

Modelos de fragilidade aplicados a análise de fatores contribuintes na morte de pacientes portadores de leucemia**Models of fragility applied to analysis of contributing factors in the death of patients with leukemia**

DOI:10.34117/bjdv6n8-046

Recebimento dos originais:08/07/2020

Aceitação para publicação:07/08/2020

Manoel Joaquim Isidro

Bacharel em Estatística pela Universidade Estadual da Paraíba

Instituição: Universidade Estadual da Paraíba

Endereço: Avenida Juvêncio Arruda, 240 A202 - Novo bodocongó, Campina Grande – PB, Brasil

E-mail: isidro.mj02@gmail.com

Roseane de Alcântara Isidro

Bacharel em Estatística pela Universidade Estadual da Paraíba

Instituição: Universidade Estadual da Paraíba

Endereço: Avenida Juvêncio Arruda, 240 A202 - Novo bodocongó, Campina Grande – PB, Brasil

E-mail: roseanealcantara96@gmail.com

Alisson de Lima Brito

Mestrando em Estatística e Experimentação Agropecuária

Universidade Federal de Lavras, Departamento de Ciências Exatas, Setor de Estatística e Experimentação.

Endereço: Avenida Central UFLA, Aquecida Sol-37200000 - Lavras, MG - Brasil

E-mail: alisson.limab95@gmail.com

Tiago Almeida de Oliveira

Doutor em Estatística e Experimentação Agronômica - ESALQ/USP

Instituição: Universidade Estadual da Paraíba

Endereço: Avenida Juvêncio Arruda, 240 A202 - Novo bodocongó, Campina Grande – PB, Brasil

E-mail: tadolive@servidor.uepb.edu.br

RESUMO

A análise de sobrevivência é uma das áreas da Estatística mais utilizadas na medicina, dada sua utilidade em estimar o risco dos pacientes curar-se ou morrer ao longo do estudo. Dentre as ferramentas da análise de sobrevivência para se avaliar a influência de covariáveis no risco dos indivíduos sofrerem um evento de interesse, o modelo de Cox é de longe, o mais utilizado. Apesar de apresentar adequação em muitos problemas práticos, o modelo de Cox não considera a presença de uma variável aleatória não observável que denota o grau de heterogeneidade presente em cada indivíduo, o que pode levar a diferentes taxas de falha, uma vez que uma característica genética por exemplo pode ser favorável ou desfavorável a ocorrência do evento. Neste contexto, um modelo que pode ser utilizado para considerar a presença dessa heterogeneidade é o modelo de fragilidade.

Diante disto, o presente estudo objetivou-se em identificar os principais fatores contribuintes para a morte de pacientes portadores de Câncer. Os dados foram obtidos do Institute for Health & Society da universidade de Wisconsin (Medical College of Wisconsin), o banco de dados é constituído de um total de 137 pacientes (38 LLA, 99 LMA) os quais apresentam uma distinção entre os tipos de câncer, leucemia mieloide aguda (LMA) de auto e baixo risco e leucemia linfoblástica aguda (LLA), os pacientes foram tratados em quatro hospitais. Na realização deste trabalho foram estimadas as curvas de sobrevivência por meio de estimador não paramétrico de Kaplan-Meier para os 3 grupos de pacientes com diferentes tipos de leucemia, em seguida o teste log-rank foi utilizado para investigar se existe diferença significativa entre as curvas. Os modelos Exponencial, Weibull e Log-normal foram ajustados para avaliar a influência das covariáveis no tempo de sobrevivência dos pacientes. Por fim, o modelo de Cox e os modelos de fragilidades Gama e Log-normal foram usados para modelar o risco de morte dos pacientes e identificar os fatores contribuintes para o desfecho. O estimador de Kaplan-Meier e o teste log-rank revelaram diferença significativa nas curvas de sobrevivência nos diferentes tipos de câncer. Entre os modelos ajustados, o modelo de fragilidade apresentou melhor adequabilidade, com resultados mais fidedignos, sendo este, o modelo a ser utilizado para as inferências.

Palavras-chave: Câncer, Análise de sobrevivência, Modelos de fragilidades.

ABSTRACT

Survival analysis is one of the most widely used areas of statistics in medicine, given its usefulness in estimating the risk of patients curing or dying throughout the study. Among the tools of survival analysis to assess the influence of covariates on the risk of individuals suffering an event of interest, the Cox model is by far the most used. Although the Cox model is suitable for many practical problems, it does not consider the presence of an unobservable random variable that denotes the degree of heterogeneity present in each individual, which can lead to different failure rates, since a genetic characteristic, for example, may be favorable or unfavorable for the event to occur. In this context, a model that can be used to consider the presence of such heterogeneity is the fragility model. In view of this, the present study aimed to identify the main factors contributing to the death of patients with Cancer. The data were obtained from the Institute for Health & Society of the University of Wisconsin (Medical College of Wisconsin), the database consists of a total of 137 patients (38 LLA, 99 LMA) who present a distinction between types of cancer, auto and low risk acute myeloid leukemia (AML) and acute lymphoblastic leukemia (APL), patients were treated in four hospitals. In this study, survival curves were estimated using a Kaplan-Meier non-parametric estimator for the 3 groups of patients with different types of leukemia, followed by the log-rank test to investigate if there is a significant difference between the curves. The Exponential, Weibull and Log-normal models were adjusted to evaluate the influence of covariates on the survival time of patients. Finally, the Cox model and the Gamma and Log-normal fragility models were used to model the risk of death of patients and to identify factors contributing to the outcome. Kaplan-Meier's estimator and log-rank test revealed significant differences in survival curves for different types of cancer. Among the adjusted models, the fragility model showed better suitability, with more reliable results, being this the model to be used for inferences.

Keywords: Cancer, Survival analysis, Fragility models.

