this function compares the CP-s of two groups         #
# it is based on the test presented in the paper        #
# from Statistics in Medicine, 1993, by Pepe and Mori   #
#########################################################
ttau=unlist(by(time,group,max))
tau=min(ttau)
ng=table(group)
lg=labels(ng)$group
dd=table(time,cens,group)
if (dim(dd)[3]!=2) stop("Pepe-Mori test is for two groups")
if (dim(dd)[2]>3) stop("All competing risks should be grouped under  code
2")
```

```
if (dim(dd)[2]<=2) stop("There are either no censored obs or \nno competing
risks or no events of interest")
tt=sort(unique(time))
tt1=c(tt[2:length(tt)],NA)
nt=table(tt1<=tau)[2]
deltat=tt1-tt
dd=dd[1:nt,,]
deltat=deltat[1:nt]
tt=tt[1:nt]

for (i in 1:2)
{

dd1=dd[,,i] ## table of censor, ev, cr
dd2=apply(dd1,1,sum) ## sum for each time point
nrisk=ng[i]-cumsum(c(0,dd2[1:(nt-1)]))
dev=dd1[,2]
dcr=dd1[,3]
dcens=dd1[,1]+dcr
dall=dev+dcr
si=(nrisk-dall)/nrisk
s=cumprod(si)
sminus=c(1,s[1:(length(s)-1)])
fi=dev/nrisk*sminus
f=cumsum(fi)
fcri=dcr/nrisk*sminus
fcr=cumsum(fcri)
cp=f/(1-fcr)
Ci=(nrisk-dcens)/nrisk
C=cumprod(Ci)
Cminus=c(1,C[1:(length(C)-1)])

if (i==1)
{nrisk1=nrisk
s1=s
f1=f
fcr1=fcr
C1=C
CP1=cp
C1minus=Cminus
dall1=dall
dev1=dev}
if (i==2)
{nrisk2=nrisk
s2=s
f2=f
fcr2=fcr
C2=C
CP2=cp
C2minus=Cminus
dall2=dall
dev2=dev}

}

wi=C1minus*C2minus*sum(ng)/(ng[1]*C1minus+ng[2]*C2minus)
si=wi*(CP1-CP2)*deltat
s=sqrt(ng[1]*ng[2]/sum(ng))*sum(si)

# for group 1
temp=wi*s1*deltat/(1-fcr1)**2
```

```
temp[is.na(temp)]=0
tparti=rev(cumsum(rev(temp)))

sigma1i=tparti**2*(dev1*(1-fcr1)**2+(dall1-dev1)*f1**2)/(nrisk1*(nrisk1-1))
sigma1=sum(sigma1i,na.rm=T)
# for group 2
temp=wi*s2*deltat/(1-fcr2)**2
temp[is.na(temp)]=0
tparti=rev(cumsum(rev(temp)))

sigma2i=tparti**2*(dev2*(1-fcr2)**2+(dall2-dev2)*f2**2)/(nrisk2*(nrisk2-1))
sigma2=sum(sigma2i,na.rm=T)
sigma=ng[1]*ng[2]*(sigma1+sigma2)/sum(ng)
z=s^2/sigma
pvalue=1-pchisq(z,1)
r=data.frame(chisquare=z,pvalue=pvalue)
row.names(r)=""
return(r)
}
```

Depois disso, basta chamar a função *compCP*, especificando: a coluna onde estão os tempos de sobrevivência; a coluna que indica a ocorrência de evento ou censura; e a coluna que determina os grupos que serão comparados.

```
teste_PepeMori = compCP(time = tempo., cens = eventos, group = sexo)
```

Para conseguir visualizar o resultado do teste de Pepe-Mori, basta chamar o objeto em que o teste foi armazenado. Como resultado, obteremos:

```
> teste_PepeMori
 chisquare   pvalue
 0.2705382 0.602971
```

Figura 8 - Resultado do teste de Pepe-Mori executado via *software* R

Usando a mesma função vista anteriormente (*plot.cp*), podemos visualizar as curvas de probabilidade condicional, considerando como evento de interesse o óbito por câncer para os dois grupos. O resultado obtido foi:

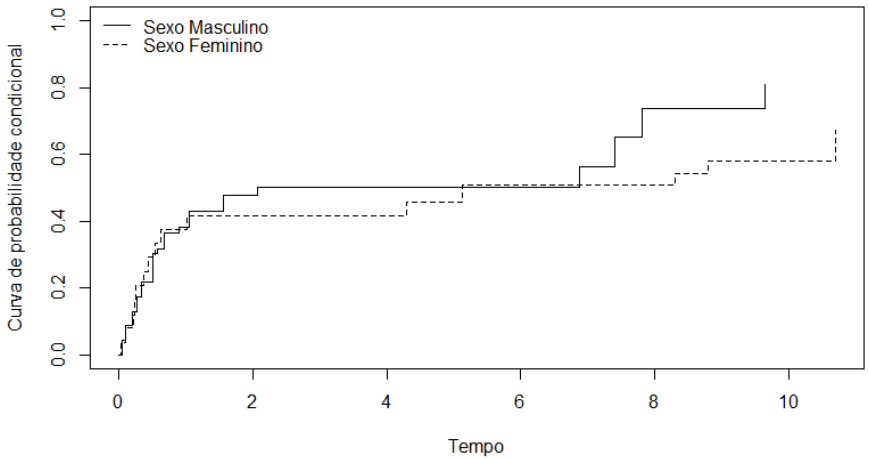


Figura 9 - Curvas de probabilidade condicional considerando os sexos masculino e feminino e o óbito por câncer como evento de interesse

Conforme sugerido pela Figura 9 e, confirmado pelo teste de Pepe-Mori (p = 0,603), não há evidencias estatísticas significativas para afirmar que exista diferença entre as curvas de probabilidade condicional de homens e mulheres, considerando o óbito por câncer como evento de interesse.

3.1.2.2.3. Modelo para a subdistribuição do risco

O foco da discussão desta seção está em situações onde há riscos competitivos e o interesse consiste em estimar o efeito de covariáveis, considerando os possíveis eventos.

Na literatura, há várias formas de se analisar problemas que contemplem riscos competitivos, todavia as abordagens mais comuns são: sobrevivência até o primeiro entre os eventos; risco específico por causa; e modelagem da subdistribuição dos riscos (Carvalho et al., 2011).

O primeiro método considera o tempo até o primeiro dos possíveis eventos, independente de qual deles tenha ocorrido. Do ponto de vista prático, faz sentido em situações onde se estudam vários possíveis efeitos de um mesmo fator de risco (Carvalho et al., 2011). Por exemplo, uma análise sobre o efeito do tabagismo em relação à ocorrência dos seguintes eventos: óbito por câncer de pulmão; e óbito por bronquite crônica e enfisema pulmonar. Porém, esse método exige que a variável independente esteja associada aos eventos competitivos na mesma direção, risco ou proteção, caso contrário, seu resultado não é interpretável (Carvalho et al., 2011).

O segundo método, risco específico por causa, busca estimar o efeito de uma covariável sobre um evento específico, tratando os demais possíveis eventos como censura não informativa, mesmo que não o sejam. A limitação dessa metodologia consiste nos seguintes pontos: não é possível estimar o efeito comum de qualquer covariável para todos os eventos competitivos (por exemplo: o efeito do tabagismo para o óbito por câncer de pulmão; por bronquite crônica e enfisema pulmonar; ou qualquer outra possível causa de óbito); a estimação da curva de sobrevivência estará inadequada, uma vez que desconsidera os eventos competitivos; e, por fim, o método exige a suposição de que o tempo entre o evento de interesse e os competitivos seja independente para que as estimativas de associação sejam válidas, o que, frequentemente, não condiz com a realidade (Carvalho et al., 2011).

A terceira abordagem, desenvolvida por Fine e Gray (Fine e Gray, 1999), propõe uma alternativa ao método do risco específico por causa usando a subdistribuição dos riscos (função de incidência acumulada), matematicamente definida por:

$$h_q(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, D = q | T \geq t \cup (T \leq t \cap D \neq q)]}{\Delta t}$$
$$h_q(t) = -\frac{\partial \ln(1 - F_q(t))}{\partial t}, \quad (40)$$

onde h_q é a função de risco da subdistribuição considerando o evento q e $F_q(t)$ é a função de incidência acumulada para o evento q (Callaghan, 2008). O modelo considerado nessa abordagem pode ser escrito como:

$$h_q(t) = h_0(t)\exp(\boldsymbol{\beta x}) \quad (41)$$

sendo h_0 a função de risco basal, $\boldsymbol{x}$ o vetor de covariáveis e $\boldsymbol{\beta}$ o vetor de coeficientes que será estimado (Carvalho et al., 2011).

A forma da função de verossimilhança parcial utilizada para estimação dos parâmetros $\boldsymbol{\beta}$ desse modelo é similar a que foi apresentada no modelo de Cox (seção 1.2.1.2), podendo ser escrita como:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{exp(\boldsymbol{\beta x_j})}{\sum_{i \epsilon R_j} w_{ji} exp(\boldsymbol{\beta x_i})}, \quad (42)$$

sendo que esse produto é realizado sobre todos os m ($t_1 < t_2 < \cdots < t_m$) tempos onde um evento de interesse tenha ocorrido (Pintilie, 2006).

A grande diferença entre as funções de verossimilhança do modelo de Cox (equação 19) e da abordagem da subdistribuição do risco (equação 42) consiste no grupo de risco (R_j). Na primeira, o indivíduo é removido do grupo de risco depois de experimentar qualquer um dos possíveis eventos no tempo t. Já na abordagem da subdistribuição dos riscos para o evento q, somente os pacientes que sofreram o evento q no tempo t são removidos e todos os demais indivíduos permanecem no grupo de risco, inclusive os que sofreram outros eventos diferentes de q no tempo t (Callaghan, 2008).

A fim de evitar viés nas estimativas, os pacientes que sofreram outros eventos, que não q, permanecem no grupo de risco, sendo que o grau com que eles contribuem para esse grupo é dado por uma **função peso** que decresce com o tempo (Callaghan, 2008; Carvalho et al., 2011). Essa função peso, representada por $w_l(t_j)$, é matematicamente definida por (Carvalho et al., 2011; Pintilie, 2006):

$$w_l(t_j) = \begin{cases} 1, & \text{se o paciente não tiver sofrido evento ou censura;} \\ \dfrac{\hat{G}_{KM}(t_j)}{\hat{G}_{KM}(t_i)}, & \text{se o paciente sofrer o evento competitivo em } t_i < t_j; \\ 0, & \text{se o paciente for retirado do estudo.} \end{cases} \qquad (43)$$

Em termos práticos, receber peso zero equivale a retirar o paciente do banco de dados em análise a partir do momento em que ele sofre o evento de interesse q ou é censurado por qualquer motivo que não os demais possíveis eventos competitivos (Carvalho et al., 2011).

A ponderação decrescente é definida a partir da curva de sobrevivência das censuras, sendo classificada como censura tudo o que não for evento. Isso significa que, para o cálculo de $G(.)$, censura vira evento e evento vira censura e a função de sobrevivência $\hat{G}_{KM}$ é estimada via método de Kaplan-Meier (Carvalho et al., 2011). Assim, no momento em que um evento competitivo ocorre em $t_i < t_j$, o peso w_{ji} terá: no numerador, a função de sobrevivência das censuras a partir do ponto t_i onde o paciente sofreu o evento competitivo e, no denominador, o valor estimado para a sobrevivência das censuras no tempo t_i em que o paciente sofreu o evento competitivo. Esse valor será, portanto, sempre o mesmo para cada paciente (Carvalho et al., 2011).

A fim de estimar o valor de $\boldsymbol{\beta}$, representado por $\widehat{\boldsymbol{\beta}}$, é necessário maximizar o valor logaritmo da função de verossimilhança parcial, etapa que necessita de apoio computacional (Pintilie, 2006).

De forma resumida, a contribuição para o grupo em risco dos pacientes que sofreram o evento competitivo é tanto menor quanto mais distante do evento de interesse, obedecendo a estimativa de decaimento no tempo dada pela curva de Kaplan-Meier das censuras (Carvalho et al., 2011).

Ao pesquisador, cabe o cuidado com a definição de censura, visto que essas podem ser tratadas como eventos competitivos. É necessário garantir que as observações censuradas sejam realmente independentes do evento estudado (Carvalho et al., 2011).

Uma vez estimado o modelo, a função de incidência acumulada pode ser estimada para determinados valores das covariáveis usando a seguinte fórmula (Pintilie, 2006):

$$F(t) = 1 - exp\,(H(t)), \tag{44}$$

onde $H(t)$ é o risco acumulado da subdistribuição. Para o cálculo de $\widehat{H}(t)$ utiliza-se o estimador de Breslow, conforme mostrado a seguir (Pintilie, 2006):

$$\widehat{H}\left(t;\, x_0, \hat{\beta}\right) = \sum_{t_j \le t} \left\{ \frac{exp\,(x_0\hat{\beta})}{\sum_{i \in R_j} w_{ji} exp\,(x_i\hat{\beta})} \right\}. \tag{45}$$

As principais vantagens do método desenvolvido por Fine e Gray, que será o método utilizado no presente trabalho, são: a utilização da função de incidência acumulada, uma vez que isso facilita a interpretação do efeito das covariáveis; e o fato de que não é necessária a suposição de eventos competitivos independentes (Callaghan, 2008; Carvalho et al., 2011). Dentre as desvantagens da técnica, destaca-se a suposição de riscos proporcionais (Callaghan, 2008).

3.1.2.2.3.1. Praticando os conceitos abordados

Utilizaremos os dados apresentados na Tabela 8 para exemplificar a obtenção dos coeficientes do modelo que considera a abordagem da subdistribuição do risco.

O objetivo do exemplo é verificar o efeito da covariável *sexo* para cada um dos eventos estudados, considerando os riscos competitivos. Ou seja, desejamos estimar:

$$h_q(t) = h_0(t)\exp(\beta x)$$

onde x, nesse caso, representa a covariável *sexo* e β, o valor do coeficiente associado a essa covariável.

Para tal, é necessário estimar o valor dos pesos w_{ji} e β, o que envolve o processo de maximização do logaritmo natural da função de verossimilhança parcial, sendo necessário o auxílio computacional. Em função disso, apresentaremos os comandos necessários para obter tais valores utilizando o *software* R.

Para obter o modelo baseado na subdistribuição dos riscos, utilizaremos a função *crr*, do pacote *cmprsk*. No caso mais simples, como o que mostraremos a seguir, é necessário especificar como argumentos da função: a coluna do banco de dados que especifica o tempo de acompanhamento (*Time*); uma coluna que identifique o tipo de evento sofrido pelo paciente (*Status*); colunas indicando as covariáveis de interesse para ajuste do modelo (*x*); valor do código que identifica o evento de interesse (*failcode*); e o valor do código que identifica a censura (*cencode*).

Dessa forma, os comandos necessários para ajustar o modelo são:

