Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Ana Flávia do Carmo Santos
Felipe Werner

Análise Estatística de Sobrevivência: um estudo de pacientes com câncer de mama tratados no município de Curitiba

Curitiba
2016

Ana Flávia do Carmo Santos
Felipe Werner

Análise Estatística de Sobrevivência: um estudo de pacientes com câncer de mama tratados no município de Curitiba

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

Orientadora: Profa. Dra. Suely Ruiz Giolo

Curitiba
2016

Agradecimentos

Agradeço primeiramente à Deus, por ter me dado saúde e força para superar as dificuldades.

À minha mãe Lucineia, pelo apoio incondicional, incentivo nas horas difíceis, de desânimo e cansaço e por tudo que sempre fez por mim. Sempre me dando conselhos, força e coragem. Você é demais!

Ao meu namorado, Bruno, pelo companheirismo e paciência.

À orientadora, Profa. Dra. Suely Ruiz Giolo, pela paciência e conhecimento transmitido. E ao Prof. Dr. Cesar Augusto Taconeli pela disponibilidade em participar da banca deste trabalho.

Aos amigos que fiz nesses quase cinco anos de faculdade, em especial à Damiane e Cintia. Obrigada por todos os momentos em que fomos estudiosas, conselheiras e cúmplices. Esta caminhada não seria a mesma sem vocês.

Ao parceiro de projeto Felipe, pelo companheirismo.

À todos que, mesmo não estando citados aqui, tanto contribuíram para a conclusão desta etapa, o meu muito obrigada!

Ana Flávia do Carmo Santos

Agradeço aos meus pais pelo esforço feito para que eu pudesse ter a oportunidade de concluir essa etapa.

À minha noiva Beatriz, pela compreensão e companheirismo nos momentos de ausência.

Um obrigado especial à orientadora Suely Ruiz Giolo, por toda a dedicação e paciência para que pudéssemos concluir esse trabalho, também ao professor Cesar Augusto Taconeli por dedicar seu tempo para nos avaliar.

As amizades feitas durante o curso, algumas que levarei para toda a vida, em especial Bruno e Jhosefer, por todo tempo dedicado aos estudos, sem o apoio de vocês talvez eu nem tivesse chego ao fim.

À companheira de projeto Ana Flávia, pelo companheirismo e comprometimento em todas as etapas

Felipe Werner

"Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender a conhecer a influência libertadora da beleza do reino do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer".

(Albert Einstein)

Resumo

O câncer de mama foi, em 2012, o segundo tipo de câncer mais incidente no mundo. Segundo o Instituto Nacional de Câncer - INCA, são esperados 57.960 casos novos de câncer de mama para o Brasil em 2016. Em virtude disso, identificar fatores associados à sobrevida de pacientes com esse tipo de câncer se torna de suma importância. Com esse objetivo, este trabalho apresenta um estudo sobre pacientes com câncer de mama tratados em um centro médico de Curitiba. Os dados analisados foram coletados em um período de 20 anos de série histórica, compreendendo os anos de 1990 a 2009. Para a seleção das covariáveis foi levado em consideração as correlações entre as mesmas. Para as análises descritivas no contexto de análise de sobrevivência foi utilizado o estimador de Kaplan-Meier. Foram ajustados os modelos de riscos proporcionais de Cox, aditivo de Aalen e um submodelo deste último, o de riscos aditivos semiparamétrico, que são modelos usuais ao se tratar de dados de sobrevivência. As covariáveis: avaliação e extensão da doença, tratamento feito na instituição, ano de entrada no estudo e idade foram identificadas como fatores que influenciam no tempo de sobrevida de pacientes com câncer de mama. A partir dos modelos ajustados foi constatado que o risco de óbito foi maior para os pacientes em que a doença já havia se espalhado. Foi verificado também que a taxa de falha foi maior para pacientes que foram tratados com Radioterapia. Para comparação da qualidade de predição dos modelos ajustados foi utilizada uma versão tempo-dependente da área sob a curva ROC, AUC(t), que evidenciou um poder de predição um pouco melhor do modelo aditivo semiparamétrico em relação ao modelo de Cox.

Palavras-chave: Análise de Sobrevivência; Câncer de mama; Modelo aditivo de Aalen; Modelo aditivo semiparamétrico; Modelo de regressão de Cox.

Sumário

1 Introdução

O câncer de mama é um tumor maligno resultante da multiplicação de células anormais da mama. Segundo o Instituto Nacional de Câncer - INCA, esse é o tipo de doença mais comum entre as mulheres no Brasil e no mundo, tanto em países em desenvolvimento quanto em países desenvolvidos (INCA, 2015). Devido à dimensão da doença, hoje em dia existe uma das mais populares campanhas de prevenção, conhecida como "Outubro Rosa", que visa chamar atenção para a realidade atual do câncer de mama.

De acordo com a *International Agency for Research on Cancer* (IARC, 2012), o câncer de mama foi, em 2012, o segundo tipo de câncer mais incidente no mundo, cerca de 1,7 milhão, o que representa 25% de todos os tipos de câncer diagnosticados em mulheres. Para 2020, a *International Agency for Research on Cancer* (IARC, 2012) estima que o número de casos novos de câncer de mama no Brasil chegue a 83.035. Segundo o INCA (2016), são esperados 57.960 casos novos de câncer de mama para o Brasil em 2016. Desses novos casos, 10.970 seriam na região Sul, sendo 3.730 no estado do Paraná e 840 na capital Curitiba (INCA, 2016).

Importantes avanços na abordagem do câncer de mama aconteceram nos últimos anos, principalmente no que diz respeito à cirurgias menos agressivas, assim como a busca da individualização do tratamento. O tratamento varia de acordo com o estadiamento da doença, suas características biológicas, bem como das condições da paciente (idade, status menopausal e comorbidades).

O câncer é uma doença que pode se espalhar para outros órgãos do corpo, fenômeno chamado de metástase. Quando se trata da região da mama, o mais comum é que a doença afete os ossos, os pulmões, o fígado ou o cérebro (IMAMA, 2014). O quadro clínico dessa doença é agravado quando o diagnóstico é realizado em fase tardia, como na maioria das vezes, especialmente nas classes com menor poder aquisitivo. Além disso, fatores prognósticos, que são parâmetros possíveis de serem mensurados no momento do diagnóstico, servem como preditor da sobrevida do paciente. Um prognóstico considerado clássico é o tamanho do tumor, que no momento do diagnóstico é um fator determinante na indicação do tratamento (ABREU; KOIFMAN, 2002). Quando a doença é diagnosticada no início, o tratamento tem maior potencial curativo. Entretanto, quando há evidências de metástase (doença se espalhou), os objetivos principais do tratamento se tornam: prolongar a sobrevida e melhorar a qualidade de vida dos pacientes (INCA, 2014). Outro fator prognóstico usual é a idade. Segundo trabalho realizado por Abreu e Koifman (2002), a menor probabilidade de sobrevivência foi observada principalmente no grupo de mulheres com idade igual ou inferior a 35 anos e com mais de 75 anos de idade. A maior probabilidade de sobrevida foi verificada no grupo de mulheres com idade entre 45 e 49 anos.

Pesquisas recentes apontam que o fato do paciente estar casado também favorece

no tratamento do câncer. Segundo pesquisa publicada por Martínez et al. (2016), que analisou dados de quase 800 mil pessoas da Califórnia, os cônjuges levam seus parceiros a consultas médicas e sessões de quimioterapia, dão apoio em caso de depressão e lembram a eles de tomar os medicamentos, o que favorece no tratamento dos pacientes com câncer. Os autores Martínez et al. (2016) também comentam que os médicos devem perguntar a seus pacientes solteiros se há alguém em seu circulo mais próximo que possa ajudar o indivíduo a enfrentar as dificuldades físicas e emocionais durante o tratamento. Ressaltaram também que mais atenção deve ser dedicada a esse efeito adverso de ser solteiro. Outro estudo realizado com 168 pacientes diagnosticados com câncer de pulmão avançado nos Estados Unidos, mostrou que 33% dos doentes casados estavam ainda vivos após três anos de tratamento em comparação com 10% dos solteiros (NICHOLS, 2012).

A variável estado civil está disponível na base de dados utilizada neste trabalho, entretanto a mesma não foi coletada visando avaliar o aspecto mencionado, tendo em vista que essa questão tem sido alvo de pesquisas mais recentes. Desta forma, não é possível saber se os pacientes declarados solteiros têm ou não pessoas próximas que os auxiliam. Do mesmo modo, se os declarados casados têm de fato o apoio mencionado no estudo realizado por Martínez et al. (2016).

Tendo em vista que um dos objetivos do tratamento do câncer é prolongar a sobrevida dos pacientes, diversos estudos foram analisados por meio de metodologias estatísticas propostas para a análise de sobrevivência, comum em estudos na área da saúde. Como exemplo, tem-se o estudo sobre sobrevivência de mulheres com diagnóstico de câncer de mama no município do Rio de Janeiro relatado por Santos (2013a), que foi analisado por meio do modelo de regressão de Cox. De acordo com esse estudo, as covariáveis idade e tipo de tratamento foram consideradas significativas ao modelo de sobrevivência ajustado. Em outro estudo analisado por Abadi et al. (2011), os fatores associados à sobrevida de mulheres com diagnóstico de câncer também foi verificado por meio das covariáveis idade e tratamento, além do estágio da doença. Esse estudo tinha como objetivo a comparação entre os modelos de regressão de Cox e de Aalen e teve como base um conjunto de dados com 14.826 mulheres diagnosticadas com câncer da mama na Columbia Britânica, Canadá. Para a análise dos dados desse estudo, os autores ajustaram o modelo de regressão de Cox e o modelo aditivo de Aalen para poder compará-los através de uma aplicação prática. Os modelos citados permitem que seja analisado o tempo até a ocorrência de um evento levando em consideração os fatores associados, com a finalidade de investigar o seu efeito sobre a sobrevida de pacientes com câncer. As principais conclusões dos autores foram de que se a suposição de riscos proporcionais não for válida para o modelo de Cox, o modelo aditivo de Aalen é uma alternativa adequada. Contudo, se a suposição de riscos proporcionais for válida, argumentaram que ambos os modelos são adequados e que as informações produzidas por cada um deles complementam a análise.

Em virtude dos fatos mencionados, identificar fatores associados à sobrevida de pacientes com câncer se torna de suma importância. Os trabalhos de Abadi et al. (2011) e Santos (2013a), já citados, tiveram como interesse identificar tais fatores. Nesse contexto, os dados de 3.542 pacientes com câncer de mama, tratadas em um centro médico de Curitiba no período de 1990 a 2009, foram analisados neste trabalho com o objetivo de identificar fatores associados à sobrevida desses pacientes. Para tanto, foram utilizados métodos estatísticos propostos na literatura para a análise de dados de sobrevivência.

No geral, este trabalho está estruturado em três capítulos. No capítulo 2 são descritos os dados mencionados e a metodologia estatística utilizada para a análise dos mesmos. No Capítulo 3 são apresentados os resultados obtidos a partir do modelo de regressão de Cox, Aalen e riscos aditivos semiparamétricos, bem como a qualidade de predição dos mesmos. No capítulo 4 são apresentadas as considerações finais da análise realizada no presente trabalho.

2 Material e Métodos

2.1 Material

A seguir, são descritos o conjunto de dados e os métodos estatísticos que foram utilizados neste trabalho para a análise dos mesmos.

2.1.1 Conjunto de dados

Para a realização das análises, foi utilizado um banco de dados fornecido por um centro médico de Curitiba que contém informações de pacientes diagnosticados com câncer de mama. As informações foram coletadas em um período de 20 anos de série histórica, compreendendo os anos de 1990 a 2009. Foram observados 3.575 pacientes, dos quais 1.531 foram a óbito. Os demais, denominados censura, não foram a óbito ou deixaram o estudo por outras razões (Liga Paranaense de Combate ao Câncer, 2011).

Do total de pacientes no estudo observou-se que apenas 33 deles eram do sexo masculino, o que corresponde a 0,9%. Decidiu-se, desse modo, pela remoção desses pacientes do estudo, tendo em vista o câncer de mama ser mais comum em pacientes do sexo feminino. Após a exclusão, o conjunto de dados ficou com 3.542 pacientes e um percentual de 57% de censura. O Quadro 1 apresenta a descrição das covariáveis disponíveis no conjunto de dados descrito.

A maioria das variáveis disponíveis no conjunto de dados são categóricas, desta forma, neste trabalho as categorias das mesmas foram consideradas como variáveis *dummy* para a construção dos modelos ajustados. Variáveis *dummy* assumem apenas um de dois valores, em geral 0 ou 1, para indicar a presença ou ausência de determinada característica.

2.1.2 Recursos computacionais

Para as análises e tratamento do banco de dados foi utilizado o *software* livre R, versão 3.2.2 (R CORE TEAM, 2015). Tal *software* permite a utilização de códigos abertos, propiciando um amplo espaço para análises estatísticas, gerenciamento de banco de dados e também análises gráficas.

2.2 Métodos

Nesta seção, são apresentados os métodos estatísticos utilizados neste trabalho. Inicialmente, faz-se uma breve revisão de conceitos básicos em análise de sobrevivência e, em seguida, dos métodos e modelos de regressão considerados.