1 INTRODUÇÃO

A análise de sobrevivência é formada por um conjunto de técnicas estatísticas para analisar dados, que consistem nos tempos até a ocorrência de um evento de interesse, comumente conhecido como tempo de sobrevivência ou tempo de falha. São exemplos de casos de análise de sobrevivência o tempo até que um aluno possa concluir sua graduação assim como a morte do paciente após o diagnóstico de uma certa doença. Um fato que caracteriza dados de sobrevivência é a possibilidade da presença de censura, que é a ocorrência da observação parcial da resposta de interesse. Neste estudo, a variável resposta foi caracterizada pelo tempo em dias até a morte do paciente ou até o término do acompanhamento.

Leucemia Linfoblástica Aguda (LLA) é um câncer que se origina de um grupo de células precursoras dos linfócitos. Os linfócitos são glóbulos brancos que defendem o corpo contra infecções. A medula óssea cria inúmeras células subdesenvolvidas conhecidas como blastos. A LLA é resultado de um dano genético adquirido (não herdado) no DNA de um grupo de células (glóbulos brancos) na medula óssea. As células doentes substituem a medula óssea normal e os efeitos são o crescimento incontrolável e o acúmulo de linfoblasto (linfócitos imaturos) que perdem a capacidade de funcionar como células sanguíneas normais, levando a um bloqueio ou diminuição na produção de glóbulos vermelhos, plaquetas e glóbulos brancos na medula óssea. Após o diagnóstico, o tratamento da LLA é dividido em duas partes: terapia de indução e terapia pós-indução ou transplante de medula óssea. A quimioterapia utiliza medicamentos anticancerígenos para destruir as células tumorais, dependendo do tipo e do estado da leucemia, e pode ser utilizada sozinha ou combinada com a radioterapia (PEDROSA; LINS, 2002).

Depois de um transplante de medula óssea a recuperação é um processo complexo e lento. O prognóstico para recuperação pode depender de fatores de risco conhecidos, no momento do transplante, como a idade do paciente e/ou idade do doador e o sexo, a fase inicial da doença, o tempo a partir do diagnóstico para o transplante, etc. O prognóstico final pode mudar após o transplante do paciente, pois a ocorrência de eventos em momentos aleatórios durante a recuperação, como o desenvolvimento de doenças aguda ou crônica do enxerto-versus-hospedeiro (GVHD). O transplante pode ser considerado uma falha quando houver a recaída da doença (recaída de leucemia no paciente), ou quando o paciente morre enquanto está em remissão da doença (morte relacionada com o tratamento). O Transplante de Medula Óssea configura-se como um grande problema de saúde pública tanto nos países desenvolvidos como nos países em desenvolvimento. Há muitas causas diferentes para a doença, que variam de radiação de produtos químicos a causas genéticas,

tem o objetivo de destruir as células leucêmicas, para que a medula óssea volte a produzir células normais.

Diante do exposto o uso de técnicas estatísticas, mais especificamente da análise de sobrevivência pode contribuir para explicar os fatores de risco para o tempo de sobrevivência de pacientes transplantados e o uso das diversas técnicas da análise de sobrevivência permite a escolha do melhor método para o tipo de estudo em questão.

Objetivo deste trabalho é identificar os principais fatores contribuintes na morte de pacientes portadores de leucemia linfoblástica e mieloide, após o transplante de medula óssea. Dentro do qual métodos clássicos e introdutórios da análise de sobrevivência foram utilizados, como também a avaliação do efeito da fragilidade nos pacientes por meio dos modelos de fragilidade Gama e Log-normal.

2 MATERIAL E MÉTODOS

A análise de sobrevivência é uma área da estatística com aplicabilidade nas mais variadas áreas do conhecimento, e sua denominação pode variar conforme o campo de aplicação, podendo ser conhecida também como análise de confiabilidade, análise do tempo de vida, análise de duração, análise do tempo de falha, análise de transição, análise de eventos históricos, entre outros. Todavia, a denominação não altera os métodos estatísticos utilizados.

Apesar disso, é nas ciências da saúde que verse maior aplicação dos métodos da análise de sobrevivência, devido a versatilidade das ferramentas estatísticas empregadas nesse tipo de análise e ao fato de incorporar na análise observações incompletas dos indivíduos em estudo, o que chamamos de censura, observações estas que, provavelmente seriam descartadas ao se utilizar outra metodologia. Nesse tipo de análise, a principal variável de interesse é sempre o tempo até a ocorrência de um determinado evento (também conhecido como falha), comumente denotada por T . A variável tempo de falha T é representada por algumas funções matemáticas importantes, tais como, a função densidade de probabilidade $f(t)$, a função de sobrevivência $S(t)$, a função de risco $h(t)$ e a função de risco acumulado $H(t)$, as quais serão abordadas a seguir.

2.1 FUNÇÕES DE INTERESSE

Diniz e Louzada (2012) definem $f(t)$ como o limite da probabilidade da falha ocorrer em um intervalo de tempo $[t, t + \Delta t]$, em que Δt representa uma variação no tempo. Deste modo, a função densidade de probabilidade é matematicamente representada por

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t},$$

em que, $f(t) \geq 0$, para todo t , e tem área abaixo da curva igual a 1.

A função de sobrevivência é sem dúvidas uma das funções mais importantes em análise de sobrevivência. Essa função pode ser entendida como a probabilidade de um indivíduo sobreviver acima de um tempo t , dado que este ainda não tenha sofrido o evento de interesse, ou seja, não tenha falhado. A função de sobrevivência pode ser obtida como complementar da função de distribuição acumulada como segue,

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \quad (2.1)$$

A função $S(t)$ tem valor 1 no tempo zero e decresce ou permanece constante ao longo do tempo (MOORE, 2016).

De acordo com Colosimo e Giolo (2006) a função de risco ou taxa de falha é bastante útil para descrever a distribuição do tempo de vida de pacientes ao longo de um período de estudo, descrevendo a forma como a taxa de falha muda ao longo do tempo. Essa função pode ser entendida como a taxa de falha instantânea no tempo t , condicionada à sobrevivência até o tempo t e matematicamente definida como,

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T > t)}{\Delta t}.$$

A função de risco acumulado ou taxa de falha acumulada é obtida através da integral no intervalo de zero a t da função de risco, sendo portanto, definida como segue

$$H(t) = \int_0^t h(u) du.$$

Esta função não tem interpretação direta, porém apresenta um estimador não paramétrico com ótimas propriedades, ao passo que, $h(t)$ é difícil de ser estimada. Deste modo, a função de risco é então obtida através da derivada da função de risco acumulado.

