Analise de sobrevivência

Tempo até a ocorrência da tuberculose

Eliana, Luanna, Barbara, Maria Cecília, Sofia, Victor

03/02/2025

Índice

1 Introdução

Os dados utilizados neste projeto referem-se a um estudo sobre o tempo até a ocorrência de doenças oportunistas em uma coorte de pacientes HIV positivos atendidos em um Hospital Universitário. As variáveis foram obtidas a partir de prontuários clínicos. Para cada paciente, registrou-se o tempo até a ocorrência de algumas doenças ou sintomatologias caracteristicamente relacionadas à imunodepressão, como candidíase, tuberculose, sinais hematológicos, herpes zoster, pneumonia por Pneumocistis.

O banco de dados é composto por 11 variáveis: nove covariáveis (oito categóricas e uma numérica), o tempo de acompanhamento e uma variável indicadora de ocorrência de tuberculose. No estudo, o tempo até a ocorrência da tuberculose foi registrado como variável resposta, com censura em casos onde as pacientes não foram acompanhadas até o surgimento da doença (Ver Tabela 1.1).

Tabela 1.1: Descrição das variáveis utilizadas no estudo sobre tuberculose

Variável	Descrição
Sexo	1 - Masculino, 2 - Feminino.
Escolaridade	0 - Sem escolaridade, 1 - Até quatro anos de estudo, 2 - Ensino fundamental, 3 - Ensino médio, 4 - Ensino superior.
Idade	Idade em anos na entrada do estudo.
Uso de drogas injetáveis	0 - Não, 1 - Sim.
Status da doença	0 - Censura, 1 - Ocorrência da doença.
Tempo até a ocorrência	Tempo até a ocorrência da doença Tuberculose.
Candidíase	0 - Não, 1 - Sim.
Sinais hematológicos	0 - Não, 1 - Sim.
Herpes zoster	0 - Não, 1 - Sim.
Pneumonia	0 - Não, 1 - Sim.
Tuberculose	0 - Não, 1 - Sim.

2 Metodologia

Primeiramente, foi realizada uma análise descritiva das variáveis em estudo. Na análise de sobrevivência, essa etapa consiste em utilizar métodos não-paramétricos. Quase todas as covariáveis são dicotômicas, e, portanto, foi possível construir as estimativas de Kaplan-Meier para comparar as duas categorias. Isso foi feito para as 8 covariáveis categóricas, e também foi testada a hipótese de igualdade das duas curvas utilizando os testes de Wilcoxon e log-rank. Além disso, foi analisado se essas covariáveis atendem à suposição de riscos proporcionais.

A variável "idade" foi analisada utilizando o modelo de Cox para verificar a presença de risco proporcional. Ela também foi estratificada para análise de diferentes faixas etárias.

A próxima etapa da análise consistiu em modelar separadamente cada uma das covariáveis com a variável resposta. O objetivo dessa etapa foi selecionar as variáveis explicativas (covariáveis) que devem prosseguir para a modelagem. O critério utilizado neste trabalho foi manter as variáveis que apresentaram valores de p inferiores a 0,25 em pelo menos um dos testes de Wilcoxon e log-rank na comparação das curvas de sobrevivência.

No modelo de Cox, as variáveis incluídas no modelo inicial foram aquelas que apresentaram significância estatística no teste de Wilcoxon ou log-rank, além de atenderem ao pressuposto de riscos proporcionais. Após essa seleção inicial, foi realizado um ajuste passo a passo, no qual a variável com o maior p-valor foi removida iterativamente, até que o modelo final contivesse apenas variáveis estatisticamente significativas. Esse processo garantiu um modelo mais parcimonioso e robusto, mantendo apenas os preditores mais relevantes para a análise da sobrevivência.

