MODELOS DE SOBREVIVÊNCIA DE LONGA DURAÇÃO APLICADOS AO ESTUDO DO COMPORTAMENTO DE RETORNO DO DOADOR DE SANGUE VOLUNTÁRIO

Adriana de Fátima LOURENÇON[1]
Edson Zangiacomi MARTINEZ[1]
Josmar MAZUCHELI[2]
Oranice FERREIRA[3]

▪ RESUMO: Estratégias para assegurar a segurança dos estoques de sangue nos serviços de transfusão e hemoterapia, incentivando o retorno dos doadores voluntários, dependem do conhecimento do comportamento de retorno dos doadores de sangue. No presente artigo, utilizamos modelos de sobrevivência de longa duração para a modelagem do tempo entre a primeira doação voluntária de sangue e o seguinte retorno, considerando as distribuições Weibull, log-normal, log-logística e gama generalizada. Consideramos ainda que o parâmetro de escala e a proporção de indivíduos que não retornam a novas doações são dependentes de um vetor de covariáveis. Observamos que os modelos de longa duração, especialmente o modelo baseado na distribuição gama generalizada estendida, apresentaram uma grande habilidade para a representação do comportamento de retorno do doador de sangue.

▪ PALAVRAS-CHAVE: Análise de sobrevivência; modelos de longa duração; distribuição gama generalizada; doadores de sangue; bioestatística.

1 Introdução

A procura de sangue e hemoderivados, assegurando seu fornecimento de maneira segura e sustentável, é um dos maiores desafios enfrentados pelos serviços

[1]Departamento de Medicina Social, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo – FMRP/USP, Ribeirão Preto, SP, Brasil. E-mail: *adlourencon@yahoo.com.br* / *edson@fmrp.usp.br*
[2]Departamento de Matemática, Centro de Ciências Exatas, Universidade Estadual de Maringá – UEM, Maringá, PR, Brasil.
[3]Centro Regional de Hemoterapia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (Hemocentro), Universidade de São Paulo – FMRP/USP, Ribeirão Preto, SP, Brasil.

de hemoterapia. Com o aumento da população idosa, do acesso da população aos cuidados médicos e da complexidade dos procedimentos terapêuticos, a preocupação com a escassez de sangue ganha evidência, tornando-se necessários os programas de promoção do fornecimento seguro de sangue baseado na captação e fidelização de doadores voluntários vindos de populações de baixo risco. Ludwig e Rodrigues (2005) enfatizam que a busca do doador voluntário e habitual se deve principalmente à segurança e a razões econômicas, dado que doadores testados e retestados significam bolsas de sangue com maior margem de segurança para o receptor e um número menor de exames sorológicos desprezados.

Estratégias para assegurar a segurança dos estoques de sangue, incentivando o retorno dos doadores voluntários, dependem do conhecimento do comportamento de retorno dos doadores de sangue. Estimativas da chance de retorno de um indivíduo que comparece a uma primeira doação voluntária são de grande interesse para o planejamento destas estratégias. Com este propósito, James e Matthews (1996) e Ownby, Watanabe e Nass (1999) utilizaram modelos de riscos proporcionais de Cox para a modelagem do tempo até uma próxima tentativa de doação. No entanto, ao assumirmos que uma parcela dos indivíduos que comparecem para uma primeira doação não retornam para novas doações, modelos de sobrevivência de longa duração podem ser mais adequados para descrever o comportamento de retorno do doador de sangue.

No presente estudo, exploramos o uso de modelos de sobrevivência de longa duração, com funções de sobrevivência segundo distribuições Weibull, log-normal, log-logística, gama e gama generalizada, para descrever o tempo decorrido entre a primeira doação de sangue de um indivíduo e o seu retorno seguinte. Os modelos de longa duração podem ser considerados como uma mistura de duas componentes (Farewell, 1982; Perperoglou Keramopoullos e Van Houwelingen, 2007), sendo p a probabilidade de uma componente. Consideraremos este parâmetro dependente de um vetor de covariáveis. O parâmetro p estima, em muitos estudos, a proporção de indivíduos curados (Lam, Fong e Tang, 2005) ou que adquiriram imunidade ao evento de interesse. No presente estudo, p é a proporção de doadores de sangue que não retornam para uma nova doação.

Na seção 2, apresentamos uma descrição do banco de dados. Na seção 3, apresentamos uma breve revisão de conceitos básicos de análise de sobrevivência, uma descrição das distribuições de probabilidade utilizadas na modelagem dos dados, e uma apresentação do modelo de longa duração. A análise dos dados é apresentada na seção 4, e comentários finais estão presentes na seção 5.

2 O banco de dados

O Hemocentro de Ribeirão Preto é responsável por mais de 95% da coleta de sangue e distribuição de hemocomponentes em uma região com 213 municípios e 4,8 milhões de habitantes, através de uma rede de quatro Núcleos de Hematologia e Hemoterapia (Araçatuba, Fernandópolis, Franca e Presidente Prudente), quatro Unidades de Hemoterapia (Batatais, Bebedouro, Olímpia e Serrana) e Postos de

Coleta, além da sua unidade central, integrados por um sistema de informática. O presente estudo utilizou um banco de dados obtido de um levantamento retrospectivo das informações registradas no Hemocentro de Ribeirão Preto. Este levantamento constou de informações de todos os doadores cadastrados no período de julho de 1996 (quando começou a funcionar o sistema informatizado, pelo Sistema de Controle de Doadores e Laboratórios), até junho de 2005. Foram excluídos os registros de doações autólogas (doação vinda do indivíduo que doa para si próprio, por indicação e critério médico), por aférese (procedimento com separação de algum hemocomponente do sangue através de um equipamento: plaquetas, hemáceas, plasma, granulócitos e linfócitos) e aquelas classificadas como exame extra-rotina (doador que passa por todo o processo de triagem que antecede a doação mas não doa efetivamente uma bolsa de sangue, apenas colhe amostras para exames). Foram também excluídos os registros de doadores reprovados definitivamente na primeira doação de sangue, por ser um público do qual não é esperado o seu retorno.

Assim, o banco de dados final é composto por registros de 115.553 indivíduos com idade entre 18 a 65 anos, que compareceram voluntariamente para doar sangue, onde a variável de interesse é o tempo decorrido entre a primeira doação de sangue e o próximo retorno do indivíduo para uma doação. É importante observar que neste segundo momento nem sempre a doação propriamente dita foi realizada, pois o indivíduo pode ter sido reprovado na entrevista que é feita antes da doação.

3 Formulação do modelo

3.1 Modelos para dados de sobrevivência

O comportamento da variável aleatória ($v.a.$) tempo de sobrevida, $T \geq 0$, é descrito por três funções: $f(t)$, a função densidade de probabilidade; $S(t)$, a função de sobrevivência; e $h(t)$, a função risco, onde t é uma observação de T. A função de sobrevivência $S(t)$, no presente estudo, é a probabilidade acumulada de um indivíduo não ter retornado para uma nova doação de sangue em um momento t, definida por$S(t) = P(T > t) = 1 - F(t)$, onde $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \longrightarrow \infty$ e $F(t) = \int_0^t f(u)du$ representa a função distribuição acumulada. A função risco $h(t)$ é definida como o limite da probabilidade de ser observado o evento de interesse (uma segunda doação de sangue) no intervalo de tempo $[t, t+\Delta t]$, dado que o indivíduo não compareceu para a segunda doação de sangue até o tempo t, quando $\Delta t \longrightarrow 0$.