```
require(cmprsk)
modelo = crr (Ttime, Status, x, failcode, cencode)
```

O resultado do modelo ajustado pode ser visto utilizando a função *summary()*, cujo argumento é o objeto onde o modelo foi armazenado. No exemplo, obtemos:

```
> summary(modelo)
Competing Risks Regression

Call:
crr(ftime = dadosfia$Tempo, fstatus = dadosfia$Eventocat, cov1 = dadosfia$Sexo,
    failcode = 1, cencode = 0)

                coef exp(coef) se(coef)      z p-value
dadosfia$Sexo1 -0.27     0.763    0.392 -0.689    0.49

               exp(coef) exp(-coef)  2.5% 97.5%
dadosfia$Sexo1     0.763       1.31 0.354  1.65

Num. cases = 50
Pseudo Log-likelihood = -89
Pseudo likelihood ratio test = 0.45  on 1 df,
```

Figura 10 - Saída do R para ajuste do modelo baseado na subdistribuição considerando os dados da Tabela 8

A primeira parte da saída do modelo mostra o valor estimado para a covariável *sexo* (*coef*); o risco relativo (*exp(coef)*); o erro padrão (*se(coef)*); o valor da estatística *z* (*z*); e o valor de p correspondente (*p-value*). Considerando os dados desse exemplo, vemos que a covariável sexo não foi significativa (p = 0,49).

Já a segunda parte da saída apresenta o risco relativo para a covariável em análise (*exp(coef)*) e seu intervalo de 95% de confiança. Esse nível de confiança pode ser alterado para outros valores, basta especificar o valor desejado no argumento *conf.int* da função *summary()*). O risco relativo, ou razão de risco da subdistribuição, para uma covariável categórica, como é o caso, é calculado através da razão dos riscos da subdistribuição para o grupo em relação à linha de base, com todas as outras covariáveis, caso houver, sendo fixas. No caso de covariáveis contínuas, o risco relativo refere-se ao efeito do aumento de uma unidade na covariável, mantendo todas as outras covariáveis iguais (Scrucca, Santucci e Aversa, 2010). Assim, o valor de *exp(coef)*, nesse caso, demonstra o risco relativo dos homens em relação as mulheres, uma vez que o grupo de referência é o sexo 1 (masculino).

Por fim, a última parte mostra o número de casos, o logaritmo da pseudo-verossimilhança no máximo e o teste da razão da pseudo-verossimilhança. Os dois últimos têm como função verificar a diferença da função objetivo (função matemática que define a qualidade da solução em função das covariáveis) no zero global e nas estimativas finais. Como essa função objetivo não é a verossimilhança verdadeira, essa estatística de teste não tem uma distribuição assintótica χ^2 e, em consequência, a comparação de modelos baseada na abordagem da razão de verossimilhanças não pode ser realizada diretamente (Scrucca, Santucci e Aversa, 2010). Contudo, já existem metodologias propostas que resolvem esse problema, como o estudo de Scrucca, Santucci e Aversa (2010), por exemplo.

3.2. RELACIONAMENTO DE BASES DE DADOS

Relacionamento de bases de dados (*record linkage*) é a solução utilizada em problemas de reconhecimento de registros em dois arquivos diferentes que representam um mesmo indivíduo (Fellegi e Sunter, 1969). Um exemplo seria a identificação de óbitos dentre os pacientes com diagnóstico de câncer registrado em um determinado hospital utilizando a base dados desse hospital e buscando, dentre esses pacientes, quais tiveram registro de óbito no Sistema de Informações de Mortalidade (SIM). De forma geral, o relacionamento de bases de dados busca obter uma única base, que combina informações parciais existentes em cada arquivo original para construção de um arquivo único completo e com informações atualizadas (De Sousa et al., 2008).

O relacionamento de registros em diferentes bases de dados torna-se uma tarefa simples em casos onde cada uma das bases contenha pelo menos um campo em comum que seja pouco sujeito a erros: o número da Carteira de Identidade, por exemplo. Nesse caso, os arquivos poderiam ser comparados utilizando o chamado **relacionamento determinístico** (*deterministic record linkage*) (Coutinho e Coeli, 2006; De Sousa et al., 2008; Mcdonald). Já em situações onde isso não existe, é necessário recorrer a técnicas alternativas, tais como o método probabilístico, que utiliza campos em comum nas bases a serem relacionadas para localizar os pares correspondentes, chamado de **relacionamento probabilístico de registros** (*probabilistic record linkage*). Tais campos compõem um escore que representa o grau de concordância entre os registros de cada par formado (Coutinho e Coeli, 2006; Mcdonald).

O método determinístico é a técnica mais simples de relacionamento entre bases de dados e refere-se a situações onde se olha para a concordância exata de um ou mais campos entre os arquivos (Blakely e Salmond, 2002; Mcdonald). Em casos de diferença do campo em comum entre os arquivos (erro de digitação, por exemplo), o registro é perdido (Blakely e Salmond, 2002).

Já no caso do método probabilístico, utilizam-se informações de um número maior de campos em que cada um deles é comparado e atribuído a um escore baseado no quão bem os pares comparados são correspondentes. Esse escore calculado para cada campo indica, para qualquer par de registros, o quão provável é que ele se refira ao mesmo indivíduo. A partir desses escores, é obtido um escore total composto pela soma dos escores gerados a partir dos pares individuais correspondentes. Os pares de registros são então ordenados e um ponto de corte é utilizado para distinguir as comparações boas e as ruins (Blakely e Salmond, 2002).

Neste trabalho, serão utilizados dois arquivos: base de dados de diagnósticos de câncer entre os anos de 2002 e 2009 do HCPA; e base de dados do SIM para localização de óbitos ocorridos entre os anos de 2002 e 2013. Entre elas não há um campo que identifique de forma direta cada registro. Desse modo, será utilizada a técnica de relacionamento probabilístico de registros com a finalidade de localizar casos de óbito dentre os pacientes da coorte.

3.2.1. Relacionamento probabilístico de registros

Um dos pesquisadores pioneiros a desenvolver metodologias para o relacionamento automático de registros foi Newcombe, cujo artigo elaborado, com participação de outros autores, foi publicado no ano de 1959 (De Camargo Junior e Coeli, 2000; Newcombe et al., 1959). Alguns anos depois, Fellegi e Sunter (Fellegi e Sunter, 1969) estenderam os conceitos originais e propuseram um tratamento matemático formal à metodologia, atualmente conhecida como relacionamento probabilístico de registros. Tal metodologia é baseada em três processos: **padronização**, **blocagem** e **pareamento** de registros (De Camargo Junior e Coeli, 2000).

A **padronização** de registros é o passo inicial a ser realizado em campos não estruturados (nome, por exemplo), tendo por objetivo minimizar a ocorrência de erros durante o processo de pareamento dos registros. Procedimentos comuns nessa etapa envolvem a eliminação de caracteres especiais, como acento e pontuação, espaços em branco e transformação dos caracteres minúsculos para maiúsculos (De Camargo Junior e Coeli, 2000).

Na etapa de **blocagem**, busca-se criar blocos lógicos de registros dentro dos arquivos que serão relacionados a fim de otimizar o processo de pareamento. Dessa forma, os arquivos são divididos em blocos mutuamente exclusivos, o que limita as comparações aos registros que pertencem ao mesmo bloco, sendo que os blocos são formados com o objetivo de aumentar a probabilidade de que seus registros representem pares verdadeiros (De Camargo Junior e Coeli, 2000). Neste trabalho, será utilizado como bloco o ano de diagnóstico de câncer, ou seja, buscar-se-á o óbito de pacientes com diagnóstico em 2002 na base de dados do SIM dos anos de 2002 a 2013; já para pacientes com diagnóstico em 2003, buscar-se-á o óbito na base de dados do SIM dos anos de 2003 a 2013 e assim por diante, o que reduz o espaço de busca e, consequentemente, o tempo e número de comparações.

O passo final é o **pareamento** dos registros baseado na construção de escores para os diferentes pares possíveis de serem obtidos a partir da estratégia de blocagem definida na segunda etapa do processo (De Camargo Junior e Coeli, 2000).