Quadro 1 - Descrição das covariáveis disponíveis no conjunto de dados de câncer de mama

COVARIÁVEL	DESCRIÇÃO	CATEGORIA
Tumor	Presença de mais de um tumor	- Sim - Não
Sexo	Sexo do paciente	- Feminino - Masculino
Civil	Estado civil do paciente	- Solteiro - Casado - Viúvo - Outros
Idade	Idade do paciente	- Em anos
AED	Avaliação e Extensão da Doença	- In situ + Localizado - Extensão direta - Envio por linfonodos regionais - Extensão direta com envio linfonodos regionais - Metástase - Sem Informação / Não aplicável
Topografia	Localização do Tumor	- Localização de acordo com a CID-O (Classificação Internacional de Doenças para Oncologia)
Morfologia	Tipo de tumor	- Carcinoma - Adenocarcinoma - Comedocarcinoma - Neoplasma - Sarcoma - Outros
GD	Grau de Diferenciação	- Pouco diferenciado - Moderadamente diferenciado - Bem diferenciado - Indiferenciado - Sem Informação
Estadiamento	Espalhamento do câncer pelo corpo	Tumor: - Limitado a mama, sem metástase - Pode estar estendido além da mama - Envolve regiões vizinhas - Com metástase a distância
TFI	Tratamento feito na instituição	- Cirurgia - Radioterapia - Quimioterapia - Hormonioterapia - Outros
Ano de entrada no estudo	Ano de entrada no estudo	- De 1990 a 2005

Fonte: Elaborado pelos autores (2016).

2.2.1 Análise de sobrevivência - Conceitos básicos

A análise de sobrevivência, técnica muito utilizada atualmente em estudos na área da saúde, tem como interesse avaliar o tempo até a ocorrência de um evento. Este tempo é denominado tempo de falha. Neste trabalho, o evento de interesse refere-se ao tempo, em meses, a partir da data de diagnóstico da doença até o óbito de pacientes com câncer de mama. A presença de censuras, ou seja, informação parcial da resposta, é comum em diversos estudos em que se tem interesse em avaliar o tempo de sobrevivência. Sem a

presença de censuras, técnicas clássicas de estatística poderiam ser utilizadas. Métodos de análise de sobrevivência possibilitam considerar na análise estatística a informação contida nos dados censurados (COLOSIMO; GIOLO, 2006).

A variável resposta em dados de sobrevivência é constituída pelos tempos de falha e censuras. Ou seja, os dados do indivíduo i $(i = 1, ..., n)$ são representados geralmente pela tripla $(t_i, \delta_i, \mathbf{x}_i)$, com t_i o tempo até a ocorrência do evento de interesse (tempo de falha), δ_i a variável indicadora de falha (1, se t_i é um tempo de falha e 0, se censura) e $\mathbf{x}_i$ um vetor de covariáveis (LAWLESS, 2011).

2.2.1.1 Especificação da função de Sobrevivência

Seja T uma variável aleatória denotando o tempo até a ocorrência de um evento, não-negativa e usualmente contínua. Uma das principais funções probabilísticas usadas para descrever a distribuição de T é a função de sobrevivência, denotada por $S(t)$, que fornece a probabilidade de um indivíduo sobreviver ao tempo t, isto é, $S(t) = P(T \geq t)$. Ainda, define-se a função taxa de falha de T como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.1}$$

A partir de (2.1), tem-se que $\lambda(t)$, usualmente denominada "taxa instantânea de falha", representa a taxa de falha instantânea no tempo t condicionada à sobrevivência até o tempo t. O numerador dessa função representa a probabilidade aproximada de um indivíduo que ainda não manifestou o evento até o tempo t, vir a manifestá-lo no instante imediatamente posterior a t. A função taxa de falha é importante para dados de sobrevivência, por ser, em geral, mais informativa do que a função de sobrevivência. Isso acontece porque para diferentes situações as funções de sobrevivência podem ter formas semelhantes, enquanto suas respectivas funções de taxa de falha podem diferir bastante (COLOSIMO; GIOLO, 2006).

2.2.1.2 Métodos de estimação não paramétricos para S(t)

Técnicas usuais de estatística para uma análise descritiva geralmente consistem em encontrar medidas de tendência central e variabilidade. Entretanto, como a presença de observações censuradas gera dificuldades na utilização de técnicas usuais, o principal componente para realizar a análise descritiva para os dados de tempo de vida é a função de sobrevivência. Em virtude disso, o procedimento inicial é estimar a função de sobrevivência e, então, a partir dela, estimar as estatísticas de interesse.

Neste contexto, para a análise exploratória dos dados mencionados na Seção 2.1.1 foi utilizado o estimador de Kaplan-Meier (KAPLAN; MEIER, 1958), o qual é um método muito utilizado para estimar a função de sobrevivência $S(t)$ quando o conjunto de dados

apresenta censuras. Para a utilização desse estimador não é necessário assumir nenhuma distribuição de probabilidade para a variável aleatória T, ou seja, este estimador é não paramétrico. A função de sobrevivência estimada é uma função escada, com degraus nos tempos em que foram observadas as falhas. A partir deste estimador pode-se obter estatísticas de interesse como o tempo médio, mediano e alguns percentis. O estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j \leq t} (\frac{n_j - d_j}{n_j}) = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j}), \tag{2.2}$$

em que $t_1, ..., t_k$ representam os k tempos distintos e ordenados de falha, d_j é o número de falhas em t_j e n_j é o número de indivíduos sob risco em t_j, para $j = 1, ..., k$.

Técnicas não paramétricas são muito importantes para a análise de dados de sobrevivência devido à sua simplicidade, porém, estas apresentam limitações quando há interesse em considerar um conjunto de covariáveis na análise. Em situações como essa, uma alternativa é a de se fazer uso de modelos de regressão apropriados para esse tipo de dados (COLOSIMO; GIOLO, 2006). Alguns desses modelos foram utilizados neste trabalho e serão descritos nas seções seguintes.

Para a comparação de curvas de sobrevivência associadas às categorias de cada covariável analisada pode-se utilizar o teste *logrank* (MANTEL, 1966). Este teste é muito utilizado em análise de sobrevivência quando a razão das funções taxa de falha a serem comparadas for aproximadamente constante. Entretanto, devido a grande quantidade de observações na base de dados, não se torna viável a realização desse teste, pois para quaisquer diferenças entre as curvas, mesmo que não significativas do ponto de vista prático, o teste as indicaria como sendo significativas. Desta forma, a comparação das curvas de sobrevivência foi embasada em procedimentos gráficos.

2.2.2 Modelos de regressão em análise de sobrevivência

Diversos modelos de regressão têm sido propostos na literatura para a análise de dados de sobrevivência. Alguns deles, que foram utilizados neste estudo, são apresentados brevemente a seguir.

2.2.3 Modelo de regressão de Cox

Inicialmente, foi considerado o modelo de regressão de Cox, que é amplamente utilizado em estudos clínicos devido à sua versatilidade. A expressão desse modelo em termos da função taxa de falha é dada, para o i-ésimo indivíduo em um dado tempo t,

por:

$$\lambda(t \mid \mathbf{x}_i) = \lambda_0(t)\, \exp\Big(\sum_{k=1}^{p} \beta_k\, x_{ik}\Big), \tag{2.3}$$

em que p é o número de covariáveis, x_{ik} é o valor da k-ésima covariável observada para o i-ésimo indivíduo, β_k é o parâmetro que descreve o efeito da k-ésima covariável e $\lambda_0(t)$ é uma função do tempo não-negativa e não especificada, usualmente denominada função taxa de falha de base ou basal, que corresponde a taxa comum a todos os indivíduos.

O modelo de Cox é um modelo semiparamétrico por considerar $\lambda_0(t)$ arbitrária, ou seja, por não assumir nenhuma forma paramétrica para essa função, e pela relação entre as covariáveis $\mathbf{x}$ e $\boldsymbol{\beta}$ assumir uma função paramétrica.

A suposição básica para o uso do modelo de Cox é que as taxas de falha sejam proporcionais, isto é, a razão das funções taxa de falha para os indivíduos i e j dada por:

$$\frac{\lambda(t \,|\, \mathbf{x}_i)}{\lambda(t \,|\, \mathbf{x}_j)} = \frac{\lambda_0(t)\exp\{\mathbf{x}_i^{'}\boldsymbol{\beta}\}}{\lambda_0(t)\exp\{\mathbf{x}_j^{'}\boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i^{'}\boldsymbol{\beta} - \mathbf{x}_j^{'}\boldsymbol{\beta}\}, \tag{2.4}$$

não depende do tempo. Por exemplo, se um indivíduo no início do estudo tem uma taxa de falha igual a três vezes a de um segundo indivíduo, então esta razão será a mesma para todo o período de acompanhamento (COLOSIMO; GIOLO, 2006). Tendo em vista que o modelo de regressão de Cox pode ser utilizado somente se o pressuposto de taxas de falha proporcionais for válido, foi considerado outro modelo que não exige tal suposição, o modelo aditivo de Aalen, que é apresentado a seguir.

2.2.4 Modelo aditivo de Aalen

Para o modelo aditivo de Aalen, a função taxa de falha em um dado tempo t é dada por:

$$\lambda(t \mid \mathbf{x}_i(t)) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t)\, x_{ik}(t), \tag{2.5}$$

em que p, x_{ik} e β_k são especificados como na equação (2.3), porém x_{ik} e β_k variam com o tempo. O componente $\beta_0(t)$ corresponde à função taxa de falha de base.

O modelo de Aalen permite que tanto o efeito quanto os valores das covariáveis variem com o tempo e assume que elas atuam de forma aditiva sobre a função taxa de falha. Também é denominado modelo dinâmico por permitir que as funções de regressão $\boldsymbol{\beta}(t)$ e as covariáveis $\mathbf{x}(t)$ variem ao longo do tempo.

Devido às dificuldades em estimar $\beta_k(t)$ diretamente, Aalen propôs um estimador para $\mathbf{B}(t) = (B_0(t), B_1(t)..., B_p(t))$, em que $B_k(t) = \int_0^t \beta_k(u)du$. O estimador proposto é denominado estimador de mínimos quadrados de Aalen, que é dado por:

$$\hat{\mathbf{B}}(t) = \sum_{t_i \leq t} \mathbf{Z}(t_i)\mathbf{I}(t_i), \tag{2.6}$$

em que $\mathbf{I}(t_i)$ é um vetor com o i-ésimo elemento igual a 1 se o evento ocorre para o indivíduo no tempo t e 0, caso contrário, e $\mathbf{Z}(t_i)$ é a inversa generalizada de $\mathbf{X}(t_i)$, em que $\mathbf{X}(t_i)$ representa a matriz de covariáveis para os indivíduos sob risco no tempo t. O estimador $\hat{\mathbf{B}}(t)$ encontra-se somente disponível sobre o intervalo de tempo em que $\mathbf{X}(t)$ tem posto completo, ou seja, a estimação pára quando $\mathbf{X}(t)$ (vetor dos coeficientes na forma matricial) deixa de ser uma matriz não-singular, que é uma consequência do princípio não-paramétrico. O valor de t em que tal fato ocorre é denotado por τ.

Os gráficos dos coeficientes de regressão acumulados $\hat{B}_k(t),\ k = 0, 1, \ldots, p$, versus os tempos, é o principal foco do modelo de Aalen. Através da inclinação desses coeficientes é possível verificar se uma particular covariável apresenta efeito constante ou tempo-dependente. Inclinação positiva em certo período indica que a covariável exerce efeito crescente sobre a taxa de falha e, inclinação negativa, efeito decrescente. Inclinação próxima de zero será observada nos períodos em que a covariável não tem efeito sobre a taxa de falha. Os intervalos de confiança para $B_k(t)$ são obtidos a partir de $\hat{B}_k(t) \pm z_{\frac{\alpha}{2}}\sqrt{\hat{Var}[\hat{B}_k(t)]}$, que são utilizados para concluir sobre as hipóteses nulas: (a) efeito não significativo das covariáveis, isto é, $H_{0a} : \beta_k(t) = 0$ e (b) efeito constante das covariáveis, isto é, $H_{0b} : \beta_k(t) = \beta_k$. As covariáveis contínuas são geralmente centradas em seus respectivos valores médios. Assim, $\beta_0(t)$ corresponderá à taxa de falha de base para um indivíduo com valor médio para todas as covariáveis contínuas e as categorias de referência para as categóricas.

A diferença entre os modelos de Cox e de Aalen é que o de Cox apresenta dois componentes, um não paramétrico (função taxa de falha de base) e outro paramétrico (função que relaciona as covariáveis com seus parâmetros), enquanto o de Aalen é um modelo completamente não-paramétrico no sentido que funções são ajustadas e não parâmetros, o que faz o modelo ser bastante flexível, pois permite estimar as funções usando informação local. No modelo de Aalen as funções de sobrevivência não são necessariamente monótonas (dado que o modelo não é paramétrico), ou seja, pode haver intervalos nos quais ela é crescente e intervalos nos quais é decrescente. Desta forma, a função taxa de falha pode assumir valores negativos e isto é visto como uma desvantagem desse modelo (AALEN, 1989). Outra desvantagem desse modelo é que as estimativas para a função taxa de falha acumulada encontra-se somente disponível para $t \leq \tau$.