Os modelos paramétricos para dados de sobrevivência assumem uma distribuição de probabilidade conhecida para o tempo de sobrevida (ver, por exemplo, Lee e Wang, 2003). Algumas distribuições usuais são:

(a) **Distribuição Weibull**: considerando t uma observação de uma $v.a.$ T, a função de sobrevivência e sua respectiva função risco são dadas por $S(t) = \exp[-(\lambda t)^\gamma]$ e $h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$, onde se observa que $h(t)$ é uma função monótona estritamente crescente para $\gamma > 1$, estritamente decrescente para $\gamma < 1$ e constante para $\gamma = 1$. Quando $\gamma = 1$, $f(t)$ é a função densidade de uma distribuição

exponencial. No presente estudo, consideraremos a parametrização $\lambda = 1/\exp(\mu)$ e $\gamma = \sigma^{-1}$, sendo μ um número real e $\sigma > 0$.

(b) **Distribuição log-normal**: se T é uma $v.a.$ tal que $\ln(T)$ tem distribuição normal com média μ e variância σ^2, dizemos que T tem distribuição log-normal, com função de sobrevivência, então dada por

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x} \exp\left[-\frac{1}{2\sigma^2}\left(\ln x - \mu\right)^2\right] dx.$$

Ao considerarmos $\lambda = \exp(-\mu)$, a função de sobrevivência é dada por $S(t) = 1 - \Phi\left[\ln\left(\lambda t \sigma^{-1}\right)\right]$, onde $\Phi(\bullet)$ é a função acumulada da distribuição normal.

(c) **Distribuição log-logística**: se T é uma $v.a.$ tal que $\ln(T)$ tem distribuição logística, T tem distribuição log-logística, com função de sobrevivência é dada por $S(t) = (1 + \lambda t^\gamma)^{-1}$ Quando $\gamma > 1$, a função risco é crescente até um dado momento, e em seguida, decrescente, similar à $h(t)$ quando T possui distribuição log-normal. Quando $\gamma \leq 1$, a função risco é decrescente. Consideraremos neste estudo $\lambda = \exp\left(-\frac{\mu}{\sigma}\right)$ e $\gamma = \frac{1}{\sigma}$, sendo μ real e $\sigma > 0$.

(d) **Distribuição gama e gama generalizada**: se T é uma $v.a.$ com distribuição gama, a função de sobrevivência é dada por

$$S(t) = \frac{\lambda}{\Gamma(\alpha)} \int_t^\infty (\lambda x)^{\alpha-1} \exp(-\lambda x)\, dx.$$

A função risco, obtida da relação $h(t) = f(t)/S(t)$ é cresente se $\alpha > 1$, decrescente se $0 < \alpha < 1$, convergindo para um valor constante quando t cresce de 0 a infinito. Se $\alpha = 1$, a função risco é constante.

A distribuição gama generalizada é caracterizada por três parâmetros, com função densidade de probabilidade dada por

$$f(t) = \frac{\gamma\lambda^{\gamma\lambda}}{\Gamma(\alpha)} t^{\gamma\alpha-1} \exp\left[(-\lambda t)^\gamma\right],$$

onde $t > 0$, $\lambda > 0$ é o parâmetro de escala e $\gamma > 0$ e $\alpha > 0$ são parâmetros de forma. Nota-se que, a partir desta distribuição, encontramos alguns casos especiais: se $\alpha = \gamma = 1$, temos $T \sim Exp(\lambda)$; se $\alpha = 1$, temos $T \sim Weibull(\gamma, \lambda)$; e se $\gamma = 1$, temos $T \sim Gama(\alpha, \lambda)$. Além disso, a distribuição log-normal aparece como um caso limite da distribuição gama generalizada quando α tende a infinito.

Uma outra maneira de definir a função densidade de probabilidade da distribuição gama generalizada é

$$f(t) = \frac{|\gamma|\alpha^\alpha\lambda^{\alpha\gamma}}{\Gamma(\alpha)} t^{\alpha\gamma-1} \exp -\alpha\left(\lambda t\right)^\gamma$$

Esta distribuição é referida como distribuição gama generalizada extendida, denotada por $T \sim GGE(\lambda, \alpha, \gamma)$. Neste estudo, consideraremos $\lambda = \exp(-\mu)$, $\alpha = \frac{1}{\varphi^2}$ e $\gamma = \frac{\varphi}{\sigma}$.

Para expressar o relacionamento de k cováriaveis relacionadas com o perfil do doador, denotadas por $\mathbf{X}_i = (x_{1i}, x_{2i}, ..., x_{ki})$, na forma da distribuição do tempo até o próximo retorno do doador, assumimos por meio da relação $\mu = \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij}$ a expressão do efeito dessas cováriaveis no parâmetro de forma.

3.2 Modelos de longa duração

Seja T uma *v.a.* com função densidade de probabilidade $f(t)$, onde os parâmetros de escala e forma são, genericamente, λ e γ. Ao associar o tempo de sobrevida T a um vetor $\mathbf{X}$ de k covariáveis por meio do parâmetro de escala λ, a função de sobrevida é dada por $S_0(t \mid \lambda(\mathbf{x}), \gamma) = P(T > t \mid \mathbf{x})$, ou seja, a probabilidade de um indivíduo não retornar para uma nova doação até o instante t. Seja p, tal que $0 < p < 1$, a proporção de doadores de sangue que não retornam para uma nova doação. O modelo de sobrevivência de longa duração (Maller e Zhou, 1996) é dado por

$$S(t \mid \mathbf{x}) = p + (1 - p) S_0(t \mid \lambda(\mathbf{x}), \gamma),$$

onde S é a função de sobrevivência na população e S_0 é a função de sobrevivência associada aos indivíduos que retornam a novas doações. Sendo $S_0(0 \mid \lambda(\mathbf{x}), \gamma) = 1$, temos $S(0 \mid \mathbf{x}) = 1$, e, considerando $S_0(\infty \mid \lambda(\mathbf{x}), \gamma) = 0$, temos $S(\infty \mid \mathbf{x}) = p$, ou seja, à medida que as observações de t assumem valores grandes, a função de sobrevivência S aproxima-se de p. Se assumirmos $p = 0$, estaremos considerando o modelo paramétrico usual, onde $S(t \mid \mathbf{x}) = S_0(t \mid \lambda(\mathbf{x}), \gamma)$, ou seja, descartamos a estimação de uma proporção de indivíduos que não retornam.