Da mesma forma que o método de relacionamento probabilístico de registros, a ideia de escore iniciou-se com Newcombe (Newcombe et al., 1959) e foi aprimorada por Fellegi e Sunter (Fellegi e Sunter, 1969), que também propuseram o conceito de **escore limiar** para classificação dos pares segundo três classificações: verdadeiros, falsos e duvidosos. De forma prática, isso equivale dizer que: determina-se, a priori, um limiar superior e um inferior; valores acima do limiar superior predeterminado serão classificados como verdadeiros; valores abaixo do limiar inferior predeterminado serão classificados como falsos; e os valores cujo escore estiver entre os limiares definidos devem ser revisados de forma manual (De Camargo Junior e Coeli, 2000).

O **escore total** gerado para quaisquer pares de registros será igual à soma dos escores ponderados gerados pelo pareamento individual de cada campo utilizado no processo de pareamento. Assim, cada campo contribui de maneira diferente para o escore total do par, sendo essa contribuição distinta altamente recomendável, já que os campos apresentam diferente poder discriminatório e possuem maior ou menor probabilidade de ter seus conteúdos registrados de forma diferente (De Camargo Junior e Coeli, 2000; Mcdonald).

A construção dos escores, coração do método de relacionamento probabilístico de registros, é baseada em conceitos de testes diagnósticos utilizados pelos epidemiologistas. Para cada campo i, definem-se: $\boldsymbol{m_i}$, probabilidade de o campo concordar entre os dois pares, dado que se trata de um par verdadeiro; e $\boldsymbol{u_i}$, probabilidade de o campo concordar, visto que se trata de um par falso (Blakely e Salmond, 2002; De Camargo Junior e Coeli, 2000). Ainda utilizando a linguagem de testes diagnósticos, poderia se dizer que m_i representa a probabilidade do campo i identificar um par como verdadeiro quando ele de fato é verdadeiro (sensibilidade) e u_i representa a probabilidade do campo i identificar um par como verdadeiro quando ele na verdade é falso ($1 - especificidade$) (De Camargo Junior e Coeli, 2000).

Com as duas quantidades (m_i e u_i) estimadas, constroem-se dois **fatores de ponderação**, um para o caso de concordância, quando o campo do primeiro registro concorda com o campo do segundo registro, e um para o caso de discordância, quando o campo do primeiro registro não concorda com o campo do segundo registro (De Camargo Junior e Coeli, 2000). O **fator de ponderação de concordância** é matematicamente definido por:

$$wc_i = log_2 \left[\frac{m_i}{u_i}\right], \quad (46)$$

onde wc_i representa o fator de ponderação de concordância do campo i e $\log_2$ representa o logaritmo de base 2. Já o **fator de ponderação de discordância** é calculado da seguinte forma:

$$wd_i = log_2 \left[\frac{(1 - m_i)}{(1 - u_i)}\right], \quad (47)$$

sendo que wd_i representa o fator de ponderação de discordância do campo i (De Camargo Junior e Coeli, 2000).

O conjunto de probabilidades m_ie u_i, e seus correspondentes fatores de ponderação, são repetidos para todas as variáveis correspondentes. Dessa forma, chega-se ao escore total de determinado par de comparação, que é dado simplesmente pela soma dos fatores de (dis)concordância para cada campo avaliado (Blakely e Salmond, 2002; De Camargo Junior e Coeli, 2000). Esse escore será um número grande e positivo se todos, ou grande parte, das variáveis correspondentes concordarem, e será um número grande e negativo se todos, ou grande parte, das variáveis correspondentes discordarem (Blakely e Salmond, 2002).

Alguns pesquisadores alertam para o fato de que nem sempre a decisão sobre a concordância ou discordância entre dois campos de dado par é tarefa simples, o que acaba por dificultar a escolha de qual fator de ponderação deve ser atribuído como resultado da comparação de dois campos (De Camargo Junior e Coeli, 2000; Jaro, 1989). A partir disso, Jaro (Jaro, 1989) propõe, para casos onde o fator de ponderação de discordância é pequeno, atribuir o fator de ponderação de concordância, mas não de forma integral. Isso significa atribuir um valor que contribuirá positivamente para o escore total, porém cuja contribuição será menor do que aquela que seria utilizada no caso de concordância exata. A definição de fator de discordância pequeno dependerá do tipo de campo avaliado e do algoritmo de comparação utilizado.

Complementa-se que, da mesma forma que os valores de limiar superior e inferior são determinados pelo pesquisador, os valores de m_i e u_i também podem ser (De Camargo Junior e Coeli, 2000). Alguns autores sugerem que valores em torno de 0,9 para m_i e 0,1 para u_i funcionam bem na prática, embora a natureza do campo deva ser avaliada para determinação

de tal valor. Para sexo, por exemplo, sugere-se utilizar u_i igual a 0,5 (De Camargo Junior e Coeli, 2000).

Finalizando, destaca-se que o objetivo do método de relacionamento de registros é encontrar correspondentes e um relacionamento acurado dependente, principalmente, da quantidade de poder discriminatório inerente às variáveis comuns para os registros que precisam ser correspondidos e da boa qualidade dos dados (Blakely e Salmond, 2002; Mcdonald).

3.2.1.1. Praticando os conceitos abordados

Para exemplificar o processo de relacionamento probabilístico, reproduziremos o exemplo apresentado por Clark (2004): suponha que você tenha os dados apresentados nas Tabelas 16 e 17 e precise decidir quais casos da ambulância correspondem aos casos atendidos pelo departamento de emergência.

Tabela 16 - Dados hipotéticos de 10 registros de atendimentos realizados por uma ambulância

Caso	Ano	Dia	Hospital	Ano de nascimento	Aniversário	Sexo
A01	1	01/jan	X	1950	21/jan	Mas
A02	1	01/jan	X	1950	01/mai	Fem
A03	1	10/jan	Y	1975	27/dez	
A04	1	13/ago	X	1977	29/abr	Fem
A05	1	12/set	Y	1980	16/fev	Fem
A06	1	31/dez	Z	1919	16/set	Mas
A07	2	02/fev	X	1924	26/mar	Fem
A08	2	10/jun	Y	1951	29/mar	Mas
A09	2	06/ago	Y	1953	17/abr	
A10	2	21/set	Z	1956	03/jun	Fem

Como o número de registros é pequeno, nesse caso, poderíamos fazer o processo de relacionamento das duas bases de dados usando nosso próprio julgamento, baseado em experiências passadas com esse tipo de pacientes e registros. Entretanto, serão demonstrados os passos do método probabilístico considerando as seguintes suposições baseadas nas experiências anteriores ou em cálculos que consideram a maior base de dados (Tabela 18):

i. Aproximadamente 90% dos pacientes feridos transportados por ambulâncias geram uma ficha no departamento de emergência do hospital. Assim, a probabilidade, a

priori, de uma correspondência entre registros das duas bases pode ser algo em torno de 0,045 – $P(Corresponder) = \frac{(0,9 \times 10)}{10} \times \frac{1}{20} = 0,045$ – e a probabilidade de que dois registros se refiram ao mesmo paciente é 0,047 – $P(Ser\ o\ mesmo\ paciente) = \frac{0,45}{1-0,45} = 0,047$.

ii. As probabilidades de que os registros sejam verdadeiramente do mesmo ano de ingresso, data de ingresso, hospital, ano de nascimento, data de nascimento e sexo (probabilidades m_i) são: 0,99, 0,95, 0,99, 0,95, 0,99 e 0,95, respectivamente.

iii. As probabilidades de que as variáveis correspondam aos registros selecionados aleatoriamente (probabilidades u_i) são: 0,50 para o ano de ingresso; 0,0027 ($1/365$) para a data de ingresso; 0,40 para o hospital X ou Y; 0,20 para o hospital Z; 0,01 para o ano de nascimento; 0,0027 para a data de nascimento; 0,60 para sexo masculino (Mas); e 0,40 para sexo feminino (Fem).

Tabela 17 – Dados hipotéticos de 20 registros de atendimentos realizados em uma emergência

Caso	Ano	Dia	Hospital	Ano de nascimento	Aniversário	Sexo
E01	1	01/jan	X	1950	21/jan	Mas
E02	1	10/jan	Z	1987	17/jul	Mas
E03	1	23/fev	X	1992	19/out	Mas
E04	1	22/abr	Y	1979	09/mai	Mas
E05	1	02/mai	X	1929	12/nov	Fem
E06	1	23/mai	Y	1964	01/dez	Mas
E07	1	01/jun	X	1950	01/mai	Fem
E08	1	14/ago	X	1977	29/abr	Fem
E09	1	12/set	Y	1980	19/fev	Fem
E10	1	21/out	Y	1985	12/mar	Mas
E11	2	01/jan	Z	1919	16/set	Mas
E12	2	10/jan	Y	1975	27/dez	Fem
E13	2	02/fev	X	1924	26/mar	Fem
E14	2	16/mai	X	1924	12/out	Mas
E15	2	10/jun	Y	1951	29/mar	Mas
E16	2	04/jul	Z	1982	12/jun	Mas
E17	2	05/ago	Y	1953	17/abr	Mas
E18	2	06/ago	Y	2002	17/abr	Fem
E19	2	21/set	Z	1956	03/jun	Fem
E20	2	22/nov	X	1917	29/mai	Mas

Comparando as duas bases de dados, percebemos que os registros A10 e E19 parecem se referir exatamente a mesma pessoa, uma vez que todos os campos são iguais, e possuem, por essa razão, fator de ponderação de concordância alto, que é calculado de acordo com a equação 46:

$$log_2\left[\frac{0{,}99}{0{,}50}\right] + log_2\left[\frac{0{,}95}{0{,}0027}\right] + log_2\left[\frac{0{,}99}{0{,}01}\right] + log_2\left[\frac{0{,}95}{0{,}01}\right] + log_2\left[\frac{0{,}99}{0{,}0027}\right] + log_2\left[\frac{0{,}95}{0{,}40}\right] = 32{,}41.$$