Como pode haver casos em que nem todas as covariáveis apresentem necessariamente efeito variando com o tempo, uma extensão do modelo aditivo de Aalen, a qual

permite que parte dos coeficientes variem com o tempo e parte não, foi proposta por McKeague e Sasieni (1994). Tal modelo, denominado modelo de riscos aditivos semiparamétrico, é expresso por:

$$\lambda(t \,|\, \mathbf{x}_i(t),\, \mathbf{z}_i(t)) = \beta_0(t) + \mathbf{x}_i^T(t)\boldsymbol{\beta}(t) + \mathbf{z}_i^T(t)\boldsymbol{\gamma}, \tag{2.7}$$

em que $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))$ e $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{il}(t))$ são vetores de covariáveis de dimensão p e l, respectivamente, assim como $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_l)$ são, respectivamente, os vetores de coeficientes de regressão dependentes e independentes do tempo. Ressalta-se que nem todas as covariáveis em $\mathbf{x}_i(t)$ e $\mathbf{z}_i(t)$ precisam ser dependentes do tempo.

Os estimadores para $\beta_0(t)$, $\boldsymbol{\beta}(t)$ e $\boldsymbol{\gamma}$ são obtidos a partir de equações de mínimos quadrados (HUFFER; MCKEAGUE, 1991). Entretanto, devido às dificuldades em se estimar $\beta_k(t)$ diretamente, estimam-se as respectivas funções de regressão acumuladas, $B_k(t) = \int_0^t \beta_k(u)du,\ k = 0, 1, \ldots, p$. O modelo expresso em (2.6) é um submodelo do modelo aditivo de Aalen (2.5), logo, para verificar se o mesmo se ajusta bem aos dados são utilizados procedimentos similares aos usados para o modelo de Aalen, que são apresentados na Seção 2.2.6.

2.2.5 Métodos de seleção de covariáveis

Devido ao número de covariáveis disponíveis no banco de dados, foi utilizado o método de seleção de covariáveis proposta por Collett (2003). De acordo com esse método, são ajustados, em um passo inicial, um modelo para cada uma das covariáveis, sendo que apenas as que forem significativos ao nível de 0,10 permanecem para a segunda etapa. As covariáveis selecionadas no primeiro passo são, em um passo seguinte, ajustadas conjuntamente já que algumas podem deixar de ser significativas na presença de outras devido à presença de correlação entre elas. Caso isso ocorra, são ajustados modelos reduzidos, retirando-se uma covariável de cada vez. A cada passo é também avaliado por meio do teste da razão de verossimilhanças, a contribuição de cada covariável para a função de verossimilhança. Somente são mantidas no modelo aquelas que atingirem a significância estabelecida. Por fim, ajusta-se um novo modelo contendo apenas as covariáveis mantidas no segundo passo.

Outra forma de seleção de covariáveis envolve algoritmos computacionais tais como o *backward*, *forward* e *stepwise*. A seleção *forward* parte de um modelo inicial sem covariáveis, apenas o intercepto. O intuito do método é adicionar uma covariável de cada vez, iniciando-se com a inclusão da covariável com maior correlação com a variável resposta. Uma vez que a primeira covariável foi selecionada, o passo seguinte é encontrar a próxima

covariável com a maior correlação com a variável resposta na presença da primeira no modelo e, assim por diante.

Enquanto o método *forward* começa sem nenhuma covariável no modelo, o método *backward* faz o caminho oposto. Inclui inicialmente no modelo todas as covariáveis e depois, por etapas, vai eliminando as não correlacionadas com a resposta. Por sua vez, o método *stepwise*, assim como o método *forward*, tem ínicio com a inclusão da covariável que apresentar a maior correlação com a variável resposta. Contudo, nos passos subsequentes uma covariável que tenha sido incluída em uma passo anterior pode vir a ser excluída em uma passo seguinte. O algoritmo é realizado até que não possa ser incluída ou excluída nenhuma covariável (FERREIRA, 2012).

2.2.6 Adequação dos modelos ajustados

Como em qualquer outro modelo estatístico, a avaliação da adequação do modelo ajustado é muito importante. Para esse propósito, foram utilizados, tanto para o modelo de Cox quanto para o modelo aditivo de Aalen e o modelo aditivo semiparamétrico os resíduos de Cox e Snell (1968). Esses resíduos são definidos por:

$$\hat{e}_i = \hat{\Lambda}(t_i \,|\, \mathbf{x}_i) = \begin{cases} \hat{\Lambda}_0(t_i)\exp\left\{\sum_{k=1}^{p} x_{ik}\hat{\beta}_k\right\} & \text{se modelo de Cox} \\ \hat{B}_0(t_i) + \mathbf{x}_i^T\hat{\mathbf{B}}(t_i) & \text{se modelo aditivo de Aalen} \\ \hat{B}_0(t_i) + \mathbf{x}_i^T\hat{\mathbf{B}}(t_i) + \mathbf{z}_i^T\hat{\boldsymbol{\gamma}}\, t_i & \text{se modelo adit. semiparamétrico} \end{cases} \tag{2.8}$$

em que $\hat{\Lambda}(.)$ é a função taxa de falha acumulada obtida a partir do modelo ajustado. Se o modelo for adequado, os resíduos $\hat{e}_i$ devem seguir uma distribuição exponencial padrão (LAWLESS, 2011). Para verificar tal fato, é comum o uso de técnicas gráficas. Uma proposta é construir o gráfico das probabilidades de sobrevivência dos resíduos $\hat{e}_i$, obtidas pelo estimador de Kaplan-Meier, versus as probabilidades de sobrevivência destes resíduos obtidas pelo modelo exponencial padrão que deve ser aproximadamente uma reta com inclinação 1 para que o modelo seja considerado adequado.

Para verificar a presença de observações atípicas e também a forma funcional das covariáveis (linear, quadrática, etc.) são construídos gráficos dos resíduos *martingal* e *deviance* versus os tempos. Se o modelo for apropriado, os resíduos devem apresentar um comportamento aleatório em torno de zero. Os resíduos *martingal* e *deviance* são expressos para $i = 1, ..., n$, respectivamente, por:

$$\hat{m}_i = \delta_i - \hat{e}_i \tag{2.9}$$

e

$$\hat{d}_i = \text{sinal}(\hat{m}_i)[-2(\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i))]^{1/2}, \tag{2.10}$$

em que δ_i é a variável indicadora de falha e $\hat{e_i}$ o resíduo de Cox-Snell.

Para avaliar a suposição de taxas de falha proporcionais do modelo de Cox, algumas técnicas encontram-se disponíveis na literatura. Uma das técnicas que foi utilizada neste trabalho consiste em um método gráfico descritivo. Este método baseia-se em dividir os dados em m estratos de acordo com as categorias de cada covariável e, em seguida, estimar $\hat{\Lambda}_{0_j}(t)$ para cada estrato usando o estimador:

$$\hat{\Lambda}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l\in R_j} \exp\{\mathbf{x}_l'\hat{\boldsymbol{\beta}}\}}, \tag{2.11}$$

em que d_j é o número de falhas calculados para cada um dos m estratos em t_j. Se a suposição for válida, as curvas do logaritmo de $\hat{\Lambda}_{0_j}(t)$ versus t, ou $log(t)$, devem apresentar diferenças aproximadamente constantes no tempo.

Uma proposta adicional para verificar a suposição de taxas de falha proporcionais assumida para o modelo de Cox é a de analisar os resíduos de Schoenfeld (1982). O uso desses resíduos é baseado em um resultado apresentado em Grambsch e Therneau (1994), em que os autores sugerem o gráfico de $s_{iq} + \hat{\beta}_q$ versus t, para $q = 1, \ldots, p$, em que s_{iq} são os resíduos padronizados de Schoenfeld, definidos por:

$$\mathbf{s}_i = [\mathcal{I}(\hat{\boldsymbol{\beta}})]^{-1}\mathbf{r}_i, \tag{2.12}$$

em que $\mathcal{I}(\hat{\boldsymbol{\beta}})$ é a matriz de informação observada avaliada em $\hat{\boldsymbol{\beta}}$ e $\mathbf{r}_i = (r_{i1}, \ldots, r_{ip})$ é o vetor de resíduos de Schoenfeld para o i-ésimo indivíduo $(i = 1, \ldots, n)$.

No gráfico sugerido por Grambsch e Therneau (1994), inclinação zero mostra evidências a favor da proporcionalidade. Entretanto, por ser um método gráfico, pode gerar conclusões subjetivas. Desse modo, foram propostas medidas estatísticas, bem como a realização de testes de hipóteses, ambos baseados nos resíduos de Schoenfeld. O coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e os tempos t ou uma função dos tempos $g(t)$, para cada covariável, é uma dessas medidas utilizadas. Valores de ρ próximos de zero mostram não haver evidências para a rejeição da suposição de taxas de falha proporcionais (COLOSIMO; GIOLO, 2006).

2.2.7 Comparação entre os modelos quanto à qualidade de predição

Abadi et al. (2011) comentaram em seu artigo sobre a dificuldade e a inexistência, até então, de métodos estatísticos que possibilitassem a comparação da qualidade de ajuste e de predição dos dois modelos citados (de Cox e de Aalen). As dificuldades citadas se devem ao fato desses dois modelos não serem encaixados, bem como pelo método de estimação do modelo de Aalen se basear na técnica de mínimos quadrados, inviabilizando

o uso de critérios como o AIC (Critério de Informação de Akaike). Contudo, Abadi et al. (2011) observaram que os modelos de Cox e Aalen não deveriam ser vistos como modelos alternativos e sim como métodos complementares, que juntos oferecem uma melhor compreensão dos dados sob estudo.

Entretanto, em um estudo recente, Raminelli (2015) propôs a obtenção da curva ROC e de sua respectiva área abaixo da curva em diversos tempos t, denotada por AUC(t), como uma ferramenta útil que permite a comparação de modelos de sobrevivência com estruturas distintas (aditiva e multiplicativa). A AUC(t), do inglês *Area Under the Curve*, consiste em uma versão tempo-dependente da área sob a curva ROC (do inglês *Receiver Operating Characteristics*), usualmente utilizada para avaliar a qualidade de predição de modelos de regressão para dados com resposta dicotômica (SANTOS, 2013b). Quanto mais próximo a AUC(t) estiver do valor 1, melhor a qualidade de predição do modelo no tempo t.

Para obtenção da curva ROC são necessárias duas medidas: a taxa de verdadeiros positivos, denominada sensibilidade, e a taxa de verdadeiros negativos, denominada especificidade. No contexto de análise de sobrevivência tais medidas tempo-dependentes foram definidas conforme proposta de Heagerty e Zheng (2005):

$$\begin{cases} sens(c,t) = sensibilidade(c,t) = P(M_i(t) > c \,|\, T_i = t) = P(M_i(t) > c \,|\, \delta_i(t) = 1) \\ esp(c,t) = especificidade(c,t) = P(M_i(t) > c \,|\, T_i = t) = P(M_i(t) > c \,|\, \delta_i(t) = 0) \end{cases}, \tag{2.13}$$

em que $M_i(t)$, $i = 1, \ldots, n$ é um marcador tempo-dependente utilizado para previsão de falha no tempo t e $c \in R$ um ponto de corte utilizado como critério para classificar a previsão como falha ou censura no tempo t. Isto posto, os indivíduos são classificados em cada tempo fixo t como falha ou censura com base no seu real status no tempo t. Desta forma, se o evento ocorreu para o indivíduo i, este assumirá o status de censura ($\delta_i(t) = 0$) para todo tempo $t < T_i$ e o status de falha ($\delta_i(t) = 1$) para $t = T_i$.

Neste contexto, para estimar a sensibilidade e a especificidade no tempo t Raminelli (2015) utilizou dois métodos descritos por Heagerty e Zheng (2005). Um deles é baseado no teorema de Bayes e no estimador de Kaplan-Meier (KAPLAN; MEIER, 1958). O outro é baseado no estimador do vizinho mais próximo (AKRITAS, 1994), denotado por NNE (do inglês *Nearest Neighbor Estimator*). O estimador baseado no NNE apresenta vantagens em relação ao primeiro, dentre elas, a sensibilidade e a especificidade são monótonas e limitadas em [0,1] (RAMINELLI, 2015). Para obtenção das estimativas dos erros padrão associados às AUC(t), Raminelli (2015) utilizou o método de reamostragem bootstrap não paramétrico. O procedimento baseado na AUC(t) mencionado, foi utilizado nesse trabalho para comparação da qualidade de predição dos modelos de Cox e de Aalen. Os métodos baseados no estimador de Kaplan-Meier (KM) e no estimador do vizinho mais próximo

(NNE), assim como a obtenção das estimativas dos erros padrão associados às AUC(t) através de reamostragem bootstrap não paramétrico, não foram estudados neste trabalho. Entretanto, mais informações podem ser encontradas em Heagerty e Zheng (2005) e Raminelli (2015), a fim de direcionar os interessados quanto a estes assuntos.