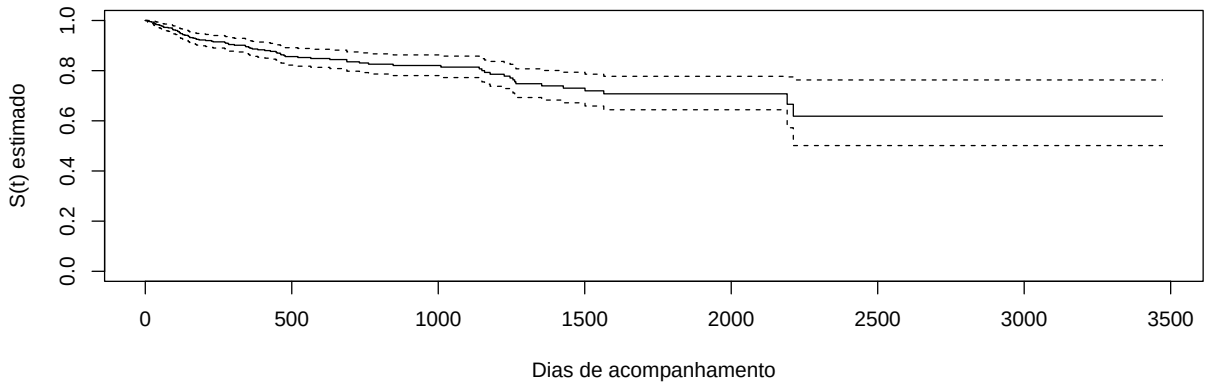
No modelo paramétrico, foi realizado um teste de comparação entre os modelos de gama generalizada, lognormal, Weibull e exponencial para escolher o melhor modelo que se ajustasse aos dados. Após essa análise, foi utilizado o método de backward selection para escolher o modelo final com as variáveis que mais explicam o tempo até a ocorrência da tuberculose.

Antes de proceder à interpretação das estimativas dos parâmetros do modelo ajustado, foram analisados os resíduos para confirmar a adequação do modelo final escolhido, tanto para o modelo paramétrico quanto para o semi-paramétrico.

3 Resultados

3.1 Analise Descritiva e Exploratória

Figura 3.1: Curva de Kaplan-Meier para tuberculose



O gráfico mostra a evolução da probabilidade de não desenvolver tuberculose ao longo do tempo (Ver Figura 3.1). Com o tempo, a probabilidade de"sobreviver" (ou seja, de não desenvolver tuberculose) diminui à medida que mais indivíduos são diagnosticados com a doença. Esse gráfico ajuda a identificar em que momento os casos de tuberculose se acumulam mais rapidamente ou se há períodos de maior risco.

A curva de Kaplan-Meier para o sexo mostra a probabilidade de não desenvolver tuberculose ao longo do tempo, separada entre homens e mulheres (Ver Figura 3.2). No gráfico, a curva das mulheres (vermelha) está acima da curva dos homens (azul), indicando que as mulheres têm uma maior probabilidade de não desenvolver a doença ao longo do tempo, ou seja, elas permanecem "saudáveis" por mais tempo. Em contraste, os homens têm um risco maior, com a curva masculina caindo mais rapidamente, sugerindo que a probabilidade de desenvolver tuberculose é maior entre eles. Esse padrão pode indicar que o sexo masculino está associado a um risco elevado de tuberculose, enquanto o sexo feminino seria um fator de proteção. Para a variável sex0, o p-valor = 0.86, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

O gráfico de Kaplan-Meier mostra as curvas de sobrevivência estratificadas por níveis de escolaridade (Ver Figura 3.3), indicando que grupos com maior escolaridade (principalmente o ensino superior) tendem a ter maior sobrevida, enquanto o grupo sem escolaridade apresenta uma curva mais baixa, embora com poucas observações (7), o que pode limitar a confiabilidade dessa estimativa. As curvas não violam o pressuposto de riscos proporcionais, permitindo o uso adequado do modelo de Cox