Outros registros com alto fator de ponderação de concordância são os pares: A01 e E01; A05 e E09; A07 e E13; A08 e E15. Por outro lado, há pares em que um ou mais elementos não correspondem, como é o caso dos registros A03 e E12. Calculando o fator de ponderação para esse caso de acordo com as equações 46 e 47 obtemos:

$$log_2\left[\frac{1-0{,}99}{1-0{,}50}\right] + log_2\left[\frac{0{,}95}{0{,}0027}\right] + log_2\left[\frac{0{,}99}{0{,}04}\right] + log_2\left[\frac{0{,}95}{0{,}01}\right] + log_2\left[\frac{0{,}99}{0{,}0027}\right] + log_2\left[\frac{1-0{,}95}{1-0{,}40}\right] = 18{,}95.$$

Nesse exemplo, o pressuposto de que quase todos os registros da ambulância devem ter uma ficha correspondente no departamento de emergência impacta em probabilidades de correspondência altas. Deve-se lembrar, porém, que podem existir comparações que necessitam de revisão manual: são os casos em que o escore total fica entre os limiares definidos pelo pesquisador. Poderíamos considerar o par A06 e E19 com um exemplo que necessitaria de revisão manual.

Em um banco de dados com muitas observações, realizar esse trabalho de forma manual certamente torna-se inviável. Nesse sentido, já existem diversos *softwares* disponíveis que relacionam campos de dois ou mais bancos de dados tais como o Link Plus, ReckLink, Link King e FRIL, sendo que o último foi o utilizado no desenvolvimento desse trabalho.

4. OBJETIVOS

4.1. GERAL

Aplicar a metodologia de riscos competitivos para estimar a letalidade e fatores associados ao óbito de pacientes diagnosticados com câncer primário entre os anos de 2002 e 2009 e atendidos no Hospital de Clínicas de Porto Alegre (HCPA).

4.2. ESPECÍFICOS

a) Descrever a metodologia utilizada em análise de sobrevivência na presença de riscos competitivos;
b) Avaliar incidências, letalidades e fatores associados ao óbito de pacientes diagnosticados com câncer primário e atendidos pelo HCPA entre 2002 e 2009 usando a abordagem de riscos competitivos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Balakrishnan N, Rao CR. Handbook of Statistics: Advances in Survival Analysis. 2004; 23.

Blakely T, Salmond C. Probabilistic Record linkage and a method to calculate the positive predictive value. International Journal of Epidemiology, 2002; 31: 1246-52.

Callaghan F M. Classification trees for survival data with competing risks [dissertation]. Department of Biostatistics: University of Pittsburgh; 2008.

Carvalho MS, Androzzi VL, Codeço CT, Campos DP, Barbosa MTS, Shimakura SE. Análise de Sobrevivência: Teoria e aplicações em saúde. Rio de Janeiro: Fiocruz; 2011.

Clark DE. Practical introduction to record linkage for injury research. Injury Prevention, 2004; 10:186-91.

Collet D. Modelling Survival Data in Medical Research. 2nd ed. London: Chapman & Hall; 2003.

Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevivência. Cadernos de Saúde Pública, 2006; 22: 2249-52.

De Camargo Junior KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cadernos de Saúde Pública, 2000; 16: 439-47.

De Glass NA, et al. Performing Survival Analyses in the Presence of Competing Risks: A Clinical Example in Older Breast Cancer Patients. Journal of the National Cancer Institute, 2015; 108(5).

De Sousa MH, Cecatti JG, Hardy E, Serruya SJ. Relacionamento probabilístico de registros: uma aplicação na área de morbidade materna grave (near miss) e mortalidade materna. Cadernos de Saúde Pública, 2008; 24: 653-62.

Dignam JJ, Kocherginsky MN. Choice and interpretation of statistical tests used when competing risks are present. Journal of Clinical Oncology. 2008; 26:4027-34.
Dos Santos Junior PC. Análise de Sobrevivência na presença de Censura Informativa [dissertação]. Departamento de Estatística: Universidade Federal de Minais Gerais; 2012.

Fellegi IP, Sunter A B. A theory for record linkage. Journal of the American Statistical Association, 1969; 64:1183-1210.

Ferlay J, et al. GLOBOCAN 2012 v1.0. Estimated cancer incidence, mortality and prevalence worldwide in 2012 [Internet]. Lyon: International Agency for Research on Cancer [citado em 22 de agosto de 2015]. Disponível em: http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association, 1999; 94:496-509.

Giordani NE. Aplicação do modelo de riscos competitivos em pacientes diagnosticados com câncer no ano de 2006 no Hospital de Clínicas de Porto Alegre [monografia]. Departamento de Estatística: Universidade Federal do Rio Grande do Sul; 2013.

Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics. 1988; 16:1141-54.

Hosmer DW, Lemeshow S. Applied Survival Analysis: Regression Modeling of Time To Event Data. New Jersey: Wiley Series in Probability and Statistics; 1999.

Instituto Nacional de Câncer José Alencar Gomes da Silva. Coordenação de Prevenção e Vigilância. Estimativa 2014: Incidência de Câncer no Brasil. Rio de Janeiro: INCA, 2014. Disponível em: http://www.inca.gov.br/estimativa/2014/estimativa-24042014.pdf

Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the America Statistical Association, 1989; 84:414-20.

Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. 2nd ed. New York: Springer Science; 2005.

Lee ET, Wang JW. Statistical Methods for Survival Data Analysis. 3th ed. New Jersey: Wiley Series in Probability and Statistics; 2003.

Lunn, M. (1998). Applying k-sample tests to conditional probabilities for competing risks in a clinical trial. Biometrics, 54, 1662–1672.

Mcdonald JM. An Introduction to Probabilistic Record Linkage [Internet]. London: Institute of Education [citado em 19 de agosto de 2015]. Disponível em: http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/problinkage.pdf

Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records. America Association for the Advancement of Science, 1959; 130:954-59.

Ortiz LP. O modelo de riscos competitivos no estudo da mortalidade infantil. Caxambu: XI Encontro de Estudos Populacionais, 1998.

Pepe MS, Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data. Statistics in Medicine, 1993; 12:737-51.

Pintilie M. Competing risks: a practical perspective. New York: John Wiley & Sons; 2006.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL hht://www.R-project.org/.

Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. British Jounal of Cancer. 2004; 91:1229-35.

Scrucca L, Santucci A, Aversa F. Regression modeling of competing risk using R: an in depth guide for clinicians. Bone Marrow Transplantation, 2010; 45:1388-95.

6. ARTIGO

Câncer entre 2002 e 2009 no Hospital de Clínicas de Porto Alegre: incidências, letalidades e fatores associados considerando eventos competitivos

Cancer between 2002 and 2009 at Hospital de Clínicas de Porto Alegre: incidence, lethality and factors associated considering competing events

Natalia Elis Giordani [1]; Isaías Prestes [1]; Luciana Neves Nunes [1, 2]; Suzi Alves Camey [1,2]

[1] Programa de Pós-Graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

[2] Instituto de Matemática e Estatística, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

A ser enviado à revista Cadernos de Saúde Pública

Resumo

O objetivo deste estudo foi avaliar incidências, letalidades e fatores associados ao óbito de pacientes com câncer primário atendidos no Hospital de Clínicas de Porto Alegre entre 2002 e 2009. Para tal, utilizou-se a abordagem de riscos competitivos. Os resultados mostram que os cânceres com maior incidência são: pele, próstata, brônquios e pulmões, mama, e sistema hematopoiético e reticuloendotelial, com número de casos variando entre 1.920 e 654. Já os com maiores taxas de letalidade, considerando o ano de diagnóstico, são: pâncreas, brônquios e pulmões, esôfago, estômago e fígado e vias biliares intra-hepáticas, com valores entre 0,74 e 0,27. Quanto a probabilidade marginal de óbito, em cinco anos, os maiores valores, de 0,57 a 0,52, referem-se aos cânceres de pâncreas, brônquios e pulmões e esôfago. Além disso, verificou-se que o risco de óbito por câncer de esôfago era maior em mulheres, enquanto que *idade* esteve associada com o risco de óbito por câncer de próstata. Com base nos achados, foi possível traçar um perfil dos casos de câncer atendidos pelo hospital.

Palavras-Chave: Neoplasias; Mortalidade; Análise de Sobrevivência

Abstract