3 Resultados e Discussão

A seguir, são apresentados os resultados obtidos neste trabalho com o auxílio dos pacotes *survival*, *timereg* e *survivalROC*, disponíveis no *software* livre R.

3.1 Análise Exploratória

Em primeiro lugar, foi realizada uma análise descritiva do conjunto de dados. Verificou-se que as covariáveis: Tumor, Topografia, Morfologia e Grau de Diferenciação (Quadro 1) apresentaram frequência elevada em uma única categoria e, sendo assim, não foram consideradas nas análises, as frequências podem ser vistas no Apêndice A. Logo, a Tabela 1 apresenta, para as covariáveis potencialmente importantes, as frequências absolutas e respectivos percentuais de pacientes, falhas e censuras. A partir da Tabela 1, observa-se que a maioria das covariáveis são categóricas, com exceção da Idade a qual foi apresentada de acordo com os seus respectivos quartis para facilitar a análise exploratória.

Considerando inicialmente a covariável Avaliação e Extensão da Doença (AED), observa-se (Tabela 1) que os maiores percentuais de pacientes estão nas categorias 4 e 1. Para o Estadiamento, 38% dos pacientes encontram-se na categoria 2 (Tumor que pode estar estendido além da mama, sem metástase) e 24% na categoria 3 (Tumor que envolve regiões vizinhas). Já para o Tratamento Feito na Instituição (TFI) tem-se que 58% dos pacientes realizaram Cirurgia com adição de outros tratamentos coadjuvantes. A idade média dos pacientes foi igual a 53 anos com desvio padrão de 13. Quanto ao ano de entrada no estudo, o maior percentual de entrada foi observado entre os anos 1995 e 2000 (45%). Por último, a covariável Estado Civil apresentou 63% dos pacientes na categoria 2 (Casado).

Quanto ao percentual de falhas, observou-se que para as covariáveis Idade e Estado Civil os percentuais estão bem distribuídos entre as categorias. Já com as covariáveis Estadiamento e AED tem-se os maiores percentuais nas categorias em que a doença está mais espalhada. Para a variável TFI observou-se que o percentual de falha foi menor para os pacientes que realizaram cirurgia. Para a covariável ano de entrada no estudo o percentual de falha diminuiu com o decorrer dos anos. Vale ressaltar que o percentual de falha é menor para os anos mais recentes de entrada no estudo devido ao tempo de acompanhamento desses pacientes ser inferior ao tempo de acompanhamento dos pacientes que entraram no estudo nos períodos mais antigos. Essa covariável foi incluída no modelo com o intuito de identificar se a tecnologia utilizada para o tratamento da doença impacta na taxa de óbito ao longo dos anos.

Para o conjunto de dados desse estudo, a curva de sobrevivência global (sem considerar covariáveis) foi obtida por meio do estimador de Kaplan-Meier. Tal curva, com

Tabela 1 – Frequências absolutas e respectivos percentuais de pacientes, falhas e censuras

Covariável	Categoria	N	N%	Censura	Falha	Cens.%	Falha%
AED	1 : In situ + Localizado	1137	32%	926	211	81%	19%
	2 : Extensão direta	213	6%	152	61	71%	29%
	3 : Envio por linfonodos regionais	95	3%	59	36	62%	38%
	4 : Extensão direta com envio por linfonodos regionais	1579	45%	808	771	51%	49%
	5 : Metástase	484	14%	67	417	14%	86%
	9 : Sem Informação	34	1%	20	14	59%	41%
Estadiamento	1 : Tumor limitado a mama, sem metástase	455	13%	405	50	89%	11%
	2 : Tumor que pode estar estendido além da mama, sem metástase	1331	38%	969	362	73%	27%
	3 : Tumor que envolve regiões vizinhas	853	24%	334	519	39%	61%
	4 : Tumor com desenvolvimento de metástase a distância	430	12%	73	357	17%	83%
	9 : Sem Informação	473	13%	251	222	53%	47%
TFI	1 : Cirurgia + Coadjuvantes	2070	58%	1368	702	66%	34%
	2 : Cirurgia	478	13%	335	143	70%	30%
	3 : Hormonioterapia + Outros	255	7%	92	163	36%	64%
	4 : Quimioterapia + Radioterapia	117	3%	45	72	38%	62%
	5 : Quimioterapia	374	11%	91	283	24%	76%
	6 : Radioterapia	248	7%	101	147	41%	59%
Idade	1 : de 18 a 42	786	22%	466	320	59%	41%
	2 : de 43 a 51	930	26%	577	353	62%	38%
	3 : de 52 a 62	921	26%	530	391	58%	42%
	4 : >62	905	26%	459	446	51%	49%
Ano de entrada no estudo	1 : de 2001 a 2005	794	22%	566	228	71%	29%
	2 : de 1995 a 2000	1609	45%	934	675	58%	42%
	3 : de 1990 a 1994	1139	32%	532	607	47%	53%
Estado Civil	1 : Solteiro	411	12%	234	177	57%	43%
	2 : Casado	2239	63%	1341	898	60%	40%
	3 : Viúvo	612	17%	301	311	49%	51%
	4 : Outros	278	8%	155	123	56%	44%

Fonte: Elaborado pelos autores (2016)

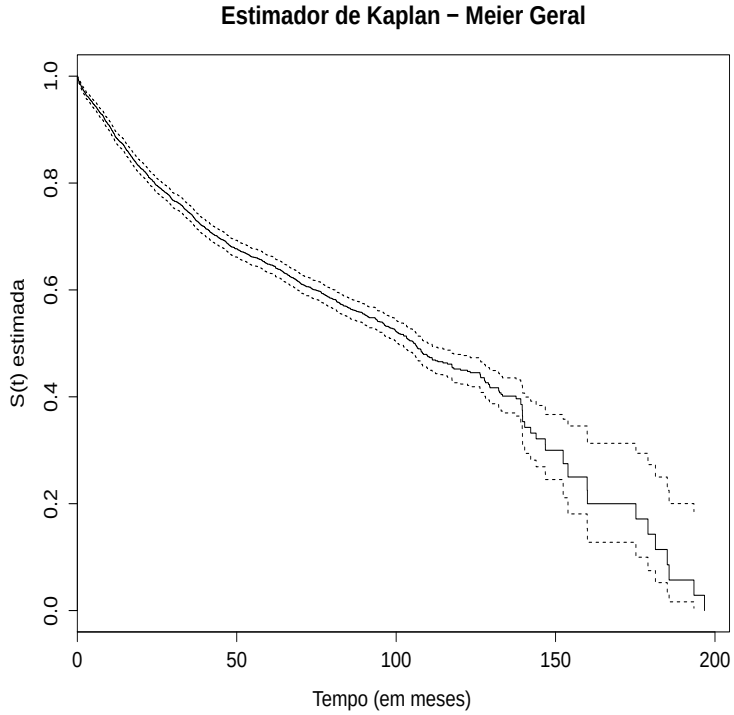
os respectivos intervalos de 95% de confiança para $0 < t < 197$, pode ser visualizada na Figura 1. O tempo mediano de vida foi estimado em 106 meses.

Para cada covariável apresentada na Tabela 1 foi utilizado o estimador de Kaplan-Meier com o propósito de verificar a associação de cada uma delas com o tempo de sobrevida auxiliando, assim, na escolha das covariáveis a serem inseridas nos modelos considerados neste trabalho. As curvas de sobrevivência estimadas para as covariáveis da Tabela 1 encontram-se na Figura 2.

A análise das curvas apresentadas na Figura 2 sugere a presença de diferença não significativa apenas para a covariável idade. Para essa covariável, as curvas das categorias estão bem próximas umas das outras o que pode indicar que as mesmas não diferem entre si ao longo do tempo. Porém devido ao grande número de observações disponíveis (amostra grande), e também o fato de que a Idade é considerada como sendo um fator prognóstico usual, a mesma foi considerada nos modelos ajustados inicialmente.

Foi verificada a presença de associação entre as covariáveis Estadiamento e AED (Tabela 2), ambas covariáveis apresentam características semelhantes com a intenção de medir o quanto o câncer já se espalhou pelo corpo. Desta forma, não foram considerados modelos com a presença das duas covariáveis ao mesmo tempo.

Figura 1 – Curva de sobrevivência global e respectivos intervalos de 95% de confiança obtidos a partir do estimador de Kaplan-Meier



Fonte: Elaborado pelos autores (2016).

Tabela 2 – Tabela de contingência entre as covariáveis Estadiamento e AED

Estad/AED	In Situ	Ext.direta	Linf. Reg	Ext.direta c/ linf	Metástase	Sem Inf.	Total
Sem metástase	375	12	3	59	2	4	455
Além da mama	553	97	56	615	8	2	1331
Regiões Vizinhas	60	72	21	637	58	5	853
Metástase	5	7	1	107	308	2	430
Sem Informação	144	25	14	161	108	21	473
Total	1137	213	95	1579	484	34	3542

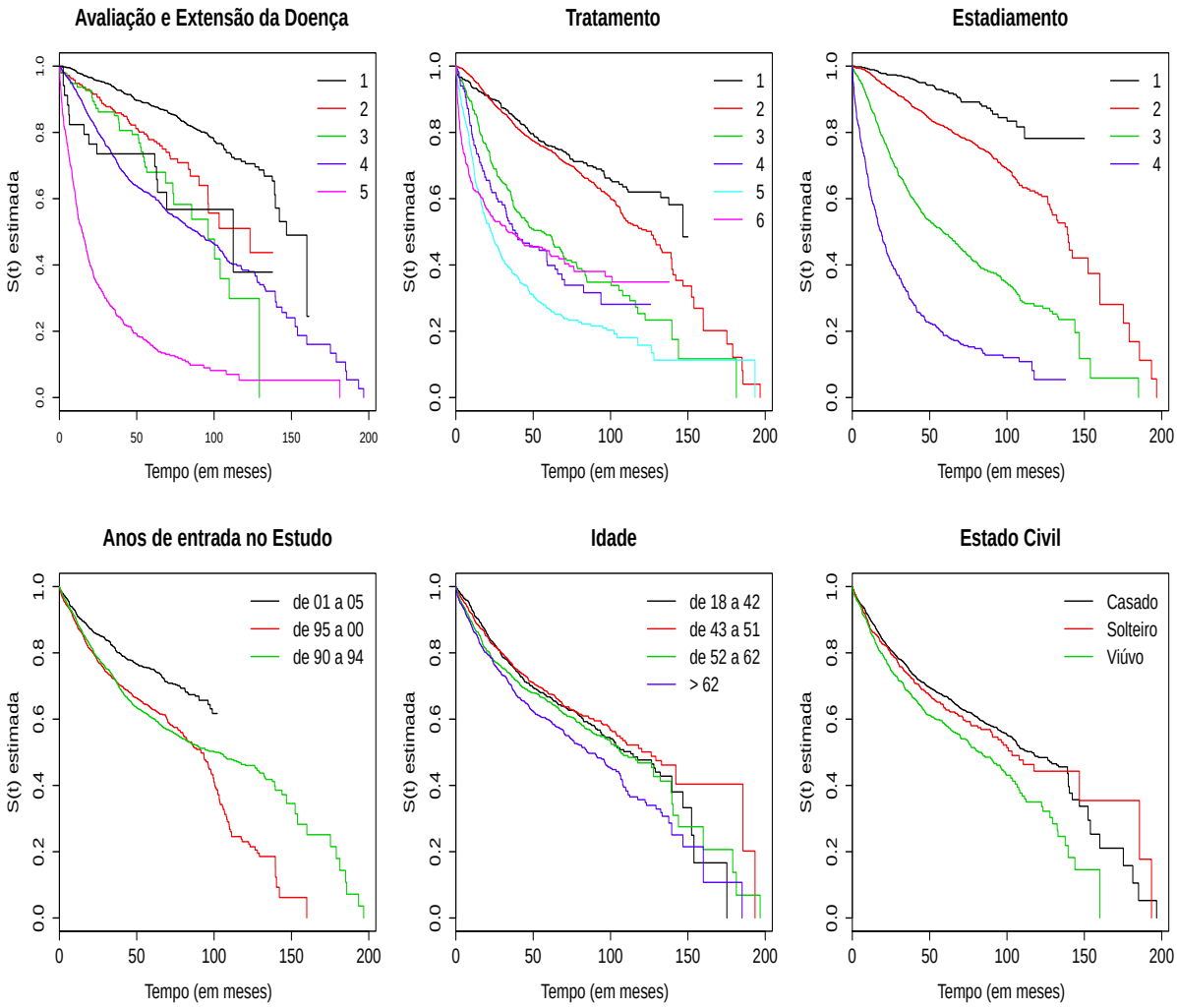
Fonte: Elaborado pelos autores (2016).

Para a construção dos modelos, as categorias de referência para as covariáveis categóricas foram consideradas da seguinte maneira: Categoria "In situ + Localizado"para a covariável AED, "Tumor limitado a mama, sem metástase"para o estadiamento, "Cirurgia + tratamentos coadjuvantes"para o TFI, "de 01 a 05"para o ano de entrada no estudo e "Solteiro"para o estado civil.