2.2 MÉTODOS NÃO PARAMÉTRICOS

As funções descritas na subseção anterior não consideram a presença de informações incompletas na variável resposta, ou seja, não permitem a presença de censura nos dados amostrais. Nesse contexto, Kaplan e Meier (1958) propuseram um estimador para obter a função de sobrevivência por meio de métodos não paramétricos, de forma que, não há necessidade de se conhecer uma distribuição probabilística para se obter a função $S(t)$. Este estimador é usualmente

chamado de estimador Kaplan-Meier ou estimador limite-produto, e é atualmente o método mais utilizado na literatura para estimar a função de sobrevivência na presença de censura. O estimador de Kaplan-Meier pode ser então obtido da seguinte maneira,

$$\hat{S}(t) = \left(1 - \frac{d_j}{n_j}\right),$$

em que, $t_j, j = 1, \dots, k$, são os k tempos distintos e ordenados de falha, d_j , representa o número de falhas no tempo t_j e n_j corresponde ao número de indivíduos sob risco no tempo t_j .

De acordo com Karin e Islam (2019), um estimador não paramétrico para a função de distribuição acumulada $F(t)$ pode ser obtido através do estimador de Kaplan-Meier, usando o resultado na equação (2.1) como segue,

$$\hat{F}(t) = 1 - \hat{S}(t), \quad t \geq 0.$$

É muito comum em estudos de sobrevivência querermos avaliar se existe diferença significativa entre as curvas de sobrevivência de determinados grupos em estudo, nesse contexto, muitos testes foram propostos na literatura para solucionar esse problema, um deles, e atualmente o teste mais utilizado para comparação entre curvas de sobrevivência é o teste log-rank.

De acordo com Emura e Chen (2018), o teste log-rank é essencialmente utilizado para testar a hipótese de igualdade entre as taxas de falha de dois grupos, sob a hipótese nula $H_0: h(t|x_i = 0) = h(t|x_i = 1), t \geq 0$, em que $x_i, i = 0, 1$ representam dois grupos distintos. O que corresponde a testar a hipótese nula $H_0: S(t, x_i = 0) = S(t, x_i = 1)$, devido a relação entre as funções de sobrevivência e de risco. Colosimo e Giolo (2006) apresentam a seguinte estatística para testar as hipóteses acima.

$$T = \frac{[\sum_{j=1}^k (d_{2j} - w_{2j})]^2}{\sum_{j=1}^k (V_j)^2},$$

em que d_{2j} representa a falha dos indivíduos do grupo 2 no tempo j , w_{2j} e V_j representam a média e a variância de d_{2j} respectivamente. Sob a hipótese nula $H_0: S(t, x_i = 0) = S(t, x_i = 1)$, para todo t no tempo de acompanhamento, T tem distribuição qui-quadrado com 1 grau de liberdade para grandes amostras, ou seja, $T \sim \chi_1^2$.

2.3 MODELOS PARAMÉTRICOS PARA O TEMPO DE FALHA

Segundo Colosimo e Giolo (2006), entre os diversos modelos probabilísticos propostos na literatura para descrever o tempo até a falha, alguns deles ocupam posições de maior relevância pela comprovada adequação a diversos problemas práticos, dentre estes, podemos citar os modelos Exponencial, Weibull e Log-normal, os quais serão abordados a seguir.

O modelo Exponencial é um dos modelos probabilísticos mais simples que pode ser utilizado para modelar o tempo de falha em estudos de sobrevivência, esse modelo, assume que o risco de sofrer o evento é constante ao longo do tempo, por esse motivo, o modelo Exponencial tem perdido competitividade, apesar de se adequar bem a determinadas situações. Se uma variável aleatória T , que representa o tempo de falha tem distribuição exponencial, então sua função densidade de probabilidade pode ser descrita como segue

$$f(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}}, \quad t \geq 0,$$

em que, λ é o tempo médio de vida. Neste caso, as funções de sobrevivência e taxa de falha são obtidas respectivamente por

$$S(t) = e^{-\frac{t}{\lambda}} \quad \text{e} \quad h(t) = \frac{1}{\lambda}.$$

O modelo Weibull é um dos modelos probabilísticos mais utilizados em análise de sobrevivência. A distribuição Weibull foi proposta por Weibull (1939), deste então, tem sido utilizada para solucionar diversos problemas práticos e tem se mostrado bastante eficaz na solução desses problemas, sendo um dos modelos probabilísticos que com alto grau de adequação. Colosimo e Giolo (2006) afirmam que sua popularidade deve-se ao fato do modelo apresentar uma grande variedade de formas e a propriedade de ter uma função taxa de falha monótona. Os autores, apresentam o modelo da seguinte forma, seja T uma variável aleatória representando o tempo de falha com distribuição Weibull, sua função densidade de probabilidade pode ser dada por

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \quad t \geq 0.$$

em que, γ é o parâmetro de forma e α , o de escala, sendo ambos positivos. As funções de sobrevivência e de risco são dessa forma, obtidas respectivamente como segue

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\} \quad \text{e} \quad h(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1},$$

para $t \geq 0$, α e $\gamma > 0$. A distribuição Weibull pode ser vista como uma generalização da distribuição exponencial, sendo esta obtida quando $\gamma = 1$.

De acordo com Lee e Wang (2003), uma variável aleatória T tem distribuição Log-normal se o logaritmo de T for normalmente distribuído. Dessa forma, se $\log T \sim N(\mu, \sigma^2)$, então, $T \sim \log N(\mu, \sigma^2)$ e sua função densidade de probabilidade é apresentada da seguinte maneira

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}, \quad (2.2)$$

em que, $\mu \in \mathbb{R}$ e $\sigma^2 > 0$.