Podemos considerar o parâmetro p dependente de l covariáveis denotadas por $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, ..., Z_{li})$, por uma relação logito, da forma $p(z_i) = \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{l} \beta_j z_{ij}\right)\right]^{-1}$. Nesse caso, a função de sobrevivência pode ser escrita como

$$S(t \mid \mathbf{x}, \mathbf{z}) = p(\mathbf{z}) + (1 - p(\mathbf{z})) S_0(t \mid \lambda(\mathbf{x}), \gamma). \tag{1}$$

Os vetores $\mathbf{X}$ e $\mathbf{Z}$ podem ou não ser iguais.

Vamos considerar uma amostra aleatória $T_1, ..., T_n$, sendo que T_i representa o tempo entre a primeira doação voluntária e o seguinte retorno do i-ésimo doador de sangue do banco de dados de n registros. A cada variável T_i podemos associar um vetor de covariáveis com observações $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{ki})'$, $\mathbf{z}_i = (z_{1i}, z_{2i}, ..., z_{li})'$, e uma variável indicadora δ_i, onde $\delta_i = 1$ se t_i é observável e $\delta_i = 0$ se t_i é uma censura à direita, ou seja, não há informação sobre o retorno do doador dentro do seguimento. A função de verossimilhança é dada por

$$L = \prod_{i=1}^{n} \left[(1 - p(\mathbf{z}_i)) f\left(t_i \mid \mathbf{x}_i\right)\right]^{\delta_i} \left[S(t_i \mid \mathbf{x}_i, \mathbf{z}_i)\right]^{1-\delta_i}, \tag{2}$$

onde $f\left(t_i \mid \mathbf{x}_i\right)$ é a função de densidade de um evento (retorno para doação de sangue) ao tempo t_i dado $\mathbf{x}_i$ e $S(t_i \mid \mathbf{x}_i, \mathbf{z}_i)$ é a função de sobrevivência (1). Estimativas de máxima verossimilhança podem ser obtidas maximizando (2) com

o auxílio do procedimento NLP do programa SAS, como sugerido por Mazucheli, Louzada-Neto e Achcar (2003), que busca pontos de máximo locais por algoritmos de otimização. Uma alternativa para o ajuste do modelo com covariáveis é o programa *gfcure*, desenvolvido por Peng, Dear e Denham (1998). O logaritmo de (2) é

$$\ln L = \sum_{i=1}^{n} \left\{ \delta_i \ln \left[(1 - p(\mathbf{z}_i)) f\left(t_i \mid \mathbf{x}_i, \mathbf{z}_i\right) \right] + (1 - \delta_i) \ln S(t_i \mid \mathbf{x}_i, \mathbf{z}_i) \right\}, \quad (3)$$

onde $S(t_i \mid \mathbf{x}_i, \mathbf{z}_i)$ é dado pela expressão (1).

Ao considerarmos o modelo Weibull, temos $f\left(t_i \mid \mathbf{x}_i\right) = \lambda(\mathbf{x}_i)\gamma(\lambda(\mathbf{x}_i)t)^{\gamma-1}$ $\exp[(-\lambda(\mathbf{x}_i)t)^{\gamma}]$ e $S_0(t \mid \lambda(\mathbf{x}), \gamma) = \exp[(-\lambda t)^{\gamma}]$. A expressão (3) é então dada por

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n} \delta_i \ln \left\{ (1 - p(\mathbf{z}_i)) \lambda(\mathbf{x}_i) \gamma (\lambda(\mathbf{x}_i) t_i)^{\gamma - 1} \exp[(-\lambda(\mathbf{x}_i) t_i)^{\gamma}] \right\} + \\
&\quad + \sum_{i=1}^{n} (1 - \delta_i) \ln \left\{ p(\mathbf{z}_i) + (1 - p(\mathbf{z})) \exp[(-\lambda(\mathbf{x}_i) t_i)^{\gamma}] \right\} = \\
&= \sum_{i=1}^{n} \delta_i \left\{ \ln(1 - p(\mathbf{z}_i)) \right\} + \\
&\quad + \sum_{i=1}^{n} \delta_i \left(\ln \gamma + \gamma \ln \lambda(\mathbf{x}_i) + \gamma \ln t_i - \ln t_i \right) - \sum_{i=1}^{n} \delta_i (\lambda(\mathbf{x}_i) t_i)^{\gamma} + \\
&\quad + \sum_{i=1}^{n} (1 - \delta_i) \ln \left\{ p(\mathbf{z}_i) + (1 - p(\mathbf{z}_i)) \exp[(-\lambda(\mathbf{x}_i) t_i)^{\gamma}] \right\}.
\end{aligned}$$

Nesta expressão, substituímos $\lambda(\mathbf{x}_i)$ por $\exp(-\mu(\mathbf{x}_i)/\sigma)$ onde $\mu(\mathbf{x}_i) = \zeta_0 + \sum_{j=1}^{l} \zeta_j z_{ij}$, γ por σ^{-1}, e $p(\mathbf{z}_i)$ por $\left[1 + exp(\beta_0 + \sum_{j=1}^{l} \beta_j z_{ij})\right]^{-1}$. De maneira análoga, encontramos expressões para o logaritmo da função de verossimilhança (3) considerando outros modelos probabilísticos.

4 Resultados

Utilizamos informações de 115.553 doadores voluntários, sendo 45.143 de sexo feminino (39,1%) e 70.410 de sexo masculino (60,9%). Observou-se que 15,7% destes doadores tinham 18 a 19 anos no momento da primeira doação. E ainda, na primeira doação, 38,1% tinham idade de 20 a 29 anos, 24,5% tinham idade de 30 a 39 anos, 15,3% tinham idade de 40 a 49 anos, 6,1% tinham idade de 50 a 59 anos e 0,3% tinham idade superior a 60 anos. Uma parcela de apenas 0,4% dos doadores é composta por analfabetos, sendo que 27,2% dos doadores tinham 1^o grau incompleto, 17,2% tinham 1^o grau completo ou 2^o grau incompleto, 46,0% tinham 2^o grau completo ou 3^o grau incompleto, e 9,2% tinham 3^o grau completo ou pós-graduação. Quanto ao tipo sanguíneo, 37,6% tinham sangue tipo A; 3,7% sangue tipo AB; 11,2% sangue tipo B e 47,4% sangue tipo O. Do total de doadores, 87,3% eram portadores de fator Rh positivo. O histórico de doação deste grupo de doadores é composto por um total de 346.211 doações, com uma taxa média retorno/doador de 1,996. O número máximo de doações realizadas por um único doador foi 40.

Considerando como a variável de interesse o tempo entre a primeira doação e o próximo retorno, as Figuras 1 e 2 mostram, respectivamente, a função empírica de sobrevivência estimada pelo método de Kaplan-Meier e a função risco, aproximada pelo método atuarial. A função de sobrevivência estima a propensão que o doador de primeira vez possui em não retornar a uma nova doação, em um certo instante t, dado que ele não realizou nova doação até este instante. Nota-se que o gráfico de Kaplan-Meier tem a forma de uma curva suave, ao invés da tradicional forma de escada, devido ao grande número de indivíduos presentes na casuística. Esta curva permance igual a 1 nos primeiros meses, correspondentes ao período em que não é permitida uma nova doação de sangue (o mínimo permitido entre doações é de dois meses), decresce nos meses seguintes, tendendo a estabilizar no final do seguimento. A função risco (Figura 2) tem um pico expressivo aos 4 meses após o início do ciclo de doações, com outros picos em 12 e 24 meses.