3.2 Resultados do Modelo de Regressão de Cox

Após a fase de análise das covariáveis potencialmente importantes é possível ajustar o modelo de regressão de Cox. Para a seleção de covariáveis optou-se por utilizar o método proposto por Collett (2003), descrito na Seção 2.2.5. O passo a passo do resultado desse método encontra-se na Tabela 3. Realizado o procedimento de seleção, o modelo

Figura 2 – Curvas de sobrevivência, para cada covariável, estimadas pelo método de Kaplan-Meier



Fonte: Elaborado pelos autores (2016).

final ficou composto das covariáveis: avaliação e extensão da doença, tratamento feito na instituição, idade e ano de entrada no estudo. Observou-se que o mesmo modelo foi sugerido pelo procedimento de seleção *stepwise*. A escolha do modelo se deu também pelo fato do mesmo indicar o menor valor do AIC, além disso a covariável AED apresenta um percentual de dados faltantes menor em comparação a covariável estadiamento. Isto posto, ajustou-se inicialmente o modelo de regressão de Cox expresso em (2.3).

Como o modelo de Cox assume taxas de falha proporcionais, esta suposição foi investigada com base nos resíduos de Schoenfeld, gráficos $\log(\hat{\Lambda}_{0_j}(t))$ versus os tempos para cada covariável do modelo e, também, do coeficiente de correlação de Pearson ρ. Os resultados obtidos para o coeficiente de correlação são apresentados na Tabela 4 e evidenciam a não violação dessa suposição para as covariáveis presentes no modelo, pois ρ está próximo de zero para a maioria das covariáveis. Já a Figura 3, que apresenta o $\log(\hat{\Lambda}_{0_j}(t))$ versus os tempos para cada covariável do modelo, evidenciam que não há violação significativa da

Tabela 3 – Seleção das covariáveis através do método proposto por Collett

Passos	**Modelo**	**-2 log L(θ)**	**TRV**	**Valor p**	**AIC**
Passo 1	Nulo	-	-	-	
	Tratamento Feito na Instituição (TFI)	22572,06	480,4	0,0000	
	Estadiamento (Est)	22118,52	934	0,0000	
	Avaliação e Extensão da Doença (AED)	22075,79	976,7	0,0000	
	Ano de Entrada no Estudo (AEE)	23003,5	48,99	0,0000	
	Idade	23027,15	25,34	0,0000	
	Estado Civil	23026,61	25,88	0,0000	
Passo 2	Todas Exceto AED	21926,48	-	-	21956
	Todas Exceto AED e TFI	22065,29	138,81	0,0000	22085
	Todas Exceto AED e Estadiamento	22480,11	553,63	0,0000	22502
	Todas Exceto AED e AEE	21947,68	21,2	0,0000	21974
	Todas Exceto AED e Idade	21948,37	21,89	0,0000	21976
	Todas Exceto AED e Estado Civil	21934,98	8,5	0,0367	21959
	Todas Exceto Estadiamento	21874,23	-	-	21906
	Todas Exceto Estadiamento e TFI	22000,2	125,97	0,0000	22022
	Todas Exceto Estadiamento e AED	22480,11	605,88	0,0000	22502
	Todas Exceto Estadiamento e AEE	21902,57	28,34	0,0000	21931
	Todas Exceto Estadiamento e Idade	21907,45	33,22	0,0000	21937
	Todas Exceto Estadiamento e Estado Civil	21879,77	5,54	0,1363	21906

Fonte: Elaborado pelos autores (2016).

suposição de taxas de falha proporcionais, tendo em vista a ausência de cruzamentos marcantes entre as curvas. O gráfico para os resíduos de Schoenfeld encontra-se no Apêndice B e também evidencia a não violação da suposição.

Tabela 4 – Coeficientes de correlação rho(ρ) associados ao modelo de Cox

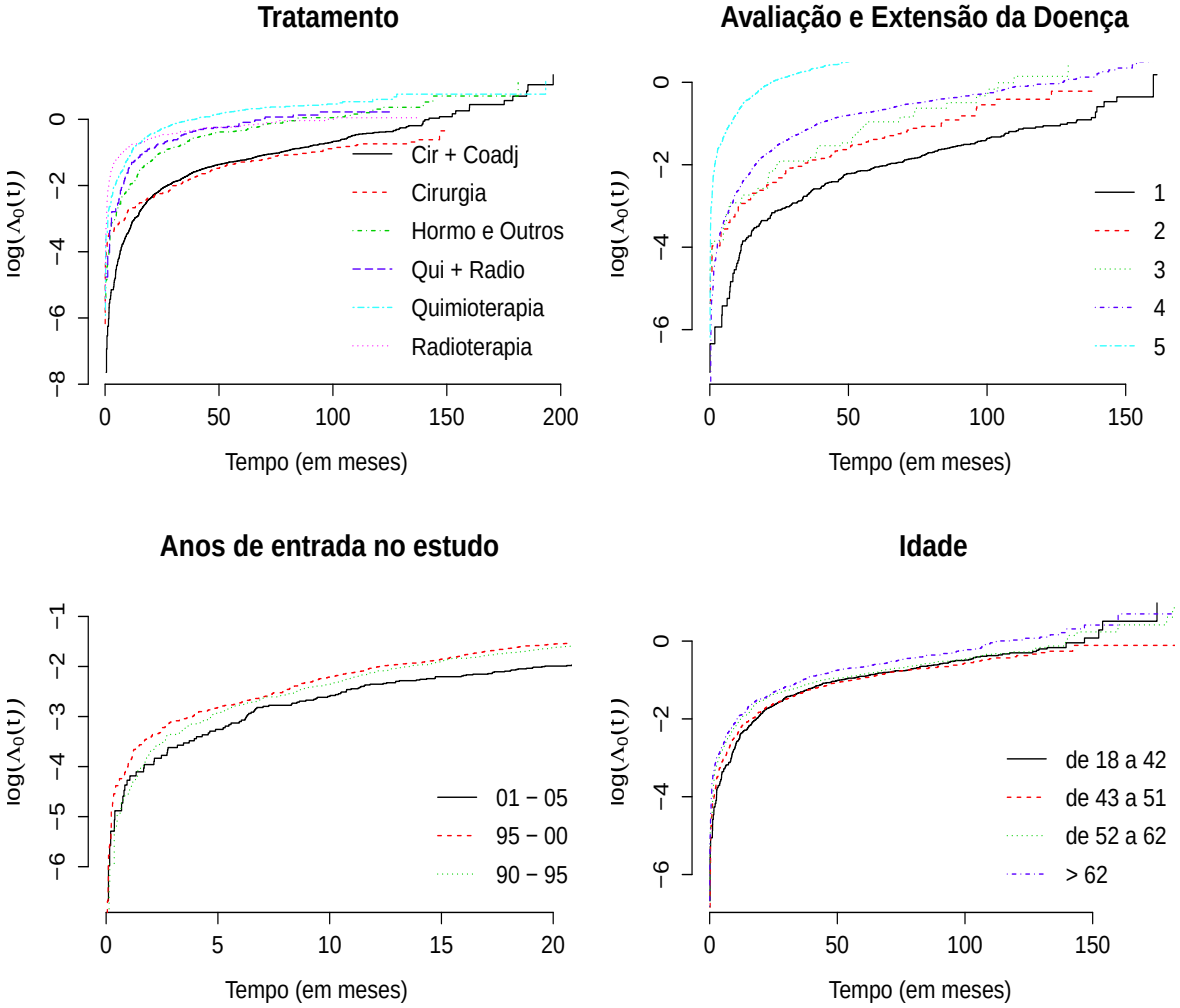
Covariável	ρ
TFI: Cirurgia + Coadjuvantes	0,063180
TFI: Hormonioterapia e Outros	0,044150
TFI: Quimioterapia e Radioterapia	0,010090
TFI: Quimioterapia	-0,009110
TFI: Radioterapia	-0,093110
AED 2: Extensão direta	-0,042430
AED 3 : Envio por linfonodos regionais	-0,016960
AED 4 : Extensão direta com envio por linfonodos regionais	-0,120810
AED 5 : Metástase	-0,141880
AED 9 : Sem Informação	-0,044370
Anos de entrada no estudo: de 1995 a 2000	0,079300
Anos de entrada no estudo: de 1990 a 1994	-0,004900
Idade	-0,034580

Fonte: Elaborado pelos autores (2016).

Para avaliar a qualidade de ajuste global do modelo de Cox foram obtidos os resíduos de Cox-Snell. A partir da Figura 4, pode-se observar que o modelo selecionado apresenta um bom ajuste, uma vez que os resíduos de Cox-Snell seguem distribuição exponencial padrão.

Para avaliar a existência de observações atípicas e também a forma funcional das covariáveis foram obtidos os gráficos dos resíduos *martingal* e *deviance*. A partir da

Figura 3 – $\text{Log}(\hat{\Lambda_{0j}}(t))$ versus tempo para as covariáveis do modelo de Cox final



Fonte: Elaborado pelos autores (2016).

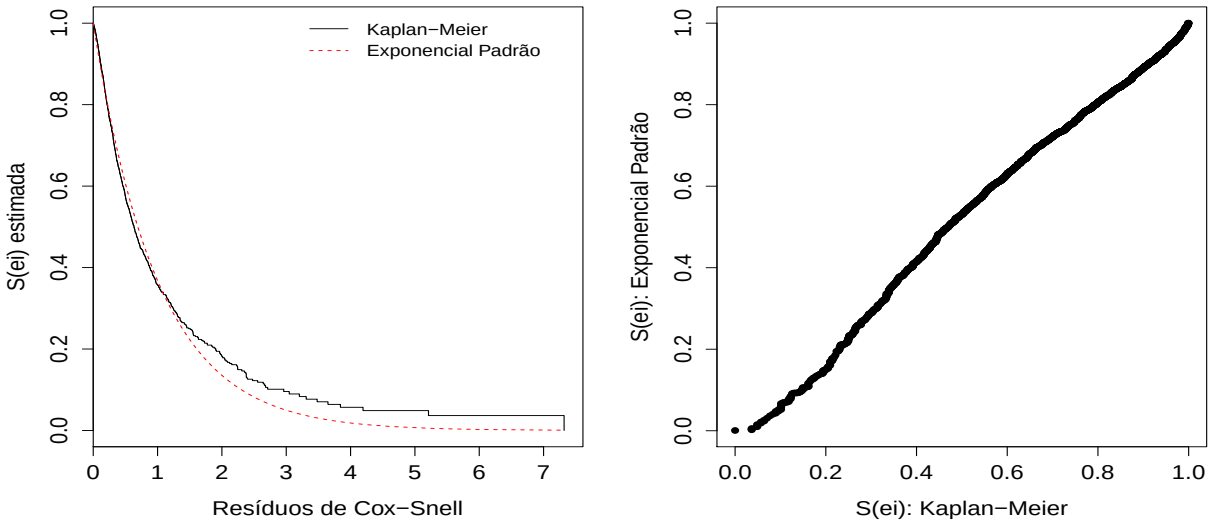
Figura 5, pode-se observar que os resíduos Deviance, que são resíduos mais simétricos, não sugerem a existência de pontos que possam ser considerados atípicos (outliers), tendo em vista que grande parte está distribuído em torno de zero.

Quanto à qualidade de predição do modelo de Cox, esta foi avaliada por meio da AUC(t). Os resultados das áreas estimadas estão na Tabela 5 e estas sugerem que o modelo apresenta qualidade de predição satisfatória e decrescente com o aumento do tempo t, o que é esperado tendo em vista que o tamanho amostral vai decrescendo no decorrer do estudo.

De modo geral, pode-se dizer com base nos resultados obtidos que o modelo de Cox apresentou ajuste global e qualidade de predição bastante razoáveis. Após a verificação da qualidade do ajuste e predição do modelo selecionado, procedeu-se a fase de interpretação das estimativas.

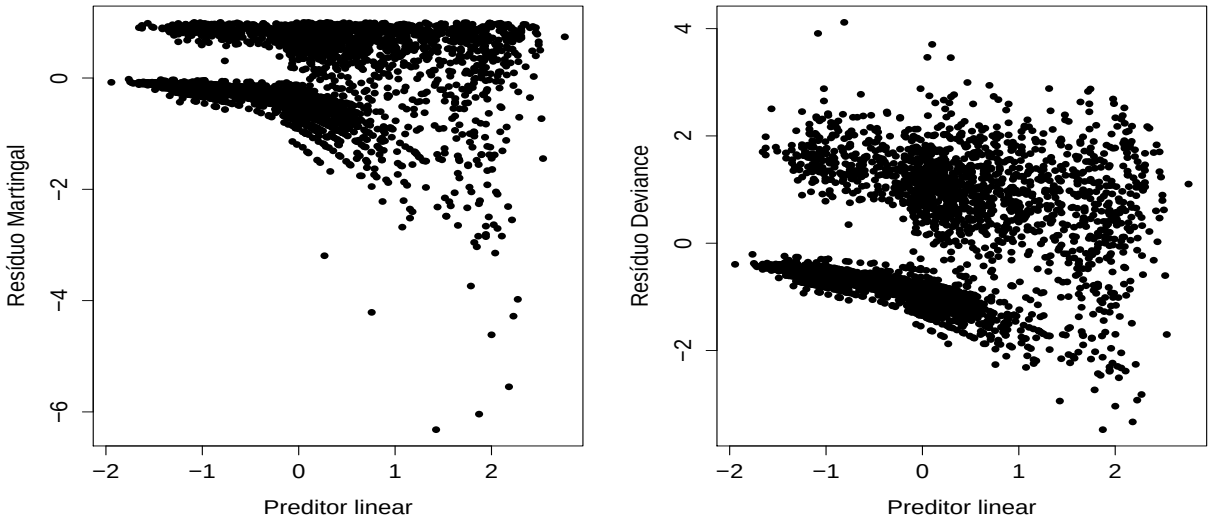
Para o modelo selecionado foram obtidas as estimativas, erros padrão e valores

Figura 4 – Sobrevivências dos resíduos de Cox-Snell do modelo de Cox ajustado, estimadas pelo método de Kaplan-Meier e pelo modelo exponencial padrão (gráfico à direita) e respectivas curvas de sobrevivências estimadas (gráfico da esquerda)



Fonte: Elaborado pelos autores (2016).