Colsimo e Giolo (2006) afirmam que o modelo Log-normal é bastante utilizado para descrever situações clínicas, como o tempo de vida de pacientes com leucemia, ainda segundo os autores, as funções de sobrevivência e de risco não apresentam uma forma analítica fechada e podem ser obtidas como segue

$$S(t) = \Phi\left(\frac{\mu - \log(t)}{\sigma}\right) \quad \text{e} \quad h(t) = \frac{f(t)}{S(t)},$$

em que, $\Phi(\cdot)$ é a função de distribuição acumulada de uma normal padrão.

2.4 O MODELO SEMI-PARAMÉTRICO DE COX

O modelo de riscos proporcionais de Cox é sem dúvidas uma das ferramentas da análise de sobrevivência mais utilizadas para se avaliar a influência de covariáveis no risco dos indivíduos sofrer o evento de interesse. Segundo Harrell Jr (2015), o modelo de Cox é o modelo de regressão mais popular no tratamento e análise de dados de sobrevivência, isso, graças a flexibilidade que o modelo oferece, devido ao fato de não fazer suposições para forma da curva de risco dos indivíduos em estudo.

Dado um vetor de covariáveis $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{pj})$, o modelo assume que os preditores agem de forma multiplicativa na função de risco. Dessa forma o modelo fica dado como segue

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

em que, $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_p)$ é o vetor de parâmetros (a ser estimado) associado ao vetor de covariáveis \mathbf{x}_j e $h_0(t)$ é a chamada de função de risco de base, uma vez que, se desconsiderarmos o efeito das covariáveis, isto é, se tivermos $\mathbf{x}_j = \mathbf{0}$, temos que, $h(t) = h_0(t)$, o que significa que todos os indivíduos possui o mesmo risco de sofrer o evento. Para Bogaerts, Komarek e Lesaffre

(2017), o modelo de Cox, é de longe, o modelo mais popular para lidar com dados de sobrevivência em aplicações médicas desde a sua introdução por Cox (1972).

2.5 O MODELO DE FRAGILIDADE

Os métodos descritos nas seções anteriores assumem que os tempos de sobrevivência dos indivíduos em estudo são independentes, todavia, existem situações para as quais essa suposição não é válida, como por exemplo, se tivermos observando indivíduos até a ocorrência de um determinado evento, e um grupo de indivíduos apresentam certa semelhança, é razoável pensarmos que os tempos de sobrevivência dos indivíduos pertencentes a esse grupo também apresentem semelhanças, nesse caso, há a suposição de que haja correlação entre os tempos de sobrevivência destes indivíduos.

No contexto multivariado, o interesse é levar em consideração essa correlação entre os tempos de sobrevivência de indivíduos de determinados grupos, para isso alguns modelos foram propostos na literatura, tal como o modelo de fragilidade compartilhada, o qual não será abordado neste trabalho. No contexto univariado, a ideia é modelar a heterogeneidade presente entre os indivíduos em estudo. Nessas situações, podemos considerar que existe uma variável aleatória latente chamada de fragilidade e usualmente denotada por Z que denota a susceptibilidade de determinados indivíduos falharem. Essa variável é incorporada no modelo de Cox de forma multiplicativa sob a função de risco de base, tal como definido por Wienke (2010), dessa forma, a função de risco para os indivíduos em estudo fica condicionada ao efeito da fragilidade. Assim sendo, o modelo de fragilidade pode ser obtido da seguinte forma

$$h(t; \mathbf{x}_j; z_j) = z_j h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_j). \quad (2.3)$$

Ainda segundo Wienke (2010), diferentes distribuições de probabilidade podem ser usadas para modelar o efeito da fragilidade, e ainda, a variância da distribuição da fragilidade denota o grau de heterogeneidade presente nas observações. Neste estudo, duas distribuições foram usadas para avaliar o efeito da fragilidade, a distribuição Gama, e a distribuição Log-normal, cuja função densidade de probabilidade é apresentada na equação (2.2).

De acordo com Colosimo e Giolo (2006) a distribuição Gama, é a distribuição de probabilidade mais utilizada para se modelar o efeito da fragilidade Z no modelo (2.3). Segundo Duchateau e Janssen (2008), é comum que a distribuição assumida para se modelar o efeito aleatório (fragilidade) tem média 1 e variância desconhecida, comumente representada na literatura pelo parâmetro θ . Neste caso, a distribuição Gama sofre uma reparametrização, de maneira que, o

parâmetro de forma é igual ao parâmetro de escala. Portanto, dado que Z tem distribuição Gama uniparamétrica com variância θ , ou seja, $Z \sim G\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$ sua função densidade de probabilidade é dada por

$$f(z) = \frac{z^{\frac{1}{\theta}-1} \exp\left(-\frac{z}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) \theta^{\frac{1}{\theta}}}.$$

Em estudos clínicos, considerar o efeito da fragilidade no modelo de regressão é de fato importante, uma vez que cada paciente pode apresentar diferentes resultados que podem variar de acordo com a genética do indivíduo, dessa forma, o modelo de fragilidade é uma ferramenta muito importante e que pode apresentar melhorias nos resultados obtidos pelo modelo.

2.6 SELEÇÃO DE MODELOS

Em estatística, muitos métodos de comparação de modelos podem ser utilizados para identificar o modelo mais adequado para o problema em questão, um dos métodos bastante recorrente em trabalhos científicos é o Critério de Informação de Akaike (AIC), que consiste em selecionar o modelo mais parcimonioso possível, ou seja, com menor número de parâmetros possível. Desta forma o critério penaliza o modelo com maior número de parâmetros, de maneira que o modelo selecionado explique o maior percentual da variabilidade presente nas observações. A ideia básica é identificar o modelo tal que a quantidade abaixo seja mínima,

$$AIC = -2\ell(\hat{\beta}) + 2\kappa,$$

em que, $\ell(\hat{\beta})$ é a log-verossimilhança do modelo e κ é o número de parâmetros. Para Moore (2016), o cálculo do AIC é uma das melhores maneiras de avaliar modelos estatísticos. E segundo Klein e Moeschberger (2005) a inclusão de variáveis no modelo causa uma diminuição no valor de AIC, todavia, em certo ponto o critério passa a aumentar indicando que a inclusão de determinadas variáveis não contribui para as estimativas dos parâmetros.