4.1 Modelos paramétricos

Na Tabela 1 estão os resultados obtidos para os modelos ajustados com as distribuições de Weibull, log-logística, log-normal e gama generalizada extendida (GGE) considerando uma fração de doadores de primeira vez que não retornaram nula (modelos paramétricos, onde $p = 0$) e não nula (modelos paramétricos de longa duração, onde $0 < p < 1$, assumindo $p = [1 + exp(\beta_0)]^{-1}$). Podemos observar que dentre os modelos considerando fração de doadores de primeira vez nula ($p = 0$), aquele com distribuição GGE apresenta um menor valor para o critério de informação de Schwarz (1978) (ou BIC, *Bayesian information criterion*, apresentado também na Tabela 1), indicando ser este o que traz um melhor ajuste aos dados dentre os quatro modelos. Considerando que o presente estudo considera uma casuística bastante grande, o BIC é mais apropriado que o tradicional critério de Akaike (AIC, *Akaike Information Criterion*). A contribuição relativa do "termo de penalidade" associada ao número de parâmetros na determinação do AIC torna-se pequena quando comparada ao valor da log-verossimilhança ln L, enquanto o BIC inclui o tamanho amostral em seu "termo de penalidade". A Figura 3 mostra as funções de sobrevivência estimadas por modelo sobrepostas às respectivas estimativas de Kaplan-Meier, onde observa-se um bom ajuste para o modelo com distribuição GGE. Os demais modelos superestimam a propensão ao não retorno no início do seguimento e subestimam no final do seguimento. A Figura 4 traz gráficos das funções de sobrevivência estimadas por Kaplan-Meier *versus* as funções de sobrevivência estimadas pelos modelos, onde é reforçada a evidência de que o modelo com distribuição GEE é aquele que melhor se ajusta aos dados.

Também entre os modelos de longa duração, o modelo com distribuição GGE é o que melhor se ajusta aos dados, apresentando menor valor para o BIC (Tabela 1) e com estimativas da propensão ao não retorno mais similares àquelas obtidas empiricamente pelo Kaplan-Meier (Figuras 5 e 6). Sendo β_0 estimado em $1,408$, o parâmetro p é estimado em aproximadamente 19,6% (dado que $\widehat{p} = \left[1 + exp(\widehat{\beta}_0)\right]^{-1}$). No entanto, a interpretação deste parâmetro como a fração

esperada de doadores que não retornam a nova doação é subjetiva, considerando os dados aqui utilizados. Dado que a curva de sobrevida tende a p quando $t \to \infty$, observamos que a curva aproxima-se de 19,6% somente para valores de t maiores de 10 mil dias (ou 27 anos), um tempo demasiadamente grande quando pretende-se estimar o intervalo até uma nova doação. Observando o ajuste do modelo com fração nula de doadores de primeira vez, podemos observar que após aproximadamente quatro anos do início do seguimento, o decréscimo da propensão de retorno já não é mais constante, tendendo a estabilizar muito lentamente. Tomando este período como ponto de corte, temos que a propensão de não retorno estimada após quatro anos de seguimento é de aproximadamente $37,8\%$. Já no final do seguimento (nove anos) é de aproximadamente $31,8\%$, o que corresponde a um aumento do retorno dos doadores para sua segunda tentativa de doação de apenas 6%.

De maneira geral, os modelos de longa duração tiveram um ajuste mais satisfatório que os modelos paramétricos, possibilitando acomodar melhor a alta taxa de eventos censurados e fazendo com que a curva caísse mais suavemente em todos os casos, acompanhando melhor a curva de Kaplan-Meier.

4.2 Modelo de longa duração incluindo um vetor de covariáveis

Para investigar o efeito de covariáveis sobre o comportamento de retorno do doador de sangue, ajustamos modelos baseados nas quatro distribuições descritas na Seção 3.1. Em um primeiro passo, ajustamos modelos de longa duração considerando cada uma das distribuições, e uma covariável de cada vez. Consideramos o parâmetro de escala, λ e o parâmetro p relacionados à covariável em questão. Em cada ajuste, construímos gráficos das funções de sobrevivência estimadas por Kaplan-Meier *versus* as funções de sobrevivência estimadas pelos modelos. Para todas as covariáveis aqui consideradas, estes gráficos (não mostrados) sugeriram que as funções de sobrevivência segundo o modelo GGE são aquelas mais próximas às estimativas de Kaplan-Meier. E ainda, comparando os ajustes dos modelos entre as quatro diferentes distribuições de probabilidade, notamos que o critério de informação de Schwarz (BIC) é menor quando utilizada a distribuição GGE.

Em um segundo passo, ajustamos modelos de longa duração considerando todas as covariáveis simultaneamente. Outra vez, consideramos tanto o parâmetro de escala, λ, quanto o parâmetro responsável pela fração de doadores de primeira vez, p, relacionados a um vetor de covariáveis, assumindo $\mathbf{x}$ e $\mathbf{z}$ presentes na equação (1) iguais. Por simplicidade, denotaremos $\lambda(\mathbf{x})$ e $p(\mathbf{x})$, onde as covariáveis em questão são: sexo (masculino ou feminino), cor da pele (negra, branca ou amarela), faixa etária (18 a 19 anos completos, 20 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 59 anos e 60 anos ou mais), estado civil (casado, viúvo, solteiro, divorciado ou outros), tipo sanguíneo segundo classificação ABO (A, B, AB e O) e fator Rh (positivo ou negativo). Todas estas variáveis categóricas foram transformadas em variáveis indicadoras (assumindo 0 e 1). Desta maneira, ajustamos um modelo para cada uma das quatro distribuições aqui apresentadas. Novamente, o modelo baseado na distribuição GGE apresentou evidências de um melhor ajuste, apresentado um

valor para o critério de informação de Schwarz menor que o observado quando consideradas as demais distribuições (BIC igual a 276467,7, contra 292548,8 para o modelo log-logístico e 289738,8 para o modelo lognormal). Os resultados obtidos com a macro *gfcure* (Peng, Dear e Deham, 1998) implementada no programa R, considerando o modelo GGE, se encontram na Tabela 2. O erro padrão do parâmetro de forma (φ) da distribuição GGE, segundo Peng, Dear e Deham (1998), tem pouca utilidade prática. Portanto, não é estimado pela macro *gfcure*.