The aim of this study was to assess incidence, lethality and factors associated with death in patients with primary cancer attended at Hospital de Clínicas de Porto Alegre from 2002 to 2009. For such we used the approach of competing risks. The results show that the cancers with higher incidence are: skin, prostate, lungs and bronchi, breast cancer, and hematopoietic and reticuloendothelial system, varying the number of cases between 1920 and 654. In relation to higher lethality rates, considering the year of diagnosis, we have: pancreas, bronchi and lungs, esophagus, stomach and liver and intrahepatic bile ducts, with rates range between 0.74 and 0.27. About the marginal probability of death in five years, the highest values, from 0.57 to 0.52, refer to pancreatic cancer, bronchi and lungs and esophagus. In addition, women has bigger risk of death for esophageal cancer, while *age* was associated with the risk of death for prostate cancer. Based on the findings it was possible to draw a profile of cancers attended by the hospital.

Keywords: Neoplasms; Mortality; Survival Analysis

Introdução

No ano de 2012, ocorreram 14,1 milhões novos casos de câncer, 8,2 milhões de mortes causadas por ele e 32,6 milhões de pessoas estavam vivendo com a doença (cinco anos após o diagnóstico) no mundo, de acordo com dados disponibilizados pelo GLOBOCAN [1]. Além disso, segundo o Instituto Nacional de Câncer (INCA), a quantidade de novos casos continuará aumentando nos países em desenvolvimento e crescerá ainda mais em países desenvolvidos se medidas preventivas não forem amplamente aplicadas [2].

Dessa forma, a análise de dados dos pacientes com câncer torna-se uma ferramenta necessária para monitorar e aprimorar as formas de prevenção, detecção e diagnóstico da doença, bem como avaliar os programas de tratamento, permitindo que os gestores da saúde consigam tomar decisões mais acertadas.

Conforme divulgado pelo INCA, as localizações primárias do tumor com maiores taxas de incidência, ajustadas por idade, por 100 mil, entre os anos de 2002 e 2004, na cidade de Porto Alegre são: traqueia, brônquios e pulmões, pele, cólon e reto, considerando ambos os sexos; câncer de próstata entre os homens; e câncer de mama entre as mulheres [3]. Já quando se trata de mortalidade sabe-se que, também nessa capital, entre os anos de 2004 e 2013, os cânceres com maiores taxas de mortalidade, segundo a localização primária do tumor, são: brônquios e pulmões; mama; cólon; pâncreas; estômago; fígado e vias biliares intra-hepáticas; e esôfago [4].

Apesar dos números apresentados se referirem à localização primária do tumor, deve-se lembrar de que metástases ou um novo câncer primário (denominado segundo tipo de câncer), mesmo que raro, podem ocorrer [5,6]. Assim, considerando o câncer de pulmão, os principais locais de metástases são: cérebro, fígado, glândula adrenal e osso. Já considerando o câncer de próstata, os locais mais comuns de metástase são: fígado, glândula adrenal, osso e pulmão. Por fim, para o câncer de mama, os locais de metástases mais frequentes são: cérebro, fígado, osso e pulmão [6].

Sendo assim, é importante que os estudos sobre câncer considerem não apenas a informação de localização primária do tumor, mas também a ocorrência de metástases ou um segundo câncer quando estimam a mortalidade, um dos parâmetros utilizados para avaliar os resultados da área oncológica. Isso é possível através de metodologias que consideram riscos competitivos, tais como a função de incidência de óbito acumulada e o modelo da subdistribuição do risco de óbito, que corrigem a superestimação ocasionada ao utilizar

abordagens que desprezam essa característica, tais como o método de Kaplan-Meier e o modelo de Cox [7].

Portanto, o objetivo principal desse trabalho foi avaliar incidências, letalidades e fatores associados ao óbito de pacientes com câncer primário atendidos no Hospital de Clínicas de Porto Alegre (HCPA) entre os anos de 2002 e 2009 usando a abordagem de riscos competitivos.

Métodos

Esse trabalho é baseado na coorte de casos primários de câncer atendidos no HCPA entre os anos de 2002 e 2009. Ao todo, são 11.832 pacientes acompanhados até 31/12/2013 a fim de garantir, no mínimo, cinco anos de acompanhamento para todos os indivíduos.

Na base de dados do HCPA, estão disponíveis as seguintes informações: sexo, data de nascimento, data da primeira consulta, data do diagnóstico e localização primária do tumor. No caso de óbito, constam: data do óbito, causa básica de morte (doença ou afecção que iniciou a cadeia de acontecimentos patológicos que conduziram diretamente à morte; ou as circunstâncias do acidente ou violência que produziu a lesão fatal [8]) e causa imediata de morte (doença, lesão ou complicação que ocorreu próximo ao momento da morte, geralmente desencadeada pela causa básica da morte [8]). Além dos óbitos registrados nessa base, foi utilizado o Sistema de Informações de Mortalidade (SIM) [9] para inclusão de mortes não ocorridas ou notificadas ao HCPA.

A identificação de óbitos registrados no SIM utilizou a técnica de relacionamento determinístico de registros (*deterministic record linkage*) [10] considerando os campos *nome* e *data de nascimento*. O *software* utilizado nessa etapa chama-se FRIL [11]. Através disso, 2.425 datas de óbito foram identificadas, totalizando 4.639 casos de morte.

As causas de morte e a localização primária do tumor estão codificadas de acordo com a Classificação Internacional de Doenças e Problemas Relacionados à Saúde (CID-10) [12], sendo que tanto para localização primária do tumor quanto para as causas de morte foram utilizados apenas os dois primeiros dígitos do código CID-10 com o intuito de agrupar topografias em locais iguais, independente da especificação do local.

A partir das informações disponíveis no banco de dados, algumas variáveis foram derivadas: a idade ao diagnóstico foi obtida a partir da diferença, em anos, entre a data de diagnóstico e a data de nascimento; o tempo de acompanhamento de casos de não óbito resulta da diferença, em anos, entre a data do término do seguimento e a data do diagnóstico;

e o tempo de acompanhamento dos casos de óbito foi obtido através da diferença, em anos, entre a data óbito e a data de diagnóstico. Para essas novas variáveis, são apresentados mediana e intervalo interquartílico (IIQ).

Dada a grande quantidade de neoplasias, optou-se por considerar para as análises os cânceres de pâncreas, brônquios e pulmões e esôfago, em função das altas taxas de letalidade, e os cânceres de mama (considerando apenas mulheres) e próstata, por possuírem altas incidências.

A fim de estimar as letalidades e avaliar os fatores associados ao óbito dos pacientes com câncer primário considerando a ocorrência de óbito por *metástase ou outros cânceres primários* ou por *causas não relacionadas ao câncer* foi utilizada a abordagem de riscos competitivos. Nela se considera que cada indivíduo pode sofrer apenas um de *D* tipos de eventos ao longo do período de seguimento, sendo que, quando um dos eventos ocorrer, não há possibilidade de observar a ocorrência dos demais [13, 14]. Essa abordagem é preferível em relação às tradicionalmente utilizadas (método de Kaplan-Meier e modelo de Cox), que resultam na superestimação da função de distribuição acumulada e em interpretações equivocadas [8]. Todas as análises foram realizadas utilizando o *software* R [15] e a biblioteca *cmprsk* [16]. Para os resultados do modelo, são exibidas as estimativas pontuais acompanhadas do intervalo de 95% de confiança.

O projeto de pesquisa foi aprovado pelo Comitê de Pesquisa e Ética em Saúde do HCPA (projeto 100056, data da versão: 29/06/2013).

Resultados

A coorte estudada é composta por 11.832 pacientes, sendo 6.515 homens (55,1%) e 5.317 (44,9%) mulheres. A idade mediana ao diagnóstico primário de câncer é de 61,2 anos (49,6; 70,9). Entre os homens, a idade mediana ao diagnóstico é de 62,3 anos (51,9; 71), já entre as mulheres é de 59,4 anos (46,9; 70,7). Em relação ao tempo de acompanhamento, o tempo de seguimento mediano foi de 5,6 anos (1,6; 8,5), sendo 5,3 anos (1,2; 8,2) entre os homens e 6,1 anos (2,5; 8,9) entre as mulheres. Cabe ressaltar que nos casos em que o tempo mediano de seguimento é inferior a cinco anos, essa informação corresponde ao tempo mediano de sobrevivência após o diagnóstico. Na Tabela 1, são apresentadas informações detalhadas para os cinco tipos de câncer avaliados nesse estudo.

Quanto aos tipos de neoplasias com as 10 maiores incidências entre os anos de 2002 e 2009 no HCPA, têm-se: pele (1.920 casos), próstata (1.080 casos), brônquios e pulmões (950