Outra maneira de realizar a seleção de modelos é trabalhar com modelos encaixados. Ou seja, é preciso identificar um modelo geral tal que os modelos em investigação sejam casos particulares deste. Alguns métodos estatísticos são possíveis para lidar com esse tipo de situação, um deles é a utilização do Teste da Razão de Verossimilhanças (TRV). O teste consiste em avaliar

o logaritmo da função de verossimilhança no modelo geral $\ell(\widehat{\beta}_G)$ e no modelo reduzido $\ell(\widehat{\beta}_R)$, a partir daí é possível calcular a estatística de teste da seguinte forma

$$TRV = -2 \log \left(\frac{\ell(\widehat{\beta}_R)}{\ell(\widehat{\beta}_G)} \right) = 2(\log \ell(\widehat{\beta}_G) - \log \ell(\widehat{\beta}_R)).$$

De acordo com Colosimo e Giolo (2006), sob a hipótese nula H_0 de que o modelo em investigação é adequado, a estatística TRV tem distribuição qui-quadrado com o número de graus de liberdade igual a diferença do número de parâmetros dos modelos sendo comparados. Ainda, segundo os autores, em análise de sobrevivência este teste é comumente aplicado utilizando o modelo Gama generalizado como modelo geral, uma vez que, os modelos Exponencial, Weibull, Log-normal e Gama são casos particulares deste modelo.

2.7 ANÁLISES ESTATÍSTICAS

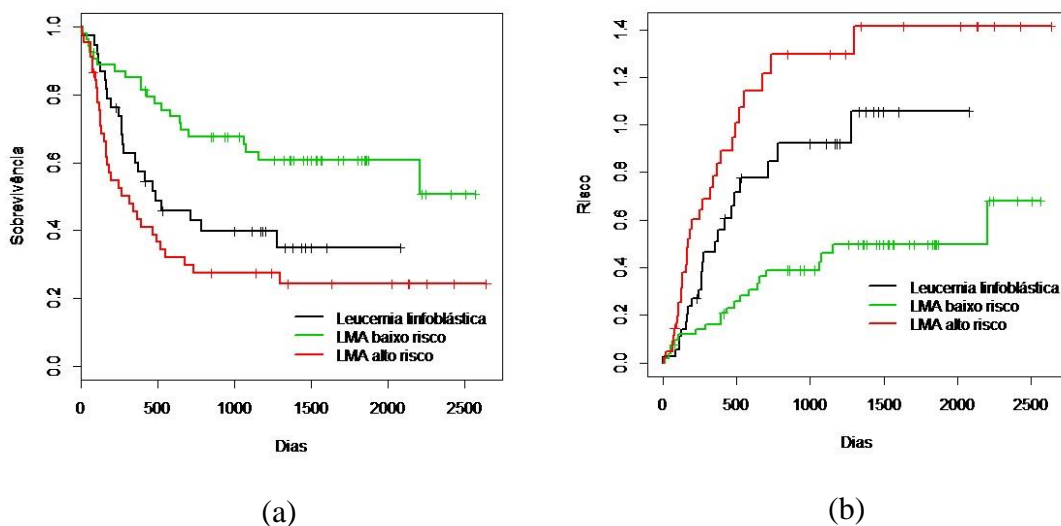
Os dados utilizados nesse trabalho são provenientes do institute for Helth & Society da universidade de Wisconsin (Medical College of Wisconsin), estes dados estão descritos detalhadamente em Klein e Moeschberger (2005). O banco de dados é constituído de um total de 137 pacientes (38 LLA, 99 LMA) os quais possuem uma distinção entre os tipos de câncer, leucemia mieloide aguda (LMA) e leucemia linfoblástica aguda (LLA), foram tratados em quatro hospitais avaliados: 76 pacientes no Hospital da Universidade Estadual de Ohio (OSU), em Columbus, 21 pacientes de Hahnemann University(HU), na Filadélfia, 23 pacientes do Hospital St. Vincent (SVH) em Sydney, Austrália, e 17 pacientes no Alfred Hospital (AH) em Melbourne. O estudo consiste de transplantes realizados nessas instituições no período de 1 de março de 1984 a 30 de Junho de 1989. O máximo de acompanhamento foi de 7 anos. Foram 42 pacientes que recaíram e 41 que morreram enquanto em remissão. Vinte e seis pacientes tiveram um episódio de GVHD aguda (doença do enxerto hospedeiro) e 17 pacientes tiveram reicidiva ou morreram em remissão, sem suas plaquetas voltarem a níveis normais. As variáveis estudadas foram: Z_1 : Idade do paciente; Z_2 : Idade do doador; Z_3 : Sexo do paciente; Z_4 : Sexo do doador; Z_5 : Citomegalovírus do paciente; Z_6 : Citomegalovírus do doador; Z_7 : Tempo de espera em dias para o transplante de medula óssea; Z_8 : O sistema de classificação francês, americano e britânico (FAB) para leucemia mielóide aguda (LMA); Z_9 : Hospital; Z_{10} : MTX medicamento administrados nos pacientes. gg_1 : Grupo de risco LLA; gg_2 : Grupo LMA baixo risco; gg_3 : Grupo de risco LMA alto risco; a = GVHD agudo; c =GVHD crônico; p = Plaquetas em níveis normais (recuperação das plaquetas).

Para a avaliação das curvas de sobrevivência foi utilizado o estimador de Kaplan-Meier, bem como o teste log-rank para realizar as comparações entre essas curvas. Para identificar a influência das covariáveis no tempo de sobrevivência dos indivíduos foram usados os modelos de regressão paramétricos Exponencial, Weibull e Log-normal. Também se utilizou o modelo de riscos proporcionais de Cox e os modelos de fragilidade Gama e Log-normal para avaliar a relação das covariáveis com o risco dos indivíduos sofrerem o evento de interesse, sendo os modelos selecionados com base nos valores do AIC e no Teste da Razão de Verossimilhanças. As análises foram conduzidas através do *software* R (R CORE TEAM, 2014), com o auxílio do pacote *survival* (THERNEAU, 2014).