Figura 5 – Resíduos martingal e deviance versus preditor linear associados ao modelo de Cox



Fonte: Elaborado pelos autores (2016).

p associados ao teste de Wald apresentadas na Tabela 6. Nota-se, desses resultados, que todas as covariáveis apresentaram efeito significativo ($p < 0{,}05$).

A partir das estimativas obtidas (Tabela 6), é possível concluir com 95% de confiança, que a taxa de óbito não apresentou grande diferença entre os pacientes que realizaram os tratamentos: 'Cirurgia mais tratamentos coadjuvantes' e 'Hormonioterapia e

Tabela 5 – AUC(t) e erros padrão (e.p.) estimados para o modelo de Cox, obtidas pelos métodos KM (estimador de Kaplan-Meier) e NNE (estimador do vizinho mais próximo)

t	AUC (NNE)	e.p. (AUC NNE)	AUC (KM)	e.p. (AUC KM)
20	0,78572	0,01306	0,83583	0,00997
30	0,77082	0,01295	0,82203	0,00967
63	0,74737	0,01209	0,79817	0,00892
84	0,73463	0,01213	0,78233	0,00908
100	0,72399	0,01271	0,76838	0,01038
130	0,72473	0,01625	0,76562	0,01538

Fonte: Elaborado pelos autores (2016).

Tabela 6 – Resultados do ajuste do modelo de Cox selecionado para os dados de câncer de mama

Covariável	Coef	Erro - Padrão	Valor p	exp(coef)	I.C (95%)
Tratamento					
Cirurgia [a]					
Cirurgia + Coadjuvantes	0,061	0,652	0,5141	1,0634	(0,884;1,279)
Hormonioterapia e Outros	0,231	1,907	0,0566	1,2602	(0,993;1,599)
Quimioterapia e Radioterapia	0,653	4,301	<0,0001	1,9204	(1,426;2,585)
Quimioterapia	0,762	6,743	<0,0001	2,1419	(1,716;2,673)
Radioterapia	0,933	7,678	<0,0001	2,5431	(2,004;3,227)
Avaliação e Extensão da Doença					
1: In situ + Localizado [a]					
2: Extensão direta	0,667	4,559	<0,0001	1,9484	(1,463;2,595)
3: Envio por linfonodos regionais	0,973	5,331	<0,0001	2,6447	(1,8496;3,781)
4: Extensão direta com envio por linfonodos regionais	1,169	14,723	<0,0001	3,2194	(2,755;3,762)
5: Metástase	2,262	23,431	<0,0001	9,6060	(7,950;11,607)
9: Sem Informação	0,698	2,502	0,0124	2,0098	(1,163;3,472)
Anos de entrada no estudo					
de 2001 a 2005 [a]					
de 1995 a 2000	0,395	5,055	<0,0001	1,4843	(1,274;1,73)
de 1990 a 1994	0,333	4,121	<0,0001	1,3956	(1,191;1,635)
Idade	0,015	7,023	<0,0001	1,0152	(1,011;1,019)

[a] Categoria de referência

Fonte: Elaborado pelos autores (2016).

outros' em relação aos que realizaram apenas cirurgia. Para os que fizeram 'Quimioterapia e em seguida Radioterapia' e os que fizeram apenas 'Quimioterapia', a taxa estimada foi de, respectivamente, 1,92 e 2,14 vezes a dos pacientes com o tratamento igual à cirurgia. Já os pacientes que realizam apenas 'Radioterapia', apresentaram risco de óbito 2,54 vezes o dos pacientes que realizaram cirurgia, podendo tal risco variar entre 2,00 e 3,23 com 95% de confiança.

Em relação a Avaliação e Extensão da Doença (AED), observou-se que os pacientes com extensão direta da doença apresentaram taxa de óbito 1,94 vezes a dos pacientes com a doença 'In Situ'. Além disso, os pacientes classificados com AED 3 e 4 apresentaram taxa de falha de, respectivamente, 2,64 e 3,22 vezes a dos pacientes classificados com AED 1, podendo tal risco variar entre 1,85 e 3,78 para a AED 3 e 2,75 e 3,76 para a AED 4, com

95% de confiança. Os pacientes classificados na AED 5 (metástase) apresentaram taxa de óbito 9,60 vezes a dos pacientes com o câncer 'In Situ', podendo variar entre 7,95 e 11,6. Em contrapartida, há um grupo em que não se tem informações sobre a AED, sendo que este grupo apresentou taxa de falha 2 vezes a dos pacientes com AED igual a 1.

Quanto ao ano de entrada no estudo, obteve-se um aumento de aproximadamente 48% da taxa de óbito dos pacientes que iniciaram o tratamento entre 1995 e 2000 em relação aos que iniciaram entre 2001 e 2005. Verificou-se também que houve um aumento de 39% da taxa de óbito dos pacientes que iniciaram o tratamento entre 90 e 94 em relação aos que iniciaram entre 2001 e 2005. que o risco de óbito aumentou em 1,5% com o aumento de uma unidade (um ano) da Idade. Por exemplo, uma paciente com 40 anos terá um aumento de aproximadamente 30% da taxa de óbito comparada à uma paciente de 20 anos de idade.

3.3 Resultados dos Modelos Aditivos

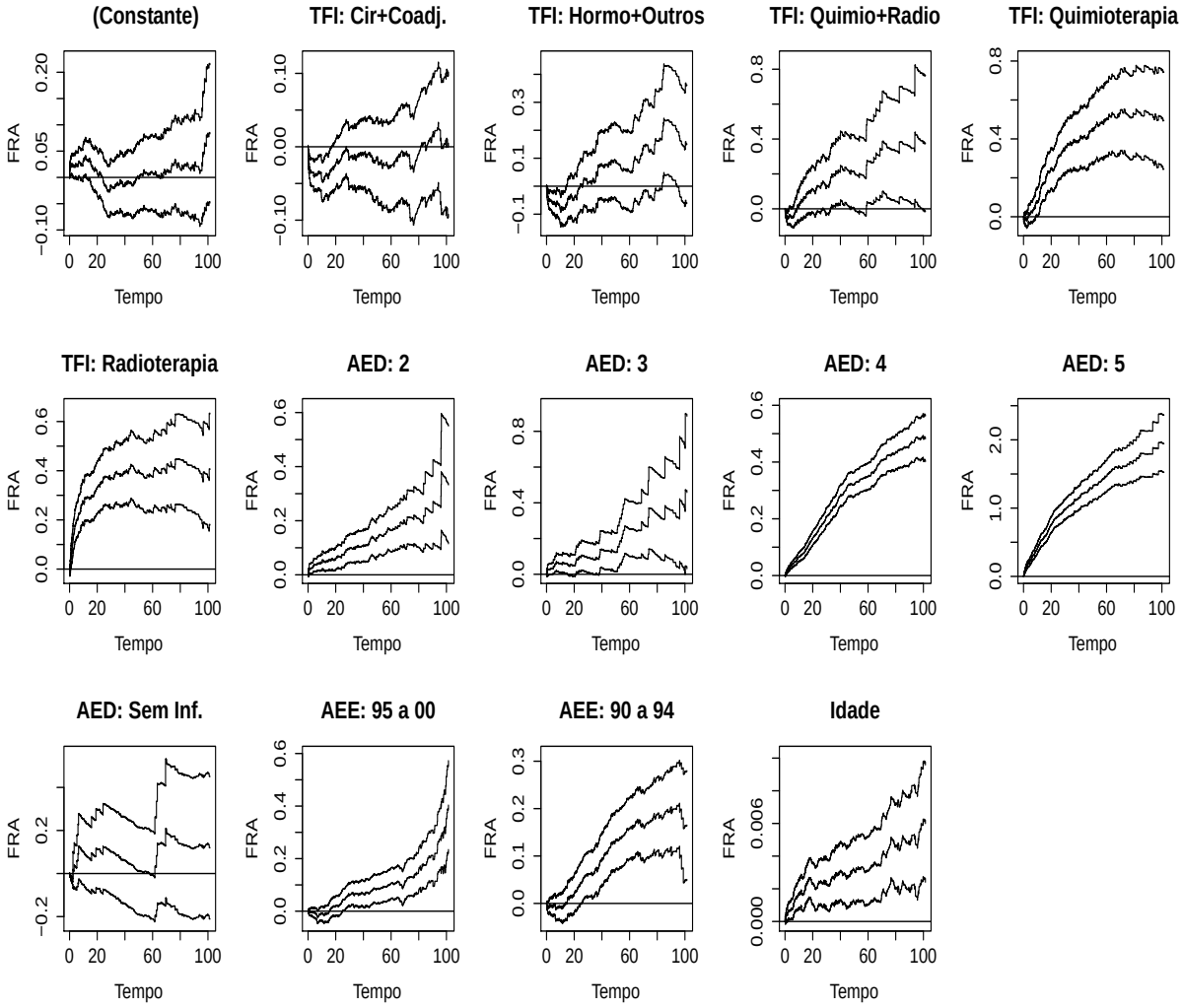
Após a análise com o modelo de regressão de Cox, o modelo aditivo de Aalen também foi cosiderado para complementar a análise dos dados de Câncer de Mama. Inicialmente foi ajustado o modelo expresso em (2.5), com as mesmas covariáveis incluídas no modelo de Cox. O tempo (t) em que estimação foi possível para esse modelo variou de 0 a 102, ou seja, $\tau = 102$.

Para verificar se alguma covariável apresenta efeito constante ao longo do tempo, foram obtidos e analisados os gráficos dos efeitos acumulados mostrados na Figura 6. A partir dessa figura, nota-se que a covariável idade apresenta efeito significativo (valor zero não pertence ao intervalo de confiança para todo t) e também apresenta efeito constante, pois é possível visualizar um comportamento linear do efeito acumulado, evidenciando que um valor constante está sendo somado a cada acréscimo de uma unidade de tempo. Para as demais covariáveis, pode-se observar efeito tempo-dependente para pelo menos uma de suas categorias (por exemplo, para as categorias: radioterapia, AED 4 e ano de entrada 90 a 94).

Dessa forma, foi ajustado um modelo de riscos aditivos semiparamétrico, expresso em (2.6), que permite parte das covariáveis com efeito dependente do tempo e parte não. O coeficiente estimado para a Idade, que foi a única covariável assumida com efeito constante ao longo do tempo, foi de 0,000078, com um erro padrão de 0,0001481. Para as covariáveis com efeito tempo-dependente, tem-se os gráficos mostrados na Figura 7, com suas bandas de 95% de confiança, que evidenciam que o efeito das covariáveis TFI, AED e Ano de Entrada no Estudo mudam com o tempo.