Observando as estimativas dos parâmetros relacionados com $\lambda(\mathbf{x})$ (Tabela 2), nota-se que a variável sexo apresenta um maior efeito, dada a magnitude de sua estimativa. Sendo este coeficiente negativo $(-0,152)$, dentre as pessoas suscetíveis ao evento (retorno a uma nova doação), há uma expectativa de que os homens retornem a uma nova doação em um espaço de tempo inferior àquele esperado para as mulheres. Às demais variáveis estão associados parâmetros com estimativas de magnitudes relativamente pequenas, próximas a zero, evidenciando que não estão relacionadas ao tempo de retorno do doador. Observando as estimativas dos parâmetros relacionados com $p(\mathbf{x})$, as variáveis cor da pele, faixa etária, estado civil e fator Rh apresentam maiores efeitos de não retorno do doador. Assim, entre os doadores que não retornaram, doadores declarados como negros ou brancos tendem a ter uma fração esperada de não retorno menor do que doadores declarados da cor amarela. Também observamos que à medida que a faixa etária aumenta a proporção de não retorno também aumenta. Notamos que indivíduos solteiros têm uma menor propensão ao retorno. Doadores com fator Rh positivo também apresentam uma proporção de não retorno esperada maior quando comparados com doadores com fator Rh positivo.

5 Discussão

Na metade do último século, Boag (1949) escreveu um estudo onde eram propostas estimativas de máxima verossimilhança para a proporção de pacientes que deixaram de portar o câncer de mama após uma terapia. A partir daí, muitos autores vêm desenvolvendo modelos voltados à análise de dados de sobrevivência sujeitos à censura, onde estão presentes uma parcela de indivíduos para os quais não ocorrerá o evento de interesse ("*long-term survivors*"). Por exemplo, em um estudo recente, Shao e Zhou (2004) desenvolveram um novo modelo paramétrico baseado na distribuição de Burr XII, de três parâmetros, com função distribuição acumulada $F(t) = 1 - [1 + (t/\phi)^{\alpha}]^{-\gamma}$, onde α, γ, e ϕ são maiores que zero. Esta distribuição é muito mais flexível que a distribuição Weibull, sendo as distribuições Weibull e Pareto casos específicos da distribuição de Burr XII.

A situação onde parte dos indivíduos é imune a um evento ocorre em muitos estudos da área médica. Por exemplo, em um estudo sobre a resposta a um tratamento oncológico, onde o evento de interesse é o óbito devido ao câncer, uma parcela dos indivíduos poderia estar curada da doença (Haybittle, 1959). No presente estudo, o evento de interesse é o retorno para uma nova doação de sangue, dado que o indivíduo compareceu a um serviço de hematologia e hemoterapia para

uma primeira doação. Como é observado que uma parcela dos indivíduos não retorna para novas doações, os modelos de sobrevivência de longa duração tornam-se adequados à modelagem dos dados.

Para a modelagem do tempo entre a primeira doação de sangue e o seguinte retorno, utilizamos modelos de sobrevivência de longa duração, considerando as distribuições Weibull, log-normal, log-logística e gama generalizada. Consideramos ainda que o parâmetro de escala e a proporção de indivíduos que não retornam a novas doações são dependentes de um vetor de covariáveis. Observamos que os modelos de longa duração, especialmente o modelo baseado na distribuição gama generalizada estendida, apresentaram uma grande habilidade para a representação do comportamento de retorno do doador de sangue. A modelagem permite previsões sobre a probabilidade de retorno de um doador de primeira vez, dentro de um determinado período, resultado este de grande utilidade para estratégias de fidelização de doadores de sangue.

Agradecimentos

Agradecemos o auxílio de Amilton Gomes de Brito, gerente de informática do Centro de Processamento de Dados do Hemocentro de Ribeirão Preto, pela montagem do banco de dados utilizado na presente pesquisa. Agradecemos ainda o apoio recebido do Prof. Dr. Dimas Tadeu Covas, diretor científico do Hemocentro de Ribeirão Preto. A pesquisa de Adriana de Fátima Lourençon recebeu auxílio financeiro da Fundação de Apoio ao Ensino, Pesquisa e Assistência, do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (FAEPA, HCFMRP/USP). O projeto de pesquisa que originou este estudo foi analisado e aprovado pelo Comitê de Ética em Pesquisa do HCFMRP/USP.

LOURENÇON, A. de F.; MARTINEZ, E. Z .; MAZUCHELI, J.; FERREIRA, O. Long-term survival models applied to a study on the return behavior of voluntary blood donors. *Rev. Mat. Estat.,* São Paulo, v.25, n.1, p.137-154, 2007.

- *ABSTRACT: Strategies to ensure safe supplies of blood in hemotherapy and transfusion services, stimulating the return of voluntary donors, depend on the knowledge of the return behavior of blood donors. In the present article, we use long-term survival models for modeling the time between the first donation of blood and the next return, considering distributions Weibull, log-normal, log-logistic and generalized gamma. We consider that the parameter of scale and the proportion of individuals that do not return for new donations are dependends on a vector of covariates. We observe that the long-term survival models, especially the model based on the generalized gamma extended distribution, adequately describe the return behavior of the blood donor.*

- *KEYWORDS: Survival analysis; long-term survival model; generalized gamma distribution; blood donors; biostatistics.*

Referências

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc., Series B*, London, v.11, p.15-45, 1949.

FAREWELL, V. T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, v.38, n.4, p.1041-1046, 1982.

HAYBITTLE, J. L. The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br. J. Radiol.*, London, v.32, p.725-733, 1959.

JAMES, R. C.; MATTHEWS, D. E. Analysis of blood donor return behaviour using survival regression methods. *Transf. Med.*, Oxford, v.6, n.1, p.21-30, 1996.

LAM, K. F.; FONG, D. Y.; TANG, O. Y. Estimating the proportion of cured patients in a censored sample. *Stat. Med.*, New York, v.24, n.12, p.1865-1879, 2005.

LEE, E. T.; WANG, J. W. *Statistical methods for survival data analysis*. Chichester: John Wiley, 2003.

LUDWIG, S. T.; RODRIGUES, A. C. M. Doação de sangue: uma visão de marketing. *Cad. Saúde Pública*, Rio de Janeiro, v.21, n.3, p.932-939, 2005.

MALLER, R. A.; ZHOU, X. *Survival analysis with long-term survivors*. Chichester: John Wiley, 1996.

MAZUCHELI, J.; LOUZADA-NETO, F.; ACHCAR, J. A. Lifetime models with nonconstant shape parameters. *Rev. Stat*, n.1, p.25-39, 2003.

OWNBY, H. E.; WATANABE, Y. T.; NASS, C. C. Analysis of donor return behavior. *Transfusion*, Paris, v.39, n.10, p.1128-1135, 1999.

PENG, Y.; DEAR, K. B.; DENHAM, J. W. A generalized F mixture model for cure rate estimation. *Stat. Med.*, New York, v.17, n.8, p.813-830, 1998.

PERPEROGLOU, A; KERAMOPOULLOS, A.; VAN HOUWELINGEN, H. V. Approaches in modelling long-term survival: an application to breast cancer. *Stat. Med.*, New York, v.26, n.13, p.2666-2685, 2007.

SHAO, Q; ZHOU, X. A new parametric model for survival data with long-term survivors. *Stat. Med.*, New York, v.23, n.22, p.3525-3543, 2004.

SCHWARZ, G. Estimating the dimension of a model. *Ann. Stat.*, Haryward, v.6, n.2, p.461-464, 1978.

Recebido em 13.02.2007.

Aprovado após revisão em 15.08.2007.