3 RESULTADOS

O início da análise se deu com a avaliação da função de sobrevivência por meio do estimador de Kaplan-Meier, considerando três diferentes grupos constituídos por (grupo 1 - LLA; grupo 2 - LMA baixo risco e o grupo 3 - LMA alto risco). Dessa forma, na Figura 1 é possível observar o resultado obtido para as curvas de sobrevivência desses grupos.

Figura 1: Função de sobrevivência (a) e Risco acumulado (b) para os tipos de Leucemia considerados no estudo



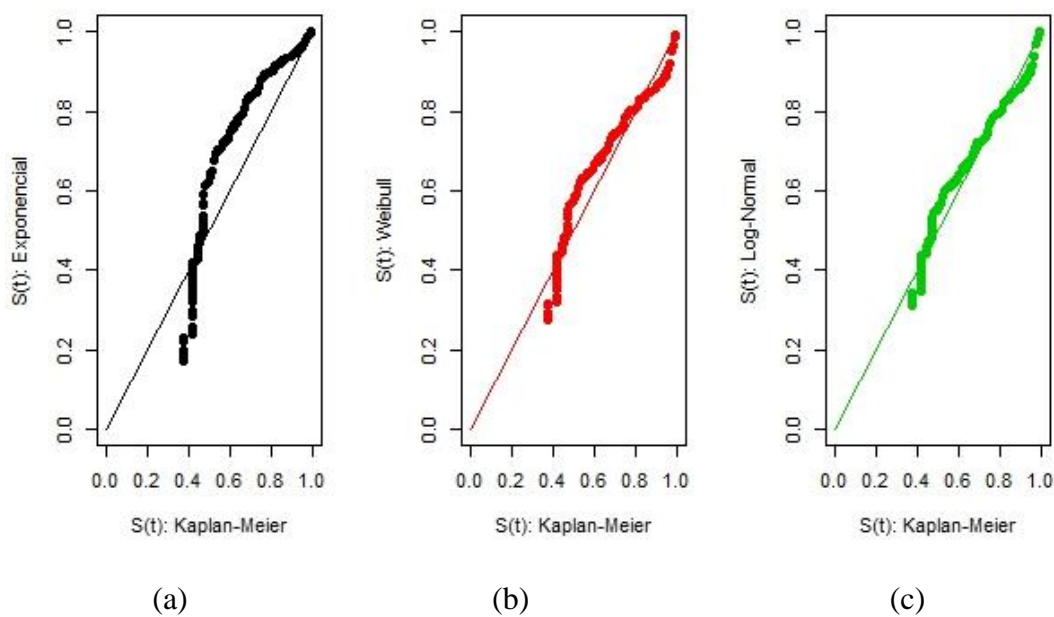
Na Figura 1 (a), é possível observar que os pacientes portadores de Leucemia mieloide (LMA) tiveram menores probabilidades de sobrevivência em relação aos pacientes com Leucemia linfoblástica (LLA), com maior destaque para o grupo dos pacientes com Leucemia mieloide aguda de alto risco, em que fica claro que esta apresentou maior agressividade quanto a morte dos indivíduos após o transplante de medula óssea. Ainda na Figura 1 (b) podemos observar que os

pacientes pertencentes aos grupos 2 e 3 (LMA) tiveram uma curva de risco muito superior aos pacientes do grupo 1, isto é, portadores de Leucemia linfoblástica, onde novamente, podemos dar maior destaque aos indivíduos do grupo 3 que apresentaram maiores taxas de falha ao longo de tempo de acompanhamento. Este comportamento revela um efeito gradual da gravidade da doença.

Após a aplicação do método de Kaplan-Meier, efetuou-se a comparação entre as curvas de sobrevivência dos grupos em questão por meio do teste log-rank, para investigar se existe diferença significativa entre as curvas, resultando numa estatística $\chi^2 = 16,3$ para 2 graus de liberdade e um p-valor igual a 0,000294, ou seja, indicando que de fato existe diferença significativa entre as três curvas, ou seja, os pacientes possuem tempos de vida diferentes entre os grupos de leucemia.

Uma vez verificada as diferenças nas funções de sobrevivência dos grupos, partiu-se então para o ajuste dos modelos. Num primeiro momento, os modelos Exponencial, Weibull e Log-normal foram ajustados a fim de identificar a influência das covariáveis no tempo de sobrevida dos pacientes sem haver neste caso distinção de grupos. Neste sentido, a Figura 2 apresenta os resultados para as curvas de sobrevivência desses modelos vs o estimador de Kaplan-Meier com o objetivo de identificar o modelo que melhor se adequa aos dados, tal como proposto em Colosimo e Giolo (2006).

Figura 2: Gráfico de adequação das curvas de sobrevivência paramétricas vs sobrevivência estimada por Kaplan Meier (a- Exponencial, b- Weibull, c – Log-Normal).



De acordo com a Figura 2 (a), o modelo Exponencial parece não ser adequado para explicar o conjunto de dados em análise. Por outro lado, os modelos Weibull e Log-Normal vistos na Figura 2 (b) e (c) respectivamente acompanham mais de perto a reta $y = x$, indicando serem estes, mais propícios para adequação aos dados.

Uma vez que as técnicas gráficas para seleção dos modelos apresentam um viés de interpretação, sendo seu resultado susceptível à erros, que podem variar de acordo com o analista, estas são usadas para dar direcionamento as análises que virão em seguida, tal como a seleção dos modelos com base nos valores do AIC e o teste da razão de verossimilhanças (TRV). Neste seguimento a Tabela 1 apresenta os resultados obtidos para os modelos em investigação.

Tabela 1: Comparação entre os modelos paramétricos com a distribuição Gama generalizada via Teste de Razão de Verossimilhança (TRV) e Critério de informação de Akaike (AIC).