Em seguida, foram obtidos os resíduos de Cox-Snell apresentados na Seção 2.2.6. Analisando a Figura 8, pode-se observar evidências bem favoráveis ao modelo de riscos aditivos nos gráficos (a) e (b). Nota-se a ausência de observações atípicas no gráfico (c),

Figura 6 – Estimativas das Funções de Regressão Acumulada (FRA) e seus respectivos intervalos de 95% de confiança para o modelo aditivo de Aalen

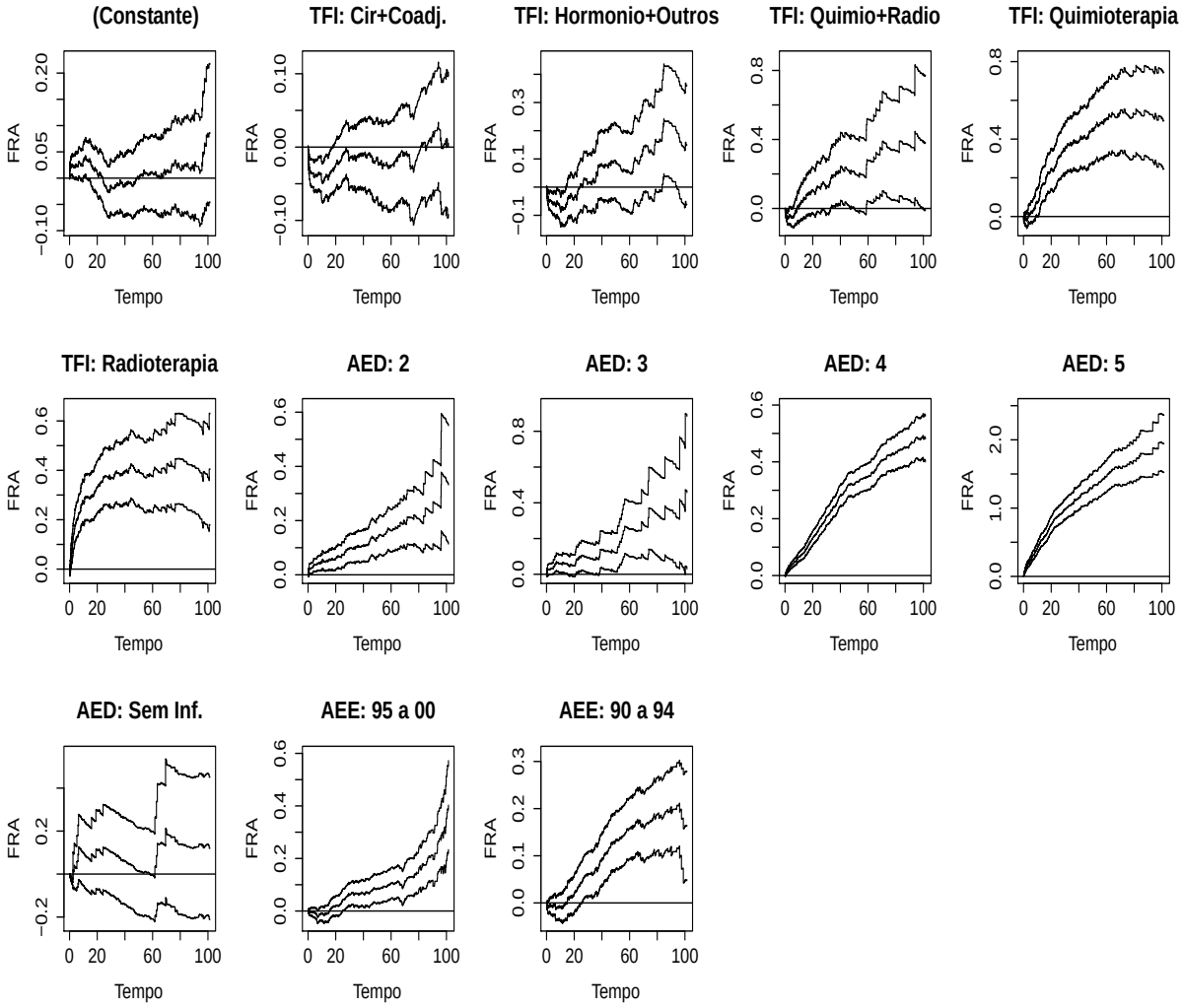


Fonte: Elaborado pelos autores (2016).

visto que os resíduos deviance estão distribuídos em torno de zero.

A partir dos resultados obtidos na Figura 7 é possível verificar no primeiro gráfico que a taxa de falha para os pacientes nas categorias de referência é muito próxima de zero em todo seu intervalo de tempo. Os pacientes que realizaram o tratamento 'Quimioterapia' ou 'Radioterapia' possuem taxa de falha elevada quando comparado a Cirurgia. Pode-se notar também que apesar de chegarem a taxas finais parecidas o efeito da covariável Radioterapia atinge uma taxa de falha elevada já no tempo 20 e depois torna-se constante enquanto o efeito da Quimioterapia se torna constante apenas no tempo 60. O grupo que realizou Cirurgia em conjunto com outros tratamentos possui uma taxa similar a categoria de referência, pois apresenta inclinação próxima de zero. Quando as pacientes em que a doença já havia se espalhado para outros órgão são comparadas com as que estavam com a doença localizada na mama, nota-se que no tempo 100 o grupo com metástase apresentou uma taxa de falha 2 vezes maior. Quanto à covariável ano de entrada no estudo, nota-se

Figura 7 – Estimativas das Funções de Regressão Acumulada (FRA) e seus respectivos intervalos de 95% de confiança para o modelo aditivo semiparamétrico

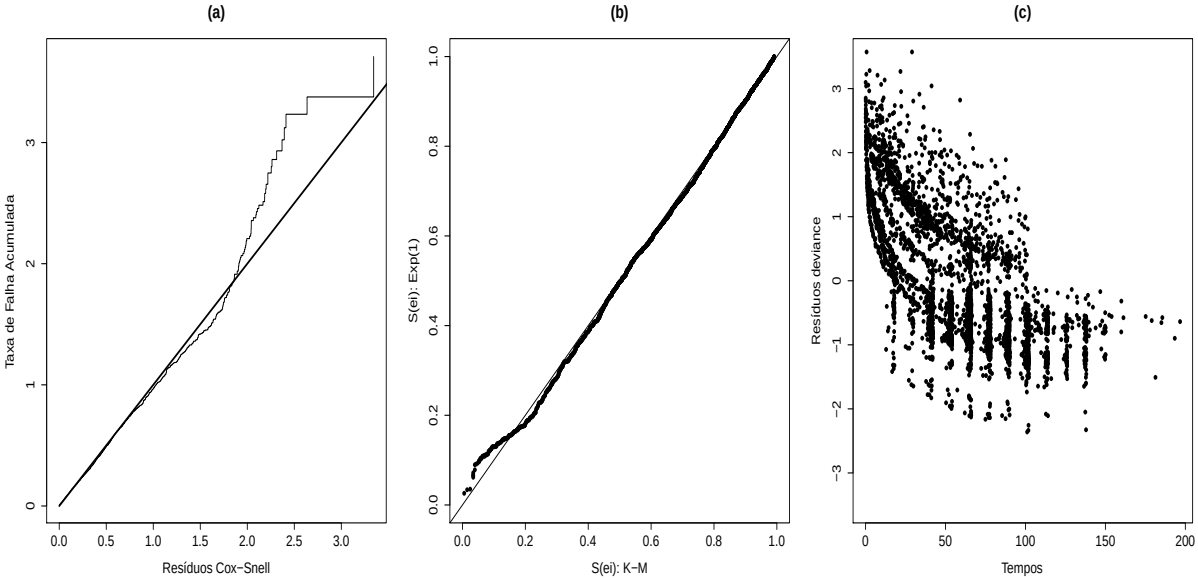


Fonte: Elaborado pelos autores (2016).

que quanto mais recente a data menor a taxa de falha, porém vale lembrar que quanto mais tarde os pacientes entraram no estudo menor o período de acompanhamento, podendo assim afetar os resultados.

Para verificar o poder de predição do modelo de riscos aditivos semiparamétrico foi calculado a AUC(t), assim como para o modelo de Cox. Os valores das áreas estimadas estão na Tabela 8, podendo-se notar que a qualidade de predição foi, assim como no modelo de Cox, bastante satisfatória.

Figura 8 – (a) resíduos $\hat{e}_i$versus taxa de falha acumulada $\hat{\Lambda}(\hat{e}_i)$, (b) pares $(\hat{S}_{KM}(\hat{e}_i), \hat{S}_{Exp}(\hat{e}_i))$ e (c) resíduos *deviance* versus tempos



Fonte: Elaborado pelos autores (2016).

Tabela 7 – AUC(t) e erros padrão (e.p.) estimados para o modelo aditivo semiparamétrico, obtidas pelos métodos KM (estimador de Kaplan-Meier) e NNE (estimador do vizinho mais próximo)

t	**AUC (NNE)**	**e.p. (AUC NNE)**	**AUC (KM)**	**e.p. (AUC KM)**
20	0,8279	0,01112	0,8298	0,01161
30	0,8143	0,00984	0,8162	0,01013
50	0,7987	0,00937	0,7994	0,00940
68	0,7936	0,009738	0,7929	0,00955
82	0,7844	0,01002	0,7830	0,01004
100	0,7589	0,01222	0,7726	0,01095

Fonte: Elaborado pelos autores (2016).

4 Considerações Finais

O câncer de mama é o tipo de doença mais comum entre as mulheres no Brasil e no mundo. Importantes avanços na abordagem do câncer de mama aconteceram nos últimos anos, e com isso verificou-se que os fatores prognósticos servem fortemente como preditor da sobrevida do paciente. Em virtude dos fatos mencionados, identificar fatores associados à sobrevida de pacientes com câncer de mama se torna de suma importância. Nesse contexto, as análises realizadas neste trabalho tiveram como objetivo identificar fatores associados à sobrevida de 3.542 pacientes com câncer de mama, tratadas em um centro médico de Curitiba no período de 1990 a 2009.

Inicialmente, foi realizado um estudo descritivo para um melhor conhecimento dos dados. Após este estudo, foi ajustado o modelo de regressão de Cox. O modelo selecionado permaneceu com as seguintes covariáveis: Tratamento feito na instituição, Avaliação e Extensão da Doença (AED), Ano de Entrada no Estudo e Idade. Foram realizadas análises gráficas a fim de verificar a suposição de riscos proporcionais e a adequação deste modelo, com ambos sendo satisfeitos. A partir dos resultados do modelo de Cox foi possível concluir, com 95% de confiança, que a taxa de falha foi maior para os pacientes que realizaram o tratamento 'Quimioterapia e em seguida Radioterapia', 'Quimioterapia' ou os que fizeram apenas 'Radioterapia', em relação aos que realizaram apenas 'Cirurgia'. Observou-se também que quanto mais espalhado o câncer estiver (AED), maior é o risco de óbito do paciente. Quanto ao ano de entrada no estudo, verificou-se que quanto mais recente tenha sido o ano, menor a taxa de falha, entretanto isso ocorre, provavelmente, devido ao período de acompanhamento ter sido menor. Também foi observado que valores mais altos da idade aumentam a taxa de falha entre os pacientes com câncer de mama.

Na sequência, foi ajustado o modelo aditivo de Aalen, ao analisar os gráficos de regressão acumulada foi observado que o efeito da covariável idade era constante ao longo do tempo. Dessa forma, uma extensão desse modelo foi utilizada, o modelo de riscos aditivos semiparamétricos, que apresentou um ajuste satisfatório aos dados utilizando as mesmas covariáveis que permaneceram no modelo de Cox. Os resultados das análises mostraram que os pacientes tratados com Radioterapia ou Quimioterapia apresentaram um risco de óbito maior em relação aos pacientes tratados com cirurgia, com esse risco sendo mais elevado nos tempos finais.Os pacientes com metástase também apresentaram risco de óbito consideravelmente maior do que os pacientes com a doença localizada na mama, independente do tempo no estudo. As covariáveis Ano de entrada no estudo e Idade não apresentaram razões de taxas de falha tão diferentes entre os grupos quanto as covariáveis 'Avaliação e extenção da doença' e 'Tratamento Feito na Instituição', mas observou-se que quanto maior a idade do paciente, maior o risco de óbito. Para o 'Ano de Entrada no Estudo' tem-se que quanto mais recente a data de entrada do paciente, menor

o risco de óbito.

Para ambos os modelos ajustados aos dados de pacientes com câncer de mama, modelo de regressão de Cox e de riscos aditivos semiparamétrico, foi possível observar que, basicamente, os modelos apresentaram as mesmas conclusões. As covariáveis TFI, AED, Ano de Entrada no Estudo e Idade foram identificadas como sendo os fatores de risco para o óbito em pacientes com câncer de mama. Concluiu-se, ainda, que o grupo de pacientes que realizaram 'Radioterapia', em relação aos demais grupos, apresentou em ambos os modelos uma taxa de falha maior. O mesmo ocorreu para o grupo de pacientes que estavam na AED classificada como 4. Vale ressaltar que o tratamento 'Cirurgia' geralmente é utilizado quando a doença ainda não está tão espalhada pelo corpo, contribuindo assim, para uma taxa de falha menor.

Analisando os gráficos apresentados na Figura 7, pode-se verificar que apesar da maioria das covariáveis terem apresentado efeito crescente ao longo do tempo, elas possuem um crescimento linear, o que sugere que a suposições do modelo de Cox não foi fortemente violada. Desta forma, pode-se concluir que o modelo de Cox é mais indicado para esse estudo do que o modelo de riscos aditivos semiparamétrico. Para comparar a eficácia dos modelos foi utilizado a AUC, descrita na Seção 2.2.7. Com base nos resultados das Tabelas 5 e 7 pode-se notar que os valores são maiores para o modelo aditivo semiparamétrico, entretanto muito similares aos valores obtidos para o modelo de Cox, o que mostra que o modelo (de Cox) além de apresentar uma interpretação mais simples também possui uma boa predição.

Ao comparar os resultados do estudo analisado neste trabalho com os estudos realizados por Santos (2013a) e Abadi et al. (2011), verificou-se que os resultados são semelhantes. Em ambos a idade foi considerada um fator que aumentava a taxa de óbito, assim como os tratamentos Quimioterapia e Radioterapia. Nesses estudos, não foi considerado a covariável Avaliação e Extensão da Doença (AED), mas sim o estadiamento. Desta forma foi observado que quanto maior a gravidade do estágio da doença maior o risco de óbito, da mesma maneira que ocorreu com a AED neste trabalho.

Referências

AALEN, O. O. A linear regression model for the analysis of life times. *Statistics in medicine*, Wiley Online Library, v. 8, n. 8, p. 907–925, 1989.