casos), mama (893 casos), sistema hematopoiético e reticuloendotelial (654 casos), cólon (573 casos), esôfago (497 casos), estômago (422 casos), neoplasia maligna secundária e não especificada dos gânglios linfáticos (360 casos) e colo do útero (328 casos).

Tabela 1 - Descrição da população considerando os tipos de cânceres analisados

Tipo de câncer	**Quantidade de pacientes n (%)**			**Idade mediana ao diagnóstico (IIQ)**			**Tempo mediano de acompanhamento (IIQ)**		
	Total	**Homens**	**Mulheres**	**Total**	**Homens**	**Mulheres**	**Total**	**Homens**	**Mulheres**
Pâncreas	254	124 (48,8)	130 (51,2)	65,7 (54,6; 73,9)	63,7 (54,6; 71,2)	67,6 (54,4; 77,3)	0,7 (0,2; 7,1)	0,6 (0,1; 6,6)	1,2 (0,2; 7,5)
Brônquios e pulmões	950	609 (64,1)	341 (35,9)	64,9 (57,1; 72,2)	66,0 (58,9; 72,4)	62,8 (54,1; 71,5)	1,1 (0,2; 6,6)	0,9 (0,2; 6,3)	1,6 (0,3; 7,0)
Esôfago	497	375 (75,5)	122 (24,6)	60,7 (53,6; 68,6)	59,9 (53,3; 67,4)	64,0 (56,0; 73,4)	1,4 (0,5; 6,4)	1,4 (0,4; 6,1)	2,0 (0,6; 7,7)
Mama	-	-	893	-	-	57,0 (47,2; 67,2)	-	-	6,9 (4,9; 9,4)
Próstata	-	1080	-	-	67,6 (61,8; 72,6)	-	-	6,9 (4,7; 8,9)	-

Já em relação à mortalidade, dos 4.639 óbitos observados entre 2002 e 2009: 2.314 (49,9%) têm como causa básica de morte o mesmo tipo de câncer da localização primária do tumor; 1.462 (31,5%) têm causa básica de morte um tipo de câncer diferente da localização primária do tumor, esses são casos que podem ser de metástase ou segundo tumor primário; e 863 (18,6%) têm causa básica de morte não relacionada ao câncer.

Na Tabela 2, são apresentadas duas informações: quantidade de casos, por ano, considerando os cinco tipos de câncer com maior incidência; e a taxa de letalidade em cinco anos, por ano de diagnóstico, para os tipos de cânceres que apresentaram os maiores valores.

Tabela 2 - Incidência e letalidade considerando o câncer primário e ano de diagnóstico

	Tipo de câncer	Data do diagnóstico							
		2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Incidência	Pele	104	158	194	218	232	310	395	309
	Próstata	90	116	144	160	143	153	120	154
	Brônquios e pulmões	93	135	104	102	100	117	157	142
	Mama	121	102	116	103	91	114	146	100
	Sistema hematopoiético e reticuloendotelial	68	104	102	93	67	57	76	87
Letalidade pelo câncer primário em 5 anos	Pâncreas	0,45	0,46	0,64	0,45	0,52	0,57	0,74	0,67
	Brônquios e pulmões	0,49	0,46	0,53	0,51	0,53	0,60	0,55	0,64
	Esôfago	0,53	0,41	0,51	0,52	0,59	0,52	0,51	0,49
	Estômago	0,28	0,53	0,49	0,37	0,41	0,51	0,49	0,39
	Fígado e vias biliares intra-hepáticas	0,27	0,46	0,57	0,38	0,41	0,47	0,41	0,39

No que diz respeito à letalidade, dentre as neoplasias com maiores taxas têm-se: pâncreas (145 óbitos; 57,1%), brônquios e pulmões (527 óbitos; 55,5%), esôfago (262 óbitos, 52,7%), estômago (186 óbitos; 44,1%), fígado e vias biliares (122 óbitos; 42,5%), vesícula biliar (23 óbitos; 42,6%), intestino delgado (12 óbitos; 40%), ovário (42 óbitos; 34,1%), cérebro (40 óbitos; 31,3%), e neoplasia maligna de outras localizações e de partes não especificadas das vias biliares (23 óbitos; 28%).

Utilizando a abordagem de riscos competitivos, a fim de evitar a superestimação ocasionada pelas metodologias tradicionais, calculou-se a função de incidência acumulada para cada um dos cinco cânceres analisados. Os resultados obtidos estão apresentados nas Figuras 1 e 2.

Através da função de incidência acumulada é possível calcular tanto a probabilidade marginal de óbito pelo câncer de interesse quanto por outros tipos de câncer. Os resultados mostram que, em cinco anos: para câncer de pâncreas, a probabilidade marginal de óbito por esse câncer é de 0,57, já por outros tipos é de 0,04; para câncer de brônquios e pulmões, a probabilidade marginal de óbito por esse mesmo câncer é de 0,54, por outras neoplasias é de 0,05; considerando o câncer de esôfago, a probabilidade marginal de óbito por essa doença é de 0,52, já para outros tipos de câncer é de 0,05; para câncer de mama, a probabilidade marginal de óbito pelo próprio câncer é de 0,12, considerando óbito por outros tipos, é de 0,02; por fim, para câncer de próstata, a probabilidade marginal de óbito por esse câncer é de 0,09 e, por outros tipos, é 0,02.

Ainda sobre os cânceres de mama e próstata, observa-se que a probabilidade marginal de óbito por outras causas é superior a de óbito por outros tipos de câncer, sendo que para o

câncer de próstata, a partir de aproximadamente 10 anos após o diagnóstico, a probabilidade marginal de óbito por outras causas é superior a de óbito por câncer de próstata (Figura 2).

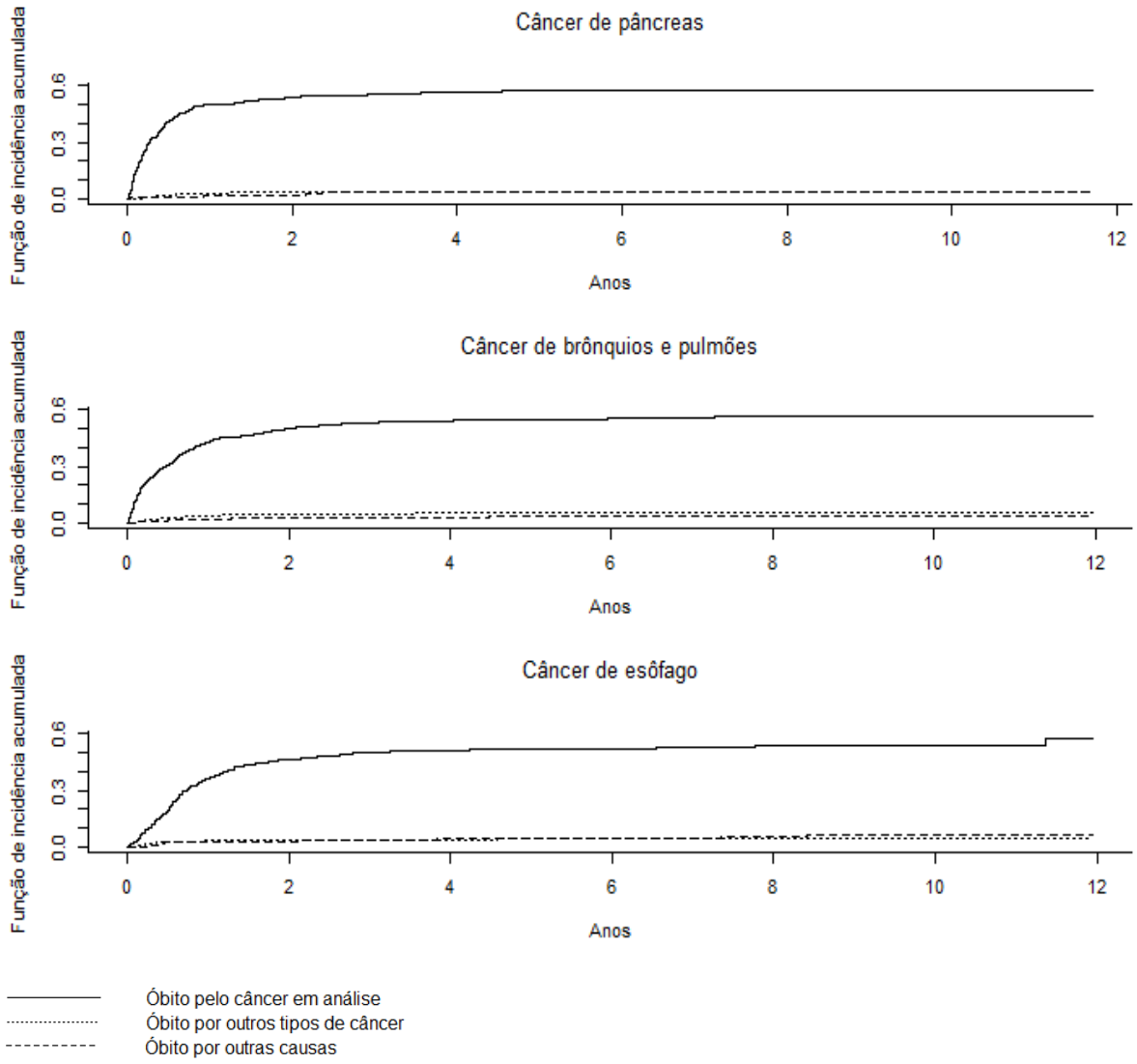


Figura 1 - Probabilidade marginal para óbito por cada um dos três tipos de câncer mais letais, considerando os eventos competitivos *óbito por outros tipos de câncer* ou *causa não relacionada ao câncer*

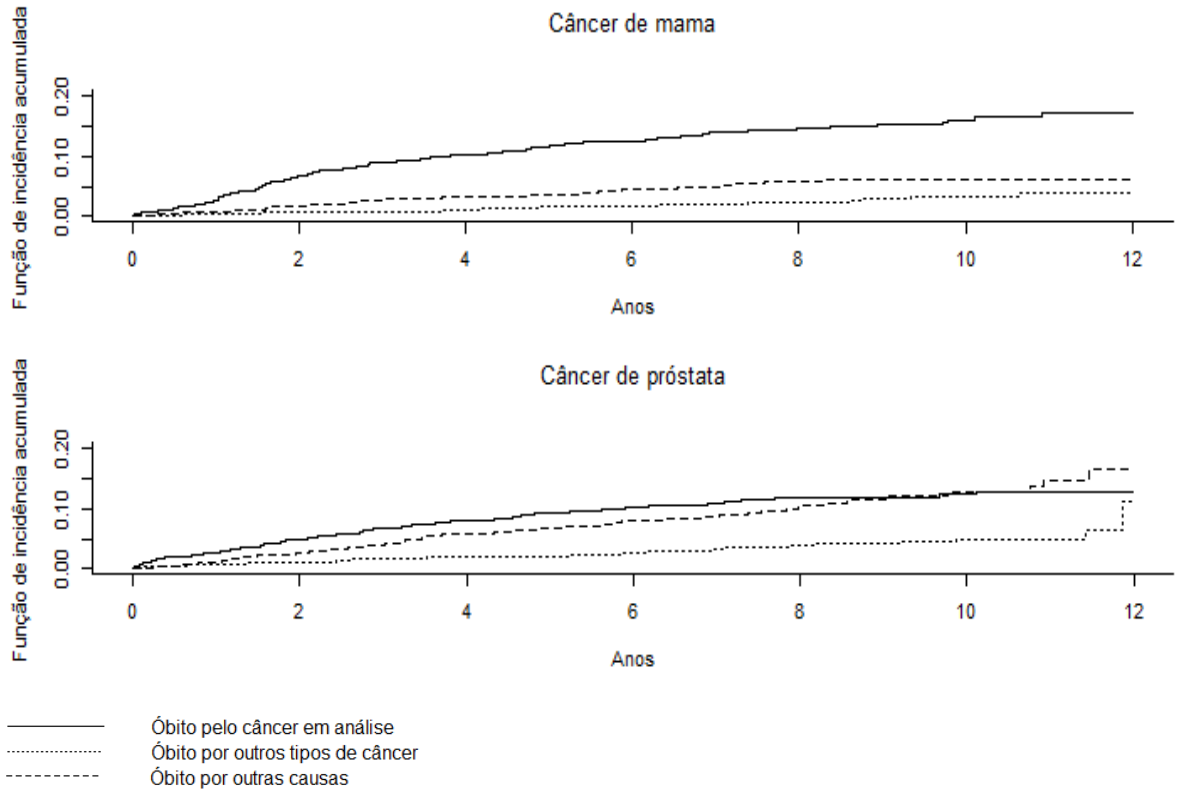


Figura 2 - Probabilidade marginal para óbito por cada um dos dois tipos de câncer mais incidentes, considerando os eventos competitivos *óbito por outros tipos de câncer* ou *causa não relacionada ao câncer*

Por fim, utilizando a abordagem da subdistribuição do risco, buscou-se avaliar o efeito de *sexo* e *idade* (covariáveis) para o risco de óbito por cada tipo de câncer analisado (o que se denominou evento de interesse), considerando os eventos competitivos. Os resultados obtidos mostram que a covariável *sexo* foi significativa apenas para o câncer de esôfago (mulheres com maior risco de óbito), enquanto que *idade* foi significativa para o câncer de próstata (Tabela 3).

Tabela 3 - Razão de azares para o evento de interesse (RA) ajustado por *sexo* e *idade*

Evento de interesse	Covariável	RA	IC95%
Pâncreas	Sexo = Masculino	0,786	0,57; 1,09
	Idade	1,006	0,99; 1,02
Brônquios e pulmões	Sexo = Masculino	0,889	0,74; 1,06
	Idade	1,005	0,99; 1,01
Esôfago	Sexo = Masculino	0,724	0,54; 0,98
	Idade	1,006	0,99; 1,02
Mama	Idade	0,995	0,98; 1,01
Próstata	Idade	1,080	1,05; 1,10

Discussão

Através desse estudo buscou-se, primeiramente, descrever os casos de câncer atendidos no HCPA entre os anos de 2002 e 2009 de acordo com a incidência, letalidade e influência de sexo e idade na letalidade. Os resultados mostraram que, considerando todos os tipos de neoplasias, o número de novos casos tem aumentado ao longo dos anos, bem como a quantidade de óbitos pela doença. A maior parte dos casos refere-se a homens cuja idade mediana ao diagnóstico é superior ao do sexo feminino e com tempo mediano de seguimento inferior. Isso pode ser consequência de diagnóstico tardio por parte dos homens ou pelo simples fato de que eles são mais velhos que as mulheres quando diagnosticados.

Quanto aos cânceres primários mais incidentes no HCPA, vemos uma diferença em relação aos dados de 2002 a 2004 publicados pelo INCA [3] para a cidade de Porto Alegre. Os cânceres de traqueia e reto, por exemplo, apontados como alguns dos mais incidentes, não constam entre os dez tipos mais frequentes do hospital. Para as taxas de mortalidade, esse cenário já é diferente: os tipos apontados pelo INCA [4] vão ao encontro do resultado obtido no HCPA.

Os cânceres de mama feminino e próstata destacaram-se pela grande quantidade de casos na população estudada. Quanto ao primeiro, Porto Alegre é a capital com maiores valores de taxas médias de incidências anuais de câncer de mama, ajustadas por idade [3]. Já o câncer de próstata é o tipo mais frequente em todas as regiões do país, e, ao longo dos anos, espera-se uma tendência de aumento em sua incidência. Isso está relacionado à evolução dos métodos diagnósticos e aumento da expectativa de vida. Além disso, sabe-se que a idade é um fator de risco estabelecido para câncer de próstata, sendo que, após os 50 anos, as taxas de incidência aumentam progressivamente [3].

Na avaliação mais detalhada de cada um dos cinco cânceres estudados, chamou atenção o fato de que o tempo mediano de sobrevivência para o sexo masculino é inferior a um ano tanto para o câncer de pâncreas (0,6 ano), quanto para o câncer de brônquios e pulmões (0,9 ano). Para o sexo feminino, o tempo mediano de sobrevivência foi de 1,2 anos para o câncer de pâncreas e 1,6 anos para o câncer de brônquios e pulmões. De fato, os dados divulgados pelo INCA apontam para altas taxas de mortalidade pelas doenças dado seu comportamento agressivo e diagnóstico tardio [2].