Modelo	$\log(L(\theta))$	TRV	Valor p	AIC (menor melhor)
Gama generalizado	-632.2	-	-	-
Weibull	-627.6	$2(627.6-632.2)=-9.2$	1	1275.34
Log-normal	-623.7	$2(623.7-632.2)=-17.0$	1	1268.42

De acordo com os resultados apresentados na Tabela 1, tendo como base o valor do AIC, tem-se que o modelo Log-Normal apresenta maior adequabilidade em relação às demais para os dados de pacientes transplantados, sendo este utilizado para dar seguimento as análises. Neste sentido, na Tabela 2 é possível observar as estimativas dos parâmetros obtidas para esse modelo. Outro modelo utilizado para ajustar o conjunto de dados neste trabalho foi o modelo de riscos proporcionais de Cox, devido à sua versatilidade e comprovada adequação à diversos problemas práticos. Deste modo, a Tabela 3 apresenta as estimativas dos parâmetros para este modelo.

Tabela 2: Variáveis selecionadas para o modelo Log-Normal com respectivas razões de risco (R.R. = $\exp(\text{coef})$) e intervalos de confiança (IC95%).

variáveis	$\exp(\text{coef})$	$\exp(-\text{coef})$	Limite inferior	Limite superior
p	0.3786	2.6413	0.2028	0.7067
c	0.4254	2.3506	0.2614	0.6923
z8	1.5954	0.6268	0.8930	2.8502
z9	0.7215	1.3861	0.5300	0.9821
z10	2.1616	0.4626	1.1157	4.1878
gg1	0.9318	1.0732	0.4718	1.8404
gg2	0.4358	2.2945	0.2398	0.7920

A partir dos resultados obtidos na Tabela 2, observamos que variável doença do hospedeiro (*c*) se mostrou como fator protetor, logo, pacientes sem a doença do hospedeiro têm risco maior de ir a óbito do que os pacientes que tem a doença do hospedeiro. Este efeito protetor pode ser interpretado como um efeito indireto, uma vez que estes pacientes podem estar tendo maior acompanhamento e controle da doença. A variável FAB (Z_8) apresentou-se como um fator de risco indicando que a classificação é importante, enquanto que a variável hospital (Z_9) atuou como fator protetor, assim como a variável MTX (Z_{10}) se mostrou como um fator de risco. O modelo apresentou um poder explicativo absoluto de 32,2%. A probabilidade de concordância estimada pelo modelo teve alto valor discriminatório ou preditivo (76,8%).

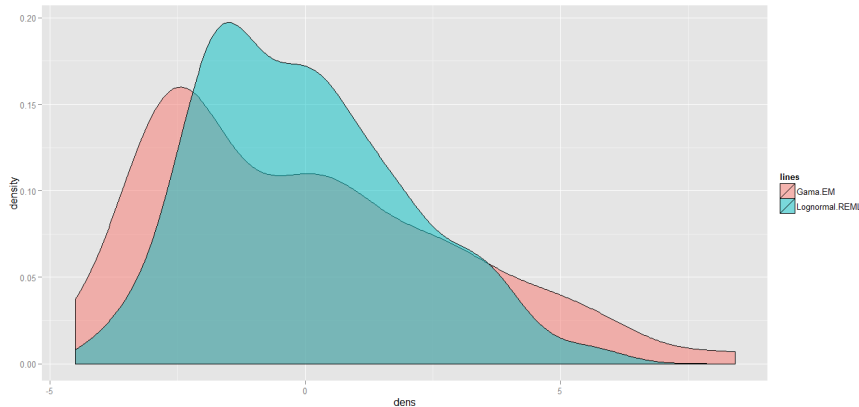
Após os ajustes dos modelos convencionais da análise de sobrevivência, também foram ajustados os modelos de fragilidade Gama e Log-normal à base de dados e feita as devidas comparações por meio do valor do AIC, sendo o modelo Log-Normal mais adequado para o estudo das fragilidades. Deste modo, a Tabela 3 apresenta as estimativas dos parâmetros obtidas para o modelo de fragilidade Log-normal.

Tabela 3: Variáveis selecionadas para o modelo Log-Normal de Fragilidade com respectivas razões de risco (R.R. = exp(coef)) e IC95%.

Fragilidade-LogNormal	exp(coef)	exp(-coef)	Limite inferior	Limite superior
p	0.08425	11.8693	0.026780	0.2651
c	0.23011	4.3457	0.104061	0.5089
z8	2.40000	0.4167	0.967981	5.9505
z9	0.63312	1.5795	0.416213	0.9631
z10	3.36885	0.2968	1.242339	9.1353
gg1	0.75457	1.3253	0.255363	2.2297
gg2	0.23945	4.1763	0.094842	0.6045

Observando os resultados nas Tabelas 2 e 3 pode-se observar que houve diferenças nas estimativas dos parâmetros dos modelos, sendo o modelo de fragilidade mais apropriado para esta base de dados e seus resultados mais confiáveis que o modelo de Cox. Para o ajuste do modelo foram necessárias 21 iterações do algoritmo de Newton-Raphson. Nos resultados obtidos verificou-se a que a variância dos efeitos aleatórios foi de 2,5249 e que o efeito da fragilidade foi significativo, o que demonstra a correta consideração desse efeito aleatório no modelo. As interpretações dos parâmetros desse modelo podem ser feitas de maneira análoga ao modelo de Cox.

Figura 3: Densidade da fragilidade nos três grupos de Câncer.



De acordo com gráfico das fragilidades dos pacientes, implementados com os modelos de fragilidade Gama e Lognormal, percebe-se que em ambos os modelos é possível destacar três picos, justamente os três grupos considerados no estudo (LLA, LMA baixo risco e alto risco), podendo se ver a distribuição da fragilidade nos diferentes grupos via modelo Lognormal considerado.