Figura 3.2: Curvas de Kaplan-Meier para sexo

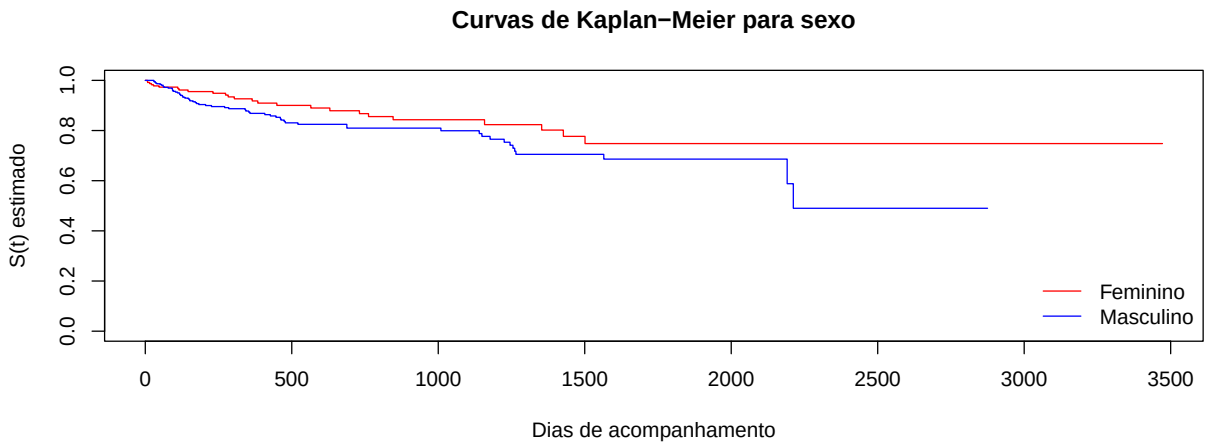
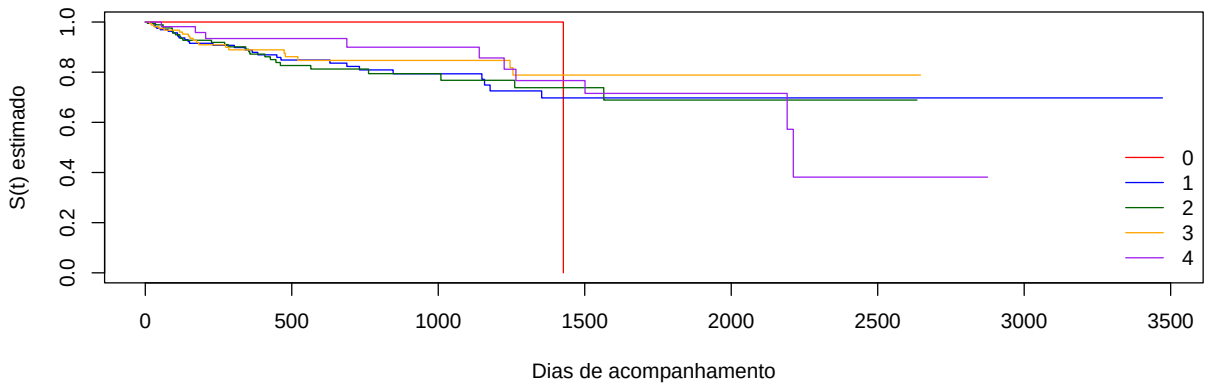
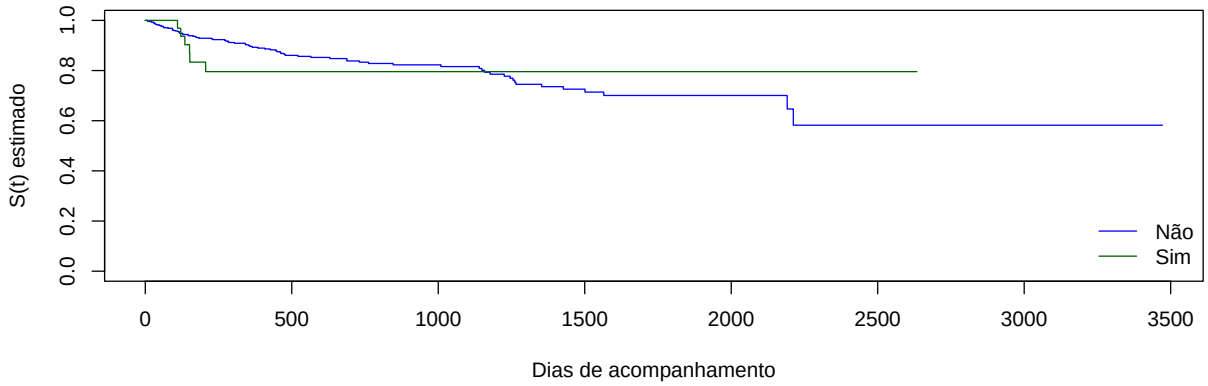