Ainda em relação aos cinco cânceres avaliados, uma análise da letalidade em cinco anos, considerando o ano de diagnóstico, apontou para valores altos e, de forma geral, crescentes. Considerando os casos de câncer de pâncreas diagnosticados em 2008, por exemplo, 74% desses pacientes faleceram por esse mesmo câncer em até cinco anos após o diagnóstico.

Os resultados obtidos através da função de incidência acumulada também mostram, pelo comportamento acentuado da curva, que a probabilidade de óbito pelos cânceres de pâncreas, brônquios e pulmões e esôfago é bastante alta, em especial nos primeiros anos após o diagnóstico. Uma comparação entre as estimativas de óbito em até cinco anos para câncer de brônquios e pulmões obtida no estudo (0,54) e divulgada pelo INCA (0,90; 0,93) [17] pode revelar os indícios de superestimação referidos na escolha do método utilizado. O mesmo vale para câncer de esôfago, onde o INCA estima que a probabilidade de óbito em até cinco anos é de 0,90 [2] e nesse estudo obtivemos 0,57.

Por fim, buscou-se avaliar o efeito das covariáveis disponíveis no risco de óbito para os cânceres estudados, também considerando a abordagem de riscos competitivos. Os resultados indicam que indivíduos do sexo masculino tem risco 0,72 (0,54; 0,98) vezes menor de falecer por câncer de esôfago do que as mulheres. Além disso, a idade aparece como fator de risco para o câncer próstata, sendo que o aumento de um ano aumenta o risco de óbito por câncer de próstata em 1,08 (1,05; 1,10) vezes.

Através desse trabalho foi possível caracterizar os tipos de cânceres com maiores incidências e letalidades (considerando os eventos competitivos) no HCPA, avaliar como *sexo* e *idade* podem contribuir para o risco de óbito de alguns tipos de câncer e comparar os resultados do hospital com os do município de Porto Alegre, onde algumas características peculiares foram verificadas.

Colaboradores

N. E. Giordani participou da concepção do estudo, da revisão da literatura, construção da base de dados, análise dos resultados e redação do texto. L. N. Nunes e S. A. Camey participaram da concepção do estudo, análise e discussão dos resultados e na revisão do texto. I. Prestes contribuiu com o relacionamento entre as bases de dados do HCPA e SIM.

Agradecimentos

À Universidade Federal do Rio Grande do Sul, pela realização do Mestrado Acadêmico em Epidemiologia e ao Departamento de Registro Hospitalar de Câncer do Hospital de Clínicas de Porto Alegre, pelos dados cedidos e dúvidas esclarecidas.

Referências

1. Ferlay J,et al. GLOBOCAN 2012 v1.0. Estimated cancer incidence, mortality and prevalence worldwide in 2012 [Internet]. Lyon: International Agency for Research on Cancer [citado em 22 de agosto de 2015]. Disponível em: http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx
2. Instituto Nacional de Câncer José Alencar Gomes da Silva. Coordenação de Prevenção e Vigilância. Estimativa 2014: Incidência de Câncer no Brasil [Internet]. Rio de Janeiro: INCA, 2014 [citado em 22 de agosto de 2015]. Disponível em: http://www.inca.gov.br/estimativa/2014/estimativa-24042014.pdf
3. Instituto Nacional de Câncer (Brasil). Coordenação de Prevenção e Vigilância. Câncer no Brasil: dados dos registros de base populacional, v. 4 [Internet]. Rio de Janeiro: INCA, 2010 [citado em 09 de setembro de 2015]. Disponível em: http://www.inca.gov.br/cancernobrasil/2010/docs/registro%20de%20base%20populacional_completo.pdf
4. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estatísticas do Câncer – Vigilância do Câncer e de Fatores de Risco. Atlas On-line de Mortalidade por Câncer [Internet]. Distribuição proporcional do total de mortes por câncer, segundo localização primária do tumor, por sexo, localidade, por período selecionado [citado em 07 de setembro de 2015]. Disponível em: https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo02/consultar.xhtml#panelResultado

5. American Cancer Society. Learn about cancer [Internet]. What are second cancer? [citado em 07 de setembro de 2015]. American Cancer Society, 2014. Disponível em: http://www.oncoguia.org.br/conteudo/segundo-cancer/7798/1/
6. Instituto Oncoguia. Câncer Avançado [Internet]. Entenda o que é o Câncer Metastático [citado em 07 de setembro de 2015]. Instituo Oncoguia, 2013. Disponível em: http://www.oncoguia.org.br/conteudo/entenda-o-que-e-cancer-metastatico/3186/357/
7. Giordani NE. Aplicação do modelo de riscos competitivos em pacientes diagnosticados com câncer no ano de 2006 no Hospital de Clínicas de Porto Alegre [monografia]. Departamento de Estatística: Universidade Federal do Rio Grande do Sul; 2013.
8. Tutorial para Preencher as Causas da Morte da Declaração de Óbito: Tutorial - Seção I - I. Conceitos e Definições [Internet]. UNIFESP - Departamento de Informática em Saúde [citado em 07 de setembro de 2015]. Disponível em: http://atestadodeobito.unifesp.br/tela.php?numero=3
9. Banco de dados do Sistema Único de Saúde - DATASUS. Brasil, Ministério da Saúde, Secretaria de vigilância em Saúde.
10. Fellegi IP, Sunter A B. A theory for record linkage. Journal of the American Statistical Association, 1969; 64:1183-1210.
11. Jurczyk, P. FRIL: Fine-grained Record Integration and Linkage. Emory University, Math&CS Department, 2009.
12. ORGANIZAÇÃO MUNDIAL DA SAÚDE. CID 10: classificação estatística internacional de doenças e problemas relacionados à saúde. 9. ed. São Paulo: ed. da EDUSP, 2003. 1 CD-ROM
13. Balakrishnan N, Rao CR. Handbook of Statistics: Advances in Survival Analysis. 2004; 23.
14. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text. 2nd ed. New York: Springer Science; 2005.
15. R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL hht://www.R-project.org/.
16. Bob Gray (2013). *cmprsk: Subdistribution Analysis of Competing Risks*. R package version 2.2-6.

17. Instituto Nacional de Câncer José Alencar Gomes da Silva. Tipos de câncer: Pulmão [Internet]. Rio de Janeiro: INCA, 2014 [citado em 17 de setembro de 2015]. Disponível em: http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao

7. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Nessa dissertação, foi abordada análise de sobrevivência considerando eventos competitivos, destacando sua vantagem quando comparada as abordagens tradicionais.

Conceitos e funções básicas da metodologia foram apresentados. Os tradicionais método de Kaplan-Meier e modelo de Cox, utilizados nos casos de eventos únicos, foram revisados e exemplos de sua utilização, aplicados na área da saúde, foram apresentados juntamente com códigos do *software* R. Eventos competitivos foram apresentados, bem como a abordagem ideal para tratá-los. Os conceitos de função de incidência acumulada, curvas de probabilidade condicional e modelo da subdistribuição do risco foram discutidos. Exemplos para cada um também foram apresentados, bem como os códigos utilizados.

Foram mencionadas as razões pelas quais, em situações onde eventos competitivos estão presentes, a abordagem tradicional deve ser evitada: a função de distribuição acumulada torna-se superestimada e a relação entre as funções básicas é perdida. Também foram comentadas as vantagens e desvantagens de utilizar a abordagem de riscos competitivos: não é necessária a suposição de que os eventos competitivos sejam independentes, porém necessita das suposições de riscos proporcionais e de que um paciente que faleceu por outra causa seja tratado como ainda pertencente ao grupo em risco para a causa de interesse mesmo que, na prática, isso possa ser biologicamente impossível.

Técnicas utilizadas para relacionar bases de dados também foram discutidas, *deterministic* e *probabilistic record linkage*, e exemplo de sua utilização foi apresentado. Esse assunto tornou-se parte da revisão uma vez que o objetivo desse trabalho consistiu em aplicar a metodologia de riscos competitivos na base de dados dos pacientes com câncer atendidos pelo HCPA e ocorreu que, em diversas situações, não havia o registro de óbito desses pacientes na base de dados do hospital. Assim, o relacionamento entre os dados do HCPA e SIM permitiu que fossem localizados mais de 2 mil casos de morte não presentes na base do hospital.

A partir de então, utilizando essa nova base de dados (resultado do *deterministic linkage*), foi aplicada a metodologia de riscos competitivos para avaliar incidências, letalidades e fatores associados ao óbito de pacientes com câncer primário atendidos no HCPA entre 2002 e 2009. Os resultados obtidos permitiram um melhor conhecimento dos tipos de cânceres com maiores incidências e letalidades (considerando os eventos competitivos) e também uma avaliação da associação das variáveis *sexo* e *idade* com risco de