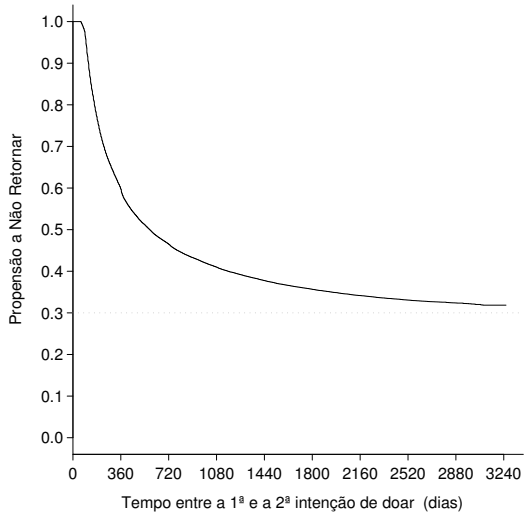

Figura 1 - Curva de Kaplan-Meier, para o tempo entre a primeira doação e o próximo retorno do doador de sangue.

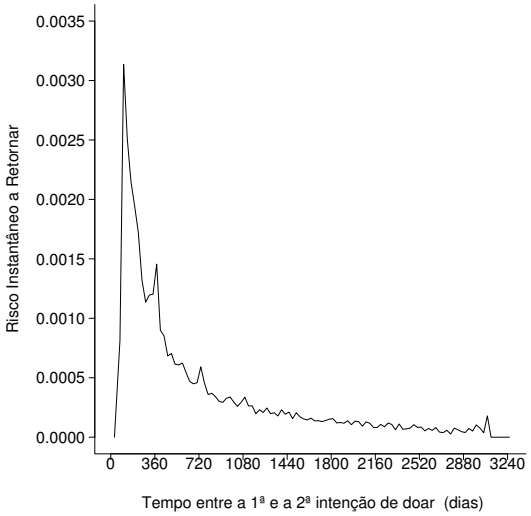

Figura 2 - Função risco instantâneo. Picos expressivos correspondem aproximadamente a 4, 12 e 24 meses após o início do ciclo de doação.

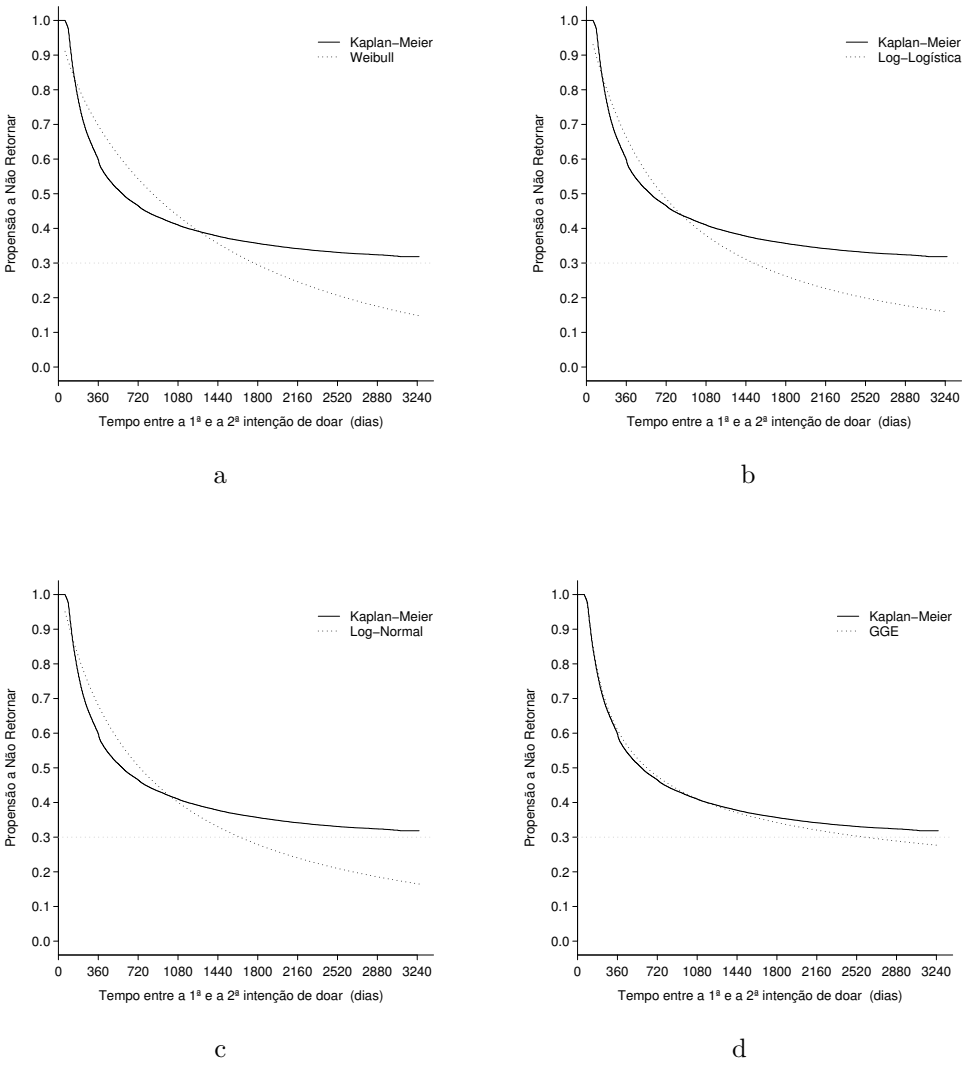

Figura 3 - Curvas de Sobrevivência estimadas pelo método de Kaplan-Meier e por modelos paramétricos assumindo as distribuições de Weibull (a), log-logística (b), log-normal (c) e gama generalizada (d).