ABADI, A. et al. Comparison of aalen's additive and cox proportional hazards models for breast cancer survival: analysis of population-based data from british columbia, canada. *Asian Pacific Journal of Cancer Prevention*, Asian Pacific Organization for Cancer Prevetion, v. 12, n. 11, p. 3113–3116, 2011.

ABREU, E. de; KOIFMAN, S. Fatores prognósticos no câncer da mama feminina. *Revista Brasileira de Cancerologia*, v. 48, n. 1, p. 113–31, 2002.

AKRITAS, M. G. Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, JSTOR, p. 1299–1327, 1994.

COLLETT, D. Modelling survival data in medical research. CRC Press, 2003.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo, Brasil: Edgard Blücher, 2006. 392 p. (ABE - Projeto Fisher). ISBN 9788521203841.

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 248–275, 1968.

FERREIRA, A. Disciplina de modelos lineares - seleção de variáveis. *UERJ - Departamento de Modelagem Computacional*, 2012.

GRAMBSCH, P. M.; THERNEAU, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, v. 81, n. 3, p. 515–526, 1994.

HEAGERTY, P. J.; ZHENG, Y. Survival model predictive accuracy and roc curves. *Biometrics*, Wiley Online Library, v. 61, n. 1, p. 92–105, 2005.

HUFFER, F. W.; MCKEAGUE, I. W. Weighted least squares estimation for aalen's additive risk model. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 86, n. 413, p. 114–129, 1991.

IARC. Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. *Lyon, France: International Agency for Research on Cancer.*, 2012. Disponível em: <http://globocan.iarc.fr>.

IMAMA. *Câncer de Mama Metastático, IMAMA - INSTITUTO DA MAMA DO RS*. Rio Grande do Sul, Brasil, 2014. Acesso em 08 de Dez de 2015. Disponível em: <http://www.imama.org.br/index.php/cancer-de-mama>.

INCA. *O câncer de mama*. Rio de Janeiro, Brasil, 2015. Acesso em 15 de Nov de 2015. Disponível em: <http://www.inca.gov.br/wcm/outubro-rosa/2015/cancer-de-mama.asp>.

INCA, E. Estimativa 2014: Incidência do câncer no brasil/instituto nacional de câncer josé alencar gomes da silva, coordenação de prevenção e vigilância. 2014. *Rio de Janeiro. Disponível em http://www. inca. gov. br/estimativa/2014/. Acesso em 10 de Abril de 2016*, v. 20, 2014.

INCA, E. Estimativa 2016: Incidência de câncer no brasil/instituto nacional de câncer josé alencar gomes da silva - coordenação de prevenção e vigilância. 2014. *Rio de Janeiro. Disponível em http://www.inca.gov.br/estimativa/2016/. Acesso em 10 de Abril de 2016*, 2016.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, v. 53, n. 282, p. 457–481, 1958.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New York: John Wiley & Sons, 2011. v. 362.

Liga Paranaense de Combate ao Câncer. *Relatório Epidemiológico: 1990 a 2009. - Curitiba: LPCC, 2011. 124 p.: il.* Curitiba, Parana - BR, 2011. Acesso em 15 de Nov de 2015. Disponível em: <http://www.erastogaertner.com.br/arquivos/rhc/DuasDecadas_RHC_HEG_1990a2009.pdf>.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, v. 50, n. 3, p. 163–170, 1966.

MARTÍNEZ, M. E. et al. Differences in marital status and mortality by race/ethnicity and nativity among california cancer patients. *Cancer*, Wiley Online Library, v. 122, n. 10, p. 1570–1578, 2016.

MCKEAGUE, I. W.; SASIENI, P. D. A partly parametric additive risk model. *Biometrika*, Biometrika Trust, v. 81, n. 3, p. 501–514, 1994.

NICHOLS, E. Secret weapon in fight against lung cancer... is being married: Singles less likely to survive after treatment. *Daily Mail - Health*, 2012.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

RAMINELLI, J. A. *Métodos de adequação e diagnóstico em modelos de sobrevivência dinâmicos*. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz" - Piracicaba, 2015.

SANTOS, R. d. S. *Sobrevivência de mulheres com diagnóstico de Câncer de Mama no município do Rio de Janeiro*. Dissertação (Mestrado) — Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro, 2013, 2013.

SANTOS, T. M. dos. *Avaliação do desempenho de modelos preditivos no contexto de análise de sobrevivência*. Tese (Doutorado) — Instituto de Matemática e Estatústica da Universidade de São Paulo, 2013.

SCHOENFELD, D. Partial residuals for the proportional hazards regression model. *Biometrika*, v. 69, n. 1, p. 239–241, 1982.

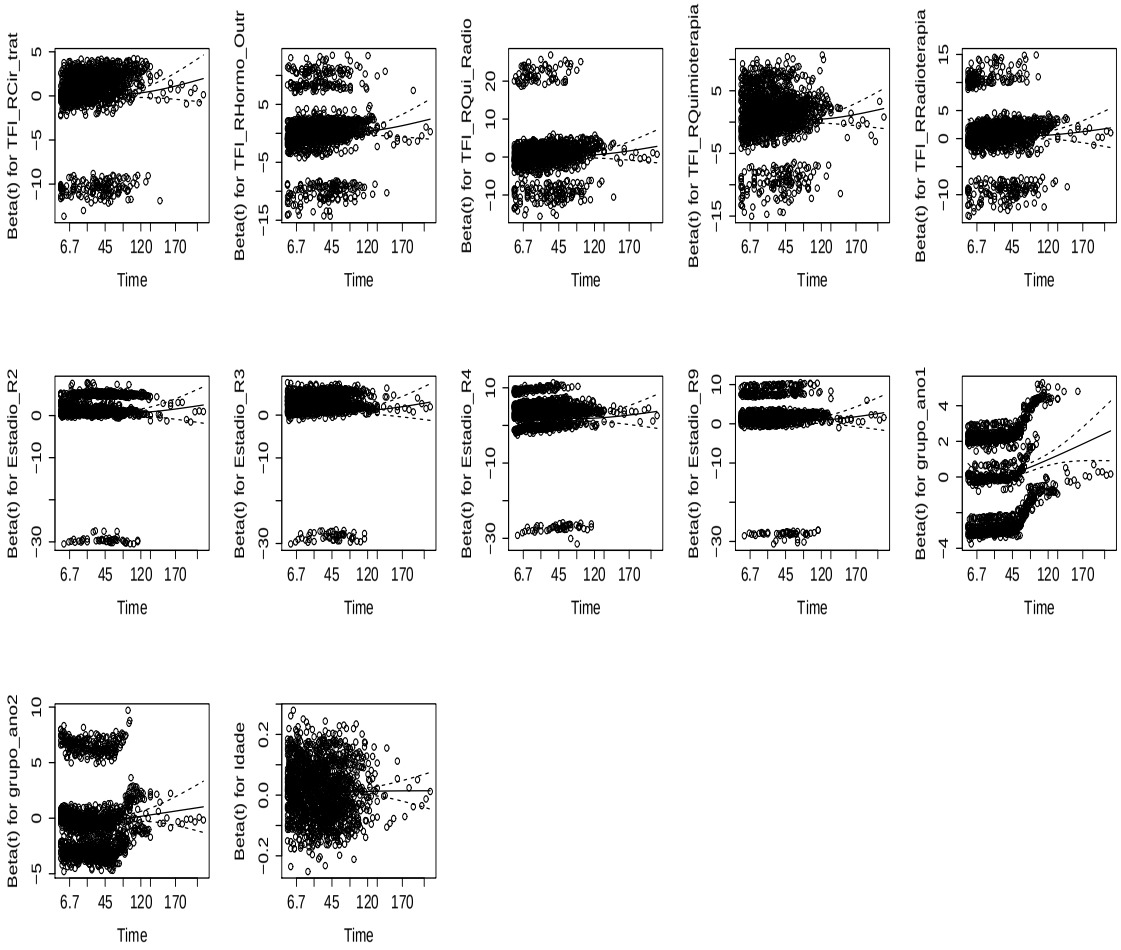
Apêndices

APÊNDICE A - Frequências absolutas e respectivos percentuais de pacientes, falhas e censuras para as covariáveis não inseridas nos modelos ajustados

Covariável	Categoria	N	N%	Censura	Falha	Cens. %	Falha%
Tumor	Apenas um Tumor	3499	99%	2015	1484	58%	42%
	Mais de um Tumor	40	1%	17	23	43%	58%
	Sem Informação	3	0%	0	3	0%	100%
Topografia	Neoplasia maligna da mama, não especificada	2200	62%	1142	1058	52%	48%
	Neoplasia maligna do quadrante superior externo da mama	707	20%	500	207	71%	29%
	Outros	635	18%	390	245	61%	39%
Morfologia	Carcinoma ductal infiltrante	3048	86%	1741	1307	57%	43%
	Carcinoma lobular	92	3%	65	27	71%	29%
	Outros	402	11%	226	176	56%	44%
GD	Bem Diferenciado	117	3%	98	19	84%	16%
	Moderadamente Diferenciado	947	27%	617	330	65%	35%
	Pouco Diferenciado	337	10%	170	167	50%	50%
	Indiferenciado	1	0%	0	1	0%	100%
	Sem Informação	2140	60%	1147	993	54%	46%

Fonte: Elaborado pelos autores (2016)

APÊNDICE B - Gráficos $\log(\hat{\Lambda}_{0_j}(t))$ *versus* os tempos



Fonte: Elaborado pelos autores (2016).

APÊNDICE C - Principais códigos em R utilizados para gerar os resultados do modelo de regressão de Cox

```
require(survival)

################### Kaplan-Meier Geral ####################
ekm <- survfit(Surv(dados$Tempo_R, dados$Censura) ~ 1);ekm
plot(ekm, xlab="Tempo (em semanas)",ylab="S(t) estimada",
     mark.time = FALSE, main = "Estimador de Kaplan - Meier Geral")

################### Modelo Final de Cox  ##################
mcoxfinal <- coxph(Surv(Tempo_R,Censura)~TFI_R+AED_R+grupo_ano_i+Idade,
                   data=dados,x = T, method="breslow")
summary(mcoxfinal)        # Estimativas
cox.zph(mcoxfinal)        # Resíduos padronizados de Schoenfeld

############ Avaliação da qualidade do ajuste  ############
#Figura 4
par(mfrow=c(1,2))
rm<-resid(mcoxfinal,type="martingale")
rc<-dados$Censura - rm
ekm <- survfit(Surv(rc, dados$Censura) ~ 1

plot(ekm, xlab="Resíduos de Cox-Snell",ylab="S(ei) estimada",
     mark.time = FALSE,conf.int=F)
rc<-sort(rc)
exp1<- exp(-rc)
lines(rc,exp1,lty=2, col= 2)
legend("topright",lty=c(1,2), col= c(1,2),c("Kaplan-Meier",
"Exponencial Padrão"),lwd=1,bty="n",cex=0.8)
st<-ekm$surv
t<-ekm$time
sexp1<-exp(-t)
plot(st,sexp1,xlab="S(ei): Kaplan-Meier",
 ylab= "S(ei): Exponencial Padrão",pch=16)

#Figura 5
par(mfrow=c(1,2))
rd<-resid(mcoxfinal,type="deviance")
```

```
rm<-resid(mcoxfinal,type="martingale")
pl<-mcoxfinal$linear.predictors
plot(pl,rm, xlab="Preditor linear", ylab="Resíduo Martingal", pch=16)
plot(pl,rd,  xlab="Preditor linear", ylab="Resíduo Deviance" , pch=16)

############################ AUC ############################
# Obtido com o auxílio dos apêndices da tese: Raminelli, 2015
```

APÊNDICE D - Principais códigos em R utilizados para gerar os resultados do modelo de regressão aditivo de Aalen e semiparamétrico

```
##### Pacotes necessários
library(timereg)
library(survivalROC)

##### Ordenando em função do Tempo
i<-order(dados$Tempo_R)
dados<-dados[i,]

##### Contornando tempos empatados
set.seed(157)
n1<-dim(dados)[1]
ei<-rnorm(n1,0,0.00001)
tempo_empate<- dados$Tempo_R + ei
dados$Tempo_R <-tempo_empate

##### Centrando a idade na média
Idade2<-dados$Idade-mean(dados$Idade)

################### Modelo aditivo de Aalen  ##################
mod2<- aalen(Surv(Tempo_R, Censura)~TFI_R+AED_R+grupo_ano_i+Idade2,
             residuals=1,max.time=102,data=dados)
#### Figura 6
par(mfrow=c(3,5))
plot(mod2, xlab="Tempo",ylab="FRA")
layout(1)

################### Modelo aditivo semiparamétrico  ##################
mod2.1<- aalen(Surv(Tempo_R, Censura)~TFI_R+AED_R+grupo_ano_i+const(Idade2),
             residuals=1,max.time=102,data=dados)
#### Figura 7
par(mfrow=c(3,5))
plot(mod2.1, xlab="Tempo",ylab="FRA")
layout(1)

############################ AUC ############################
# Obtido com o auxílio dos apêndices da tese: Raminelli, 2015
```