óbito de alguns tipos de câncer. Além disso, embora tenha sido discutido o método de *probabilistic record linkage*, os resultados apresentados, até o momento, utilizaram apenas o relacionamento determinístico de registros.

Apesar da ficha de registro de tumor (Anexo b), instrumento utilizado para registrar as informações dos pacientes com câncer, conter vários detalhes, apenas as informações de *sexo* e *data de nascimento* são inseridas no sistema de dados utilizado pelo hospital para envio ao INCA. Em função disso, o modelo que utilizamos contém apenas as duas únicas covariáveis disponíveis.

Por fim, recomenda-se aos pesquisadores que não desprezem, em seus estudos, situações com eventos competitivos. Já existem diversos livros e artigos que tratam desse tema e utilizam programas como o R, que já possui pacotes implementados para essa abordagem, em seus exemplos.

8. ANEXOS

a. Aprovação pelo Comitê de Pesquisa e Ética em Saúde

HCPA - HOSPITAL DE CLÍNICAS DE PORTO ALEGRE
Grupo de Pesquisa e Pós-Graduação
COMISSÃO CIENTÍFICA E COMISSÃO DE PESQUISA E ÉTICA EM SAÚDE

A Comissão Científica e a Comissão de Pesquisa e Ética em Saúde, que é reconhecida pela Comissão Nacional de Ética em Pesquisa (CONEP)/MS como Comitê de Ética em Pesquisa do HCPA e pelo Office For Human Research Protections (OHRP)/USDHHS, como Institucional Review Board (IRB00000921) analisaram o projeto:

Projeto: 100056

Pesquisador Responsável
BRUCE BARTHOLOW DUNCAN

Título: Projeto de Desenvolvimento para a Consolidação do Centro Colaborador para a Vigilância do Diabetes, Doenças Cardiovasculares e Outras Doenças Não Transmissíveis - Análise de Dados Primários e Secundários dos Grandes Sistemas Nacionais de Informações em Saúde do Sistema Único de Saúde

	Data da Versão:
ADENDO AO PROJETO	29/08/2013

Este documento referente ao projeto acima foi APROVADO em seus aspectos éticos e metodológicos, de acordo com as Diretrizes e Normas Internacionais e Nacionais, especialmente as Resoluções 196/96 e complementares do Conselho Nacional de Saúde.

Porto Alegre, 09 de setembro de 2013.

Prof. José Roberto Goldim
Coordenação CEP/HCPA

Data da Impressão: 09/09/2013 **Página 1 de 1**

HCPA - HOSPITAL DE CLÍNICAS DE PORTO ALEGRE
GRUPO DE PESQUISA E PÓS-GRADUAÇÃO

COMISSÃO CIENTÍFICA E COMISSÃO DE PESQUISA E ÉTICA EM SAÚDE

A Comissão Científica e a Comissão de Pesquisa e Ética em Saúde, que é reconhecida pela Comissão Nacional de Ética em Pesquisa (CONEP)/MS como Comitê de Ética em Pesquisa do HCPA e pelo Office For Human Research Protections (OHRP)/USDHHS, como Institutional Review Board (IRB00000921) analisaram o projeto:

Projeto: 100056 **Versão do Projeto:** 19/02/2010

Pesquisadores:
BRUCE BARTHOLOW DUNCAN

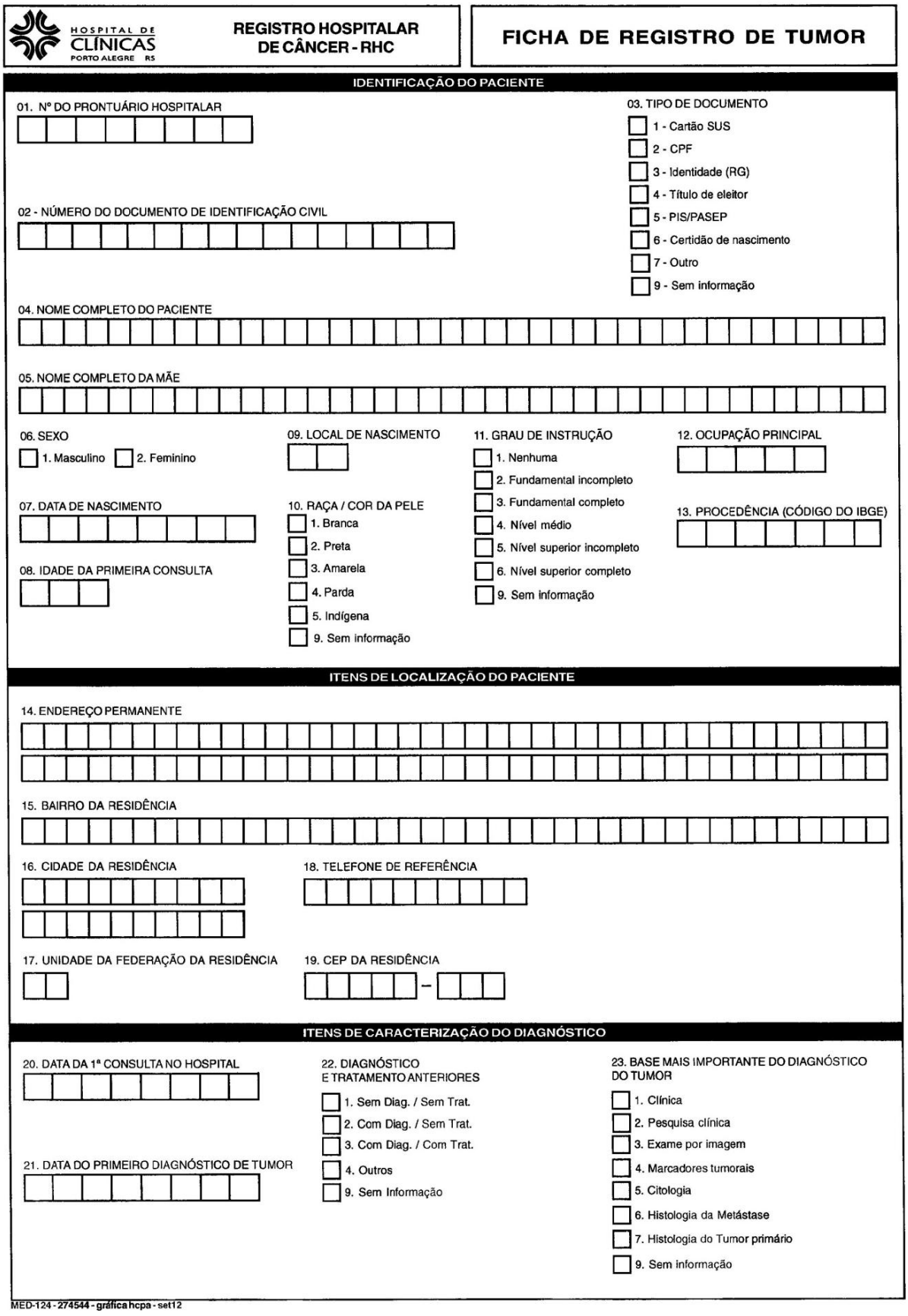
Título: Projeto de Desenvolvimento para a Consolidação do Centro Colaborador para a Vigilância do Diabetes, Doenças Cardiovasculares e Outras Doenças Não Transmissíveis - Análise de Dados Primários e Secundários dos Grandes Sistemas Nacionais de Informações em Saúde do Sistema Único de Saúde

Este projeto foi Aprovado em seus aspectos éticos e metodológicos de acordo com as Diretrizes e Normas Internacionais e Nacionais, especialmente as Resoluções 196/96 e complementares do Conselho Nacional de Saúde. Os membros do CEP/HCPA não participaram do processo de avaliação dos projetos onde constam como pesquisadores. Toda e qualquer alteração do Projeto deverá ser comunicada imediatamente ao CEP/HCPA.

Porto Alegre, 10 de março de 2010.

Profª Nadine Clausell
Coordenadora GPPG e CEP/HCPA

b. Ficha de registro de tumor usada pelo HCPA para criação do banco de dados enviado ao INCA



HOSPITAL DE CLÍNICAS
PORTO ALEGRE RS

REGISTRO HOSPITALAR DE CÂNCER - RHC

FICHA DE REGISTRO DE TUMOR

IDENTIFICAÇÃO DO PACIENTE

01. Nº DO PRONTUÁRIO HOSPITALAR

02 - NÚMERO DO DOCUMENTO DE IDENTIFICAÇÃO CIVIL

03. TIPO DE DOCUMENTO
- 1 - Cartão SUS
- 2 - CPF
- 3 - Identidade (RG)
- 4 - Título de eleitor
- 5 - PIS/PASEP
- 6 - Certidão de nascimento
- 7 - Outro
- 9 - Sem informação

04. NOME COMPLETO DO PACIENTE

05. NOME COMPLETO DA MÃE

06. SEXO
1. Masculino 2. Feminino

07. DATA DE NASCIMENTO

08. IDADE DA PRIMEIRA CONSULTA

09. LOCAL DE NASCIMENTO

10. RAÇA / COR DA PELE
- 1. Branca
- 2. Preta
- 3. Amarela
- 4. Parda
- 5. Indígena
- 9. Sem informação

11. GRAU DE INSTRUÇÃO
- 1. Nenhuma
- 2. Fundamental incompleto
- 3. Fundamental completo
- 4. Nível médio
- 5. Nível superior incompleto
- 6. Nível superior completo
- 9. Sem informação

12. OCUPAÇÃO PRINCIPAL

13. PROCEDÊNCIA (CÓDIGO DO IBGE)

ITENS DE LOCALIZAÇÃO DO PACIENTE

14. ENDEREÇO PERMANENTE

15. BAIRRO DA RESIDÊNCIA

16. CIDADE DA RESIDÊNCIA

17. UNIDADE DA FEDERAÇÃO DA RESIDÊNCIA

18. TELEFONE DE REFERÊNCIA

19. CEP DA RESIDÊNCIA –

ITENS DE CARACTERIZAÇÃO DO DIAGNÓSTICO

20. DATA DA 1ª CONSULTA NO HOSPITAL

21. DATA DO PRIMEIRO DIAGNÓSTICO DE TUMOR

22. DIAGNÓSTICO E TRATAMENTO ANTERIORES
- 1. Sem Diag. / Sem Trat.
- 2. Com Diag. / Sem Trat.
- 3. Com Diag. / Com Trat.
- 4. Outros
- 9. Sem Informação

23. BASE MAIS IMPORTANTE DO DIAGNÓSTICO DO TUMOR
- 1. Clínica
- 2. Pesquisa clínica
- 3. Exame por imagem
- 4. Marcadores tumorais
- 5. Citologia
- 6. Histologia da Metástase
- 7. Histologia do Tumor primário
- 9. Sem informação

MED-124 - 274544 - gráfica hcpa - set12

ITENS DE CARACTERIZAÇÃO DO TUMOR

24. LOCALIZAÇÃO DO TUMOR PRIMÁRIO

25. TIPO HISTOLÓGICO DO TUMOR PRIMÁRIO

26. TNM

27a. ESTADIAMENTO CLÍNICO DO TUMOR (TNM)

27b. OUTRO ESTADIAMENTO (DIFERENTE DO TNM E IDADE ATÉ 18 ANOS)

28. pTNM

29. LOCALIZAÇÃO DE METÁSTASE À DISTÂNCIA

ITENS DE CARACTERIZAÇÃO DO PRIMEIRO TRATAMENTO

30. CLÍNICA DO INÍCIO DE TRATAMENTO NO HOSPITAL

31. DATA DO INÍCIO DO 1º TRATAMENTO ESPECÍFICO PARA O TUMOR, NO HOSPITAL

32. PRINCIPAL RAZÃO PARA A NÃO REALIZAÇÃO DO TRATAMENTOANTINEOPLÁSICO NO HOSPITAL

- 1. Recusa do Tratamento
- 2. Tratamento realizado fora
- 3. Doença avançada, falta de condições clínicas ou outras doenças associadas
- 4. Abandono de Tratamento
- 5. Complicações do Tratamento
- 6. Óbito
- 7. Outras razões
- 8. Não se aplica
- 9. Sem informação

33. PRIMEIRO TRATAMENTO RECEBIDO NO HOSPITAL

- 1. Nenhum
- 2. Cirurgia
- 3. Radioterapia
- 4. Quimioterapia
- 5. Hormonioterapia
- 6. Transplante de Medula Óssea
- 7. Imunoterapia
- 8. Outras
- 9. Sem Informação

34. ESTADO DA DOENÇA AO FINAL DO PRIMEIRO TRATAMENTO NO HOSPITAL

- 1. Sem Evidência da Doença (Remissão Completa)
- 2. Remissão Parcial
- 3. Doença Estável
- 4. Doença em Progressão
- 5. Suporte terapêutico oncológico
- 6. Óbito
- 8. Não se aplica
- 9. Sem Informação

35. DATA DO ÓBITO

36. ÓBITO POR CÂNCER

- 1. Sim
- 2. Não
- 9. Ignorado

ITENS DE CARACTERIZAÇÃO DO PRIMEIRO TRATAMENTO

37. CASO ANALÍTICO

- 1. Sim
- 2. Não

38. INDICAÇÃO DE REALIZAÇÃO DE SEGUIMENTO

- 1. Sim
- 2. Não

ITENS DE IDENTIFICAÇÃO DO REGISTRADOR

39. CÓDIGO DE IDENTIFICAÇÃO DO REGISTRADOR

ITENS OPCIONAIS

40. ESTADO CONJUGAL ATUAL

- 1. Casado
- 2. Solteiro
- 3. Desquitado / divorciado
- 4. Viúvo
- 9. Sem informação

41. DATA DA TRIAGEM

42. HISTÓRICO FAMILIAR DE CÂNCER

- 1. Sim
- 2. Não
- 9. Sem informação

43. HISTÓRICO DE CONSUMO DE BEBIDA ALCÓOLICA

- 1. Nunca
- 2. Ex-consumidor
- 3. Sim
- 4. Não avaliado
- 8. Não se aplica
- 9. Sem informação

44. HISTÓRICO DE CONSUMO DE TABACO

- 1. Nunca
- 2. Ex-consumidor
- 3 - Sim
- 4 - Não avaliado
- 8 - Não se aplica
- 9 - Sem informação

45. ORIGEM DE ENCAMINHAMENTO

- 1. SUS
- 2. Não SUS
- 3 - Veio por conta própria
- 8 - Não se aplica
- 9 - Sem informação

46. CLÍNICA DE ENTRADA DO PACIENTE NO HOSPITAL

42. EXAMES RELEVANTES PARA O DIAGNÓSTICO E PLANEJAMENTO DA TERAPÊUTICA DO TUMOR

- 1. Exame Clínico e Patologia Clínica
- 2. Exames por Imagem
- 3. Endoscopia e Cirurgia Exploradora
- 4. Anatomia Patológica
- 8. Não se aplica
- 9. Sem Informação

43. LOCALIZAÇÃO PRIMÁRIA PROVÁVEL

50. LATERALIDADE DO TUMOR

- 1. Direita
- 2. Esquerda
- 3. Bilateral
- 8. Não se aplica
- 9. Sem Informação

51. OCORRÊNCIA DE MAIS DE UM TUMOR PRIMÁRIO

- 1. Não
- 2. Sim
- 3. Duvidoso

52. CUSTEIO DO DIAGNÓSTICO DO TUMOR NO HOSPITAL

- 1. Público (SUS)
- 2. Plano de saúde
- 3. Particular
- 4. Outros
- 8. Não se aplica
- 9. Sem informação

53. CUSTEIO DO TRATAMENTO DO TUMOR NO HOSPITAL

- 1. Público (SUS)
- 2. Plano de saúde
- 3. Particular
- 4. Outros
- 8. Não se aplica
- 9. Sem informação

32. CAUSA BÁSICA DA MORTE