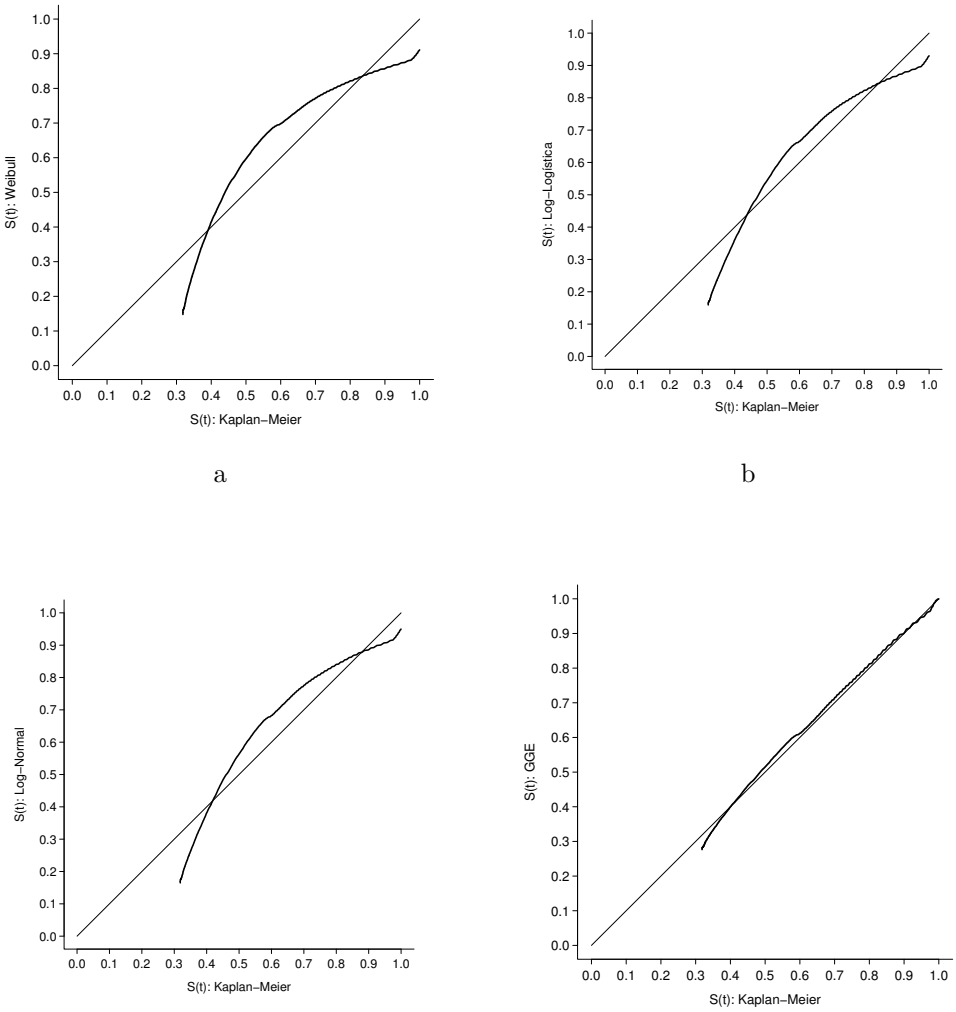

Figura 4 - Gráficos das funções de sobrevivência estimadas por Kaplan-Meier *versus* as funções de sobrevivência estimadas pelos modelos assumindo distribuição Weibull (a), log-logística (b), log-normal (c) e gama generalizada (d).

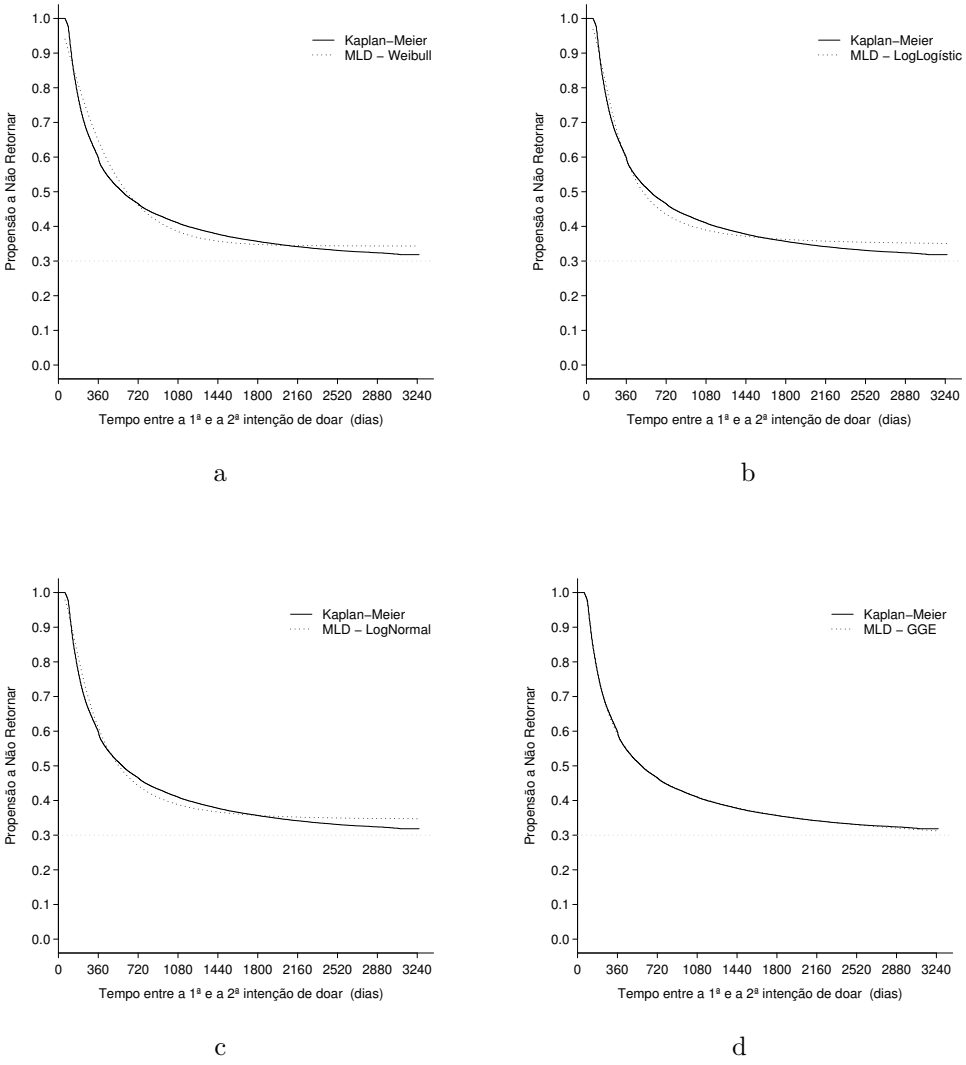

Figura 5 - Curvas de Sobrevivência estimadas pelo método de Kaplan-Meier e por modelos de longa duração (MLD), assumindo as distribuições de Weibull (a), log-logística (b), log-normal (c) e gama generalizada (d).