4 DISCUSSÃO

Apesar das diversas ferramentas existentes no campo da estatística, e do estímulo para que elas sejam utilizadas, há raras publicações especificamente em transplante de medula óssea que integrem técnicas mais elaboradas de análise de sobrevivência. É possível que o progresso metodológico mais lento nesta área seja, em parte, devido à raridade da doença. Neste sentido, Sayehmiri et al. (2009) avaliou os fatores prognósticos da sobrevida global (OS) após o transplante de células-tronco hematopoiéticas (TCTH) em pacientes com leucemia linfoblástica aguda (LLA) usando tempo de falha acelerado (AFT), risco proporcional de Cox (PH) e modelos de Cox com coeficientes variável no tempo. Os autores concluíram que as distribuições de AFT podem ser uma ferramenta útil para o reconhecimento de fatores prognósticos da OS em pacientes com leucemia linfoblástica aguda. Fuerst et al. (2016) em um estudo do tempo de vida após o transplante de células-tronco hematopoiéticas com base em uma análise de dados do Registro Alemão para Transplante de Células-Tronco, no qual foram incluídos 14.951 pacientes com leucemia mielóide aguda, leucemia linfocítica aguda, síndrome mielodisplásica e linfoma não-Hodgkin transplantado com células-tronco do sangue periférico ou enxertos de medula óssea. Os autores utilizaram modelos de regressão múltipla de Cox no qual foram testados para efeitos dependentes do tempo dentro de cada grupo de diagnóstico e apesar de ser um estudo recente com dados atualizados o efeito de fragilidade não foi avaliado. Já Lin (2018) utilizou o modelo de fragilidade paramétrico em um estudo de um modelo de sobrevivência, introduzido nos perfis de sobrevivência o efeito de

fragilidade de 11.160 pacientes submetidos a TCTHs primeiro alogênicos nos Estados Unidos entre 1995 e 2006; Neste estudo, os dados foram compartilhados pelo *Center for International Bone and Marrow Transplant Research* e os resultados encontrados pelo autor sugerem que em comparação à estratificação, a modificação permite a apresentação de duas entidades em relação ao tempo; idade e hora calendário, além de ter permitido resultados novos para a literatura, tais como duas descobertas inesperadas: i) um risco inverso em forma de J na fonte de doadores não relacionada $t \leq 100$ dias; ii) a convergência de linhas de risco não relacionadas em 100 dias $< t \leq 365$ dias sugere uma manifestação máxima de senescência entre os sobreviventes.

Estes estudos demonstram que o uso de modelos de fragilidade de Cox seja paramétrico ou semiparamétrico são escassos na literatura para estudos de transplante de medula óssea, apesar de esse tipo de estudo ter uma maior capilaridade nas aplicações a dados de sobrevivência em diversas áreas do conhecimento.

5 CONCLUSÕES

Por meio do teste log-rank foi possível comprovar diferença significativa entre as curvas de sobrevivência dos grupos estudados neste trabalho (grupo 1: LLA, grupo 2: LMA baixo risco e grupo 3: LMA alto risco), sendo seus comportamentos ao longo do tempo de acompanhamento observados através do estimador de Kaplan-Meier.

Entre os modelos ajustados, o modelo de fragilidade Log-normal apresentou melhor adequação aos dados, com menor valor de AIC e maior grau de predição. O modelo revelou que as variáveis c , $Z9$, $gg1$ e $gg2$ mostraram-se como fator de proteção nos riscos de morte dos pacientes após o transplante de medula óssea, enquanto que as variáveis $Z8$ e $Z10$ apresentaram ser fatores de agravamento para esse desfecho.

A partir das técnicas estatísticas usadas nesse trabalho, evidenciou-se que análise de sobrevivência é uma importante ferramenta na área da saúde desde que todos os critérios que cada técnica estatística possui sejam seguidos de forma correta, ajudando a melhor entender o comportamento e quais características afetam os pacientes sobre risco.

REFERÊNCIAS

- BOGAERTS, K; KOMAREK, A; LESAFFRE, E. *Survival analysis with interval-censored data: A practical approach with examples in R, SAS, and BUGS*. CRC Press, 2017.
- COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. 1. ed. São Paulo: Edgard Blucher, 2006.
- COX, D. R. Regression models and life tables (with discussion). *Journal Royal of Statistical Society, B*, v. 34, p. 187–220, 1972.
- DINIZ, C.; LOUZADA, F. *Modelagem estatística para risco de crédito*. ABE, São Paulo-SP, 2012.
- DUCHATEAU, L.; JANSSEN, P. *The Frailty Models*. 1. ed. New York: Springer Science & Business Media, 2008.
- EMURA, T.; CHEN, Y. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. 1st ed. 2018.
- FUERST, Daniel et al. Time-dependent effects of clinical predictors in unrelated hematopoietic stem cell transplantation. *haematologica*, v. 101, n. 2, p. 241-247, 2016.
- KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.
- KARIM, M. R., & ISLAM, M. A. *Reliability and Survival Analysis*. Springer Singapore. 2019.
- KLEIN, J. P.; MOESCHBERGER, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*. 2. ed. New York: Springer Science & Business Media, 2005.
- HARRELL Jr, F. E. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
- Lin Y (2018) AFT survival model to capture the rate of aging and age-specific mortality trajectories among first-allogeneic hematopoietic stem cells transplant patients. *PLOS ONE* 13(3): e0193287. <https://doi.org/10.1371/journal.pone.0193287>
- LEE, E. T.; WANG, J. *Statistical methods for survival data analysis*. John Wiley & Sons, 2003.
- MOORE, D. F. *Applied survival analysis using R*. New Jersey: Springer, 2016.
- SAYEH, MIRI K., et al. "Predictive factors of survival time after hematopoietic stem cell transplant in acute myeloid leukemia patients who received allogeneic BMT from matched sibling donors using Generalized Gamma models." *International Journal Of Hematology-Oncology And Stem Cell Research (IJHOSCR)*, Volume 3 , Number 1; 2009: 21-26.
- THERNEAU, T. *Survival: A Package for Survival Analysis in S*. R package version 2.37-7, URL <http://CRAN.R-project.org/package=survival>. 2014.

WEIBULL, W. A statistical theory of the strength of materials. *Stockholm: Generalstabens litografiska anstalts förlag*, 1939.

WIENKE, A. *Frailty Models in Survival Analysis*. Boca Raton: CRC Press, 2010