Figura 3.3: Curvas de Kaplan-Meier para escolariedade



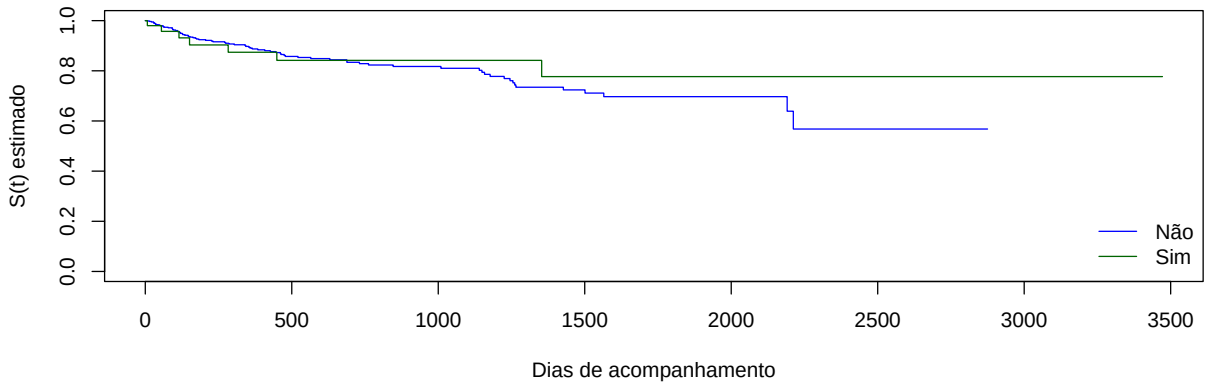
para análise mais detalhada.Para a variável escolaridade, o p-valor = 0.31, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

Figura 3.4: Curvas de Kaplan-Meier para Uso de drogas injetáveis



O gráfico de Kaplan-Meier apresenta as curvas de sobrevivência para indivíduos que usam ou não drogas injetáveis (Ver Figura 3.4). Observa-se que o grupo que não usa drogas injetáveis (linha azul) possui uma probabilidade de sobrevivência maior ao longo do tempo em comparação ao grupo que usa (linha verde), até o dia 1100, aproximadamente. Depois disso, as curvas são invertidas. A curva daqueles que usam drogas injetáveis se estabiliza perto do dia 200, uma vez que, entre aqueles que são usuários, o último que teve ocorrência da doença foi no dia 206. Apesar disso, as curvas não violam o pressuposto de riscos proporcionais, permitindo o uso do modelo de Cox para análise adicional. Para a variável escolaridade, o p-valor = 0.092, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

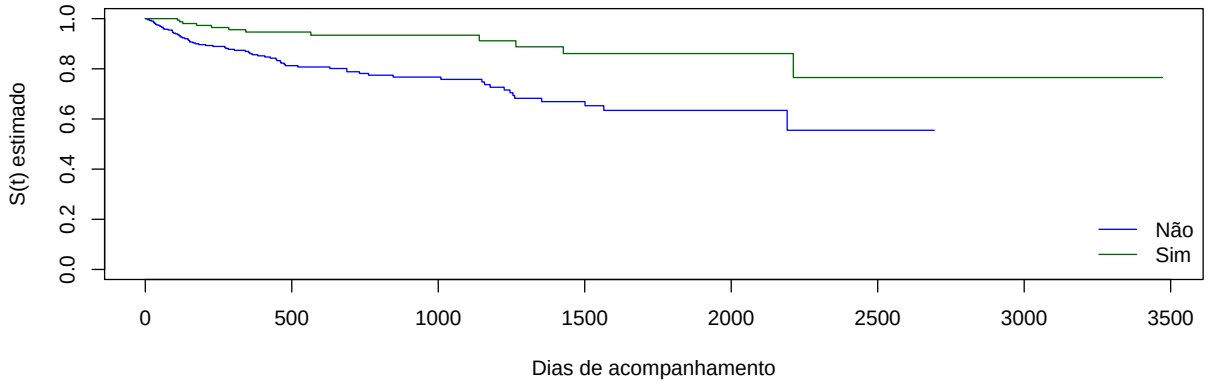
Figura 3.5: Curvas de Kaplan-Meier para sexualidade



O gráfico de Kaplan-Meier apresenta as curvas de sobrevivência para indivíduos de acordo com a orientação sexual (Ver Figura 3.5).. Observa-se que o grupo de heterossexuais (linha azul)

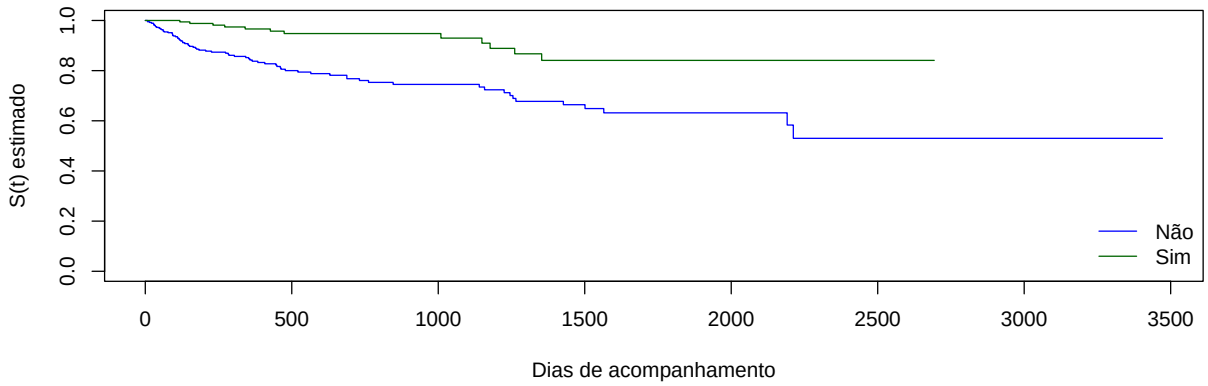
possui uma probabilidade de sobrevivência parecida ao longo do tempo em comparação aos não heterossexuais (linha verde), até o dia 600, aproximadamente. Depois disso, as curvas são invertidas. Para a variável sexualidade, o p-valor = 0.16, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

Figura 3.6: Curvas de Kaplan-Meier para Candidíase



A curva daqueles com candidíase (verde) está acima daqueles sem (azul), indicando que pessoas com candidíase têm uma maior probabilidade de não desenvolver a doença ao longo do tempo, ou seja, elas permanecem "saudáveis" por mais tempo (Ver Figura 3.6). Aqueles sem candidíase têm um risco maior, com a curva caindo mais rapidamente, sugerindo que a probabilidade de desenvolver tuberculose é maior entre eles. O pressuposto de risco proporcional é atendido. Para a variável Candidíase o p-valor = 0.25, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

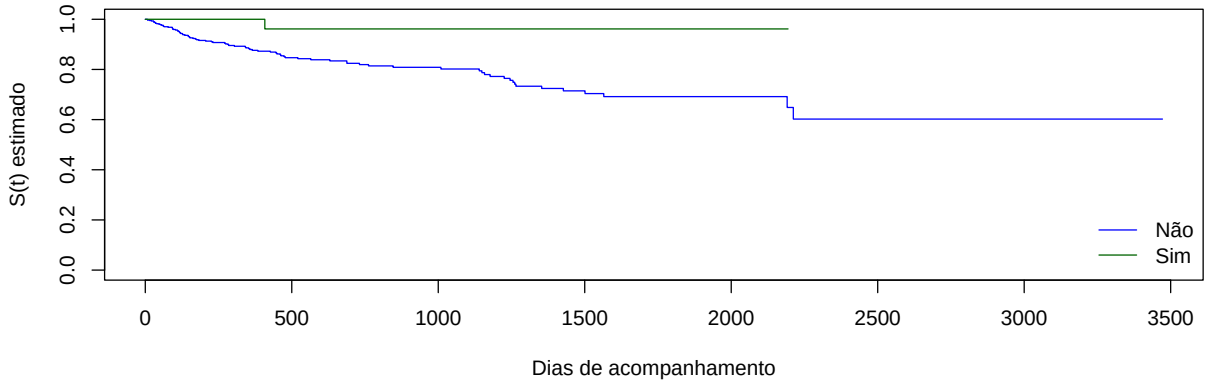
Figura 3.7: Curvas de Kaplan-Meier para Sinais hematológicos



A curva daqueles com sinais hematológicos (verde) está acima daqueles sem (azul), indicando que pessoas com sinais hematológicos têm uma maior probabilidade de não desenvolver a doença ao