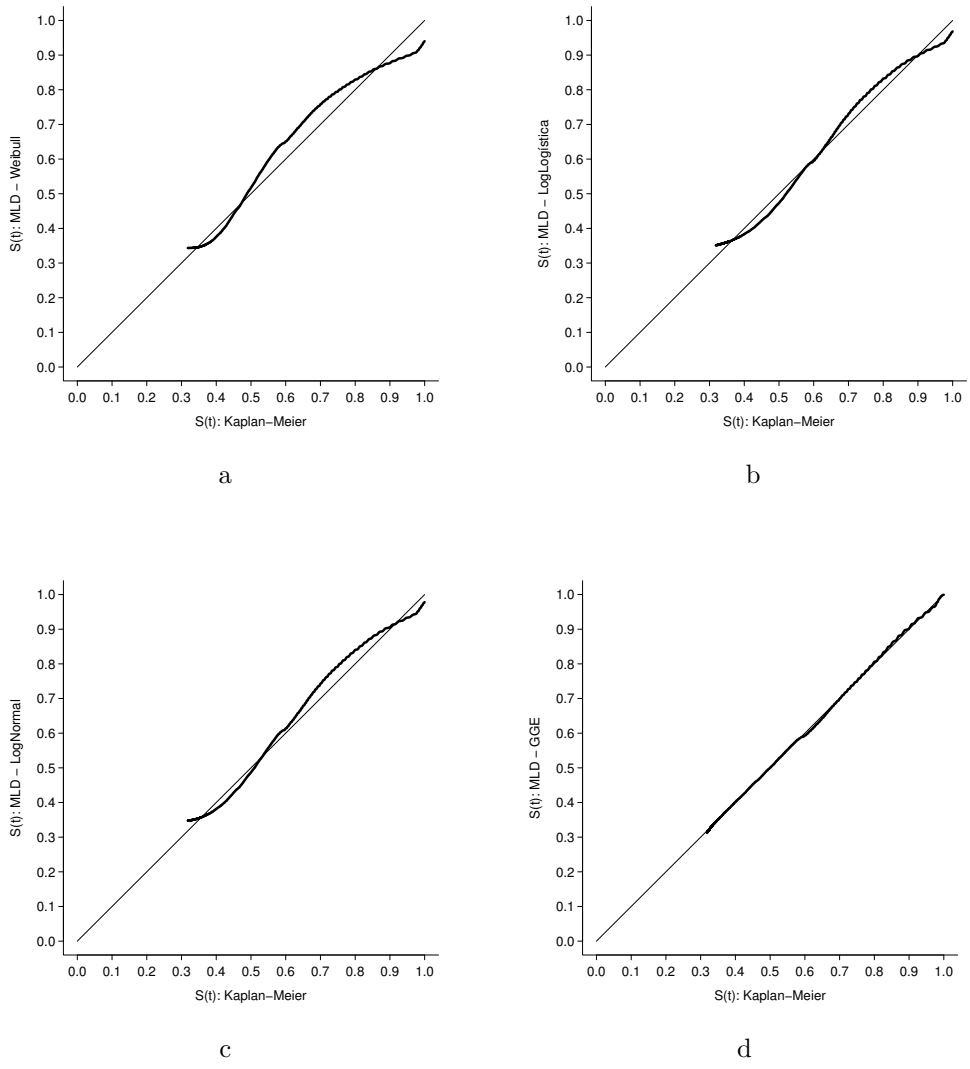

Figura 6 - Gráficos das funções de sobrevivência estimadas por Kaplan-Meier *versus* as funções de sobrevivência estimadas pelos modelos de longa duração (MLD) assumindo distribuição Weibull (a), log-logística (b), log-normal (c) e gama generalizada (d).

Tabela 1 - Estimativas dos parâmetros e comparação dos modelos pelo critério de informação de Schwarz (1978) (ou BIC), *Bayesian information criterion*

	Estimativa	Erro padrão	Log-Verossimilhança	BIC
I - Modelos paramétricos				
a) Weibull			-534271,7	1068566,7
μ	7,231	0,0053		
σ	1,322	0,0042		
b) Log-logística			-527069,4	1054162,1
μ	6,524	0,0053		
σ	0,940	0,0030		
c) Log-normal			-524213,1	1048449,5
μ	6,601	0,0052		
σ	1,525	0,0046		
d) GGE			-507889,9	1015814,8
μ	4,961	0,0064		
σ	0,790	0,0047		
φ	-3,521	0,0268		
II - Modelos de longa duração				
a) Weibull			-522156,0	1044347,0
μ	6,114	0,0040		
σ	0,859	0,0026		
β_0	0,648	0,0073		
b) Log-logística			-514976,5	1029988,0
μ	5,629	0,0039		
σ	0,515	0,0019		
β_0	0,639	0,0075		
c) Log-normal			-513499,4	1027033,8
μ	5,680	0,0038		
σ	0,864	0,0029		
β_0	0,639	0,0074		
d) GGE			-507452,9	1014952,4
μ	5,006	0,0070		
σ	0,709	0,0408		
φ	-2,528	0,0044		
β_0	1,408	0,0307		

Tabela 2 - Estimativas dos parâmetros do modelo de regressão de longa duração com distribuição gama generalizada e covariáveis relacionadas com o perfil do doador

		Estimativa	Erro Padrão
Parâmetro de Forma (φ)		-2,819	
Parâmetro de Escala ($\ln \sigma$)		-0,407	0,006
Parâmetros relacionados com $\lambda(\mathbf{x})$			
	Intercepto (ζ_0)	5,084	0,053
1-Sexo	(Masculino *vs* Feminino) (ζ_1)	-0,152	0,005
2-Cor da Pele	(Negra *vs* Amarela) (ζ_2)	-0,026	0,033
	(Branca *vs* Amarela) (ζ_3)	0,014	0,033
3-Faixa Etária	(18 a 19 anos *vs* 60 a 65 anos) (ζ_4)	-0,093	0,042
	(20 a 29 anos *vs* 60 a 65 anos) (ζ_5)	-0,040	0,042
	(30 a 39 anos *vs* 60 a 65 anos) (ζ_6)	-0,009	0,042
	(40 a 49 anos *vs* 60 a 65 anos) (ζ_7)	0,007	0,042
	(50 a 59 anos *vs* 60 a 65 anos) (ζ_8)	-0,014	0,043
4-Estado Civil	(Casado *vs* Solteiro) (ζ_9)	0,001	0,006
	(Divorciado *vs* Solteiro) (ζ_{10})	-0,046	0,011
	(Outros *vs* Solteiro) (ζ_{11})	-0,053	0,016
	(Viúvo *vs* Solteiro) (ζ_{12})	-0,061	0,021
5-Tipo Sanguíneo	(AB *vs* A) (ζ_{13})	-0,049	0,012
	(B *vs* A) (ζ_{14})	-0,013	0,008
	(O *vs* A) (ζ_{15})	-0,006	0,005
6-Fator Rh	(Negativo *vs* Positivo) (ζ_{16})	-0,027	0,007
Parâmetros relacionados com com $p(\mathbf{x})$			
	Intercepto (β_0)	-1,169	0,202
1-Sexo	(Masculino *vs* Feminino) (β_1)	-0,066	0,029
2-Cor da Pele	(Negra *vs* Amarela) (β_2)	0,913	0,123
	(Branca *vs* Amarela) (β_3)	0,753	0,120
3-Faixa Etária	(18 a 19 anos *vs* 60 a 65 anos) (β_4)	1,809	0,165
	(20 a 29 anos *vs* 60 a 65 anos) (β_5)	1,422	0,162
	(30 a 39 anos *vs* 60 a 65 anos) (β_6)	1,459	0,163
	(40 a 49 anos *vs* 60 a 65 anos) (β_7)	1,332	0,163
	(50 a 59 anos *vs* 60 a 65 anos) (β_8)	1,166	0,167
4-Estado Civil	(Casado *vs* Solteiro) (β_9)	0,731	0,036
	(Divorciado *vs* Solteiro) (β_{10})	0,916	0,086
	(Outros *vs* Solteiro) (β_{11})	0,951	0,147
	(Viúvo *vs* Solteiro) (β_{12})	0,696	0,140
5-Tipo Sanguíneo	(AB *vs* A) (β_{13})	0,195	0,077
	(B *vs* A) (β_{14})	0,072	0,046
	(O *vs* A) (β_{15})	0,108	0,029
6-Fator Rh	(Negativo *vs* Positivo) (β_{16})	0,423	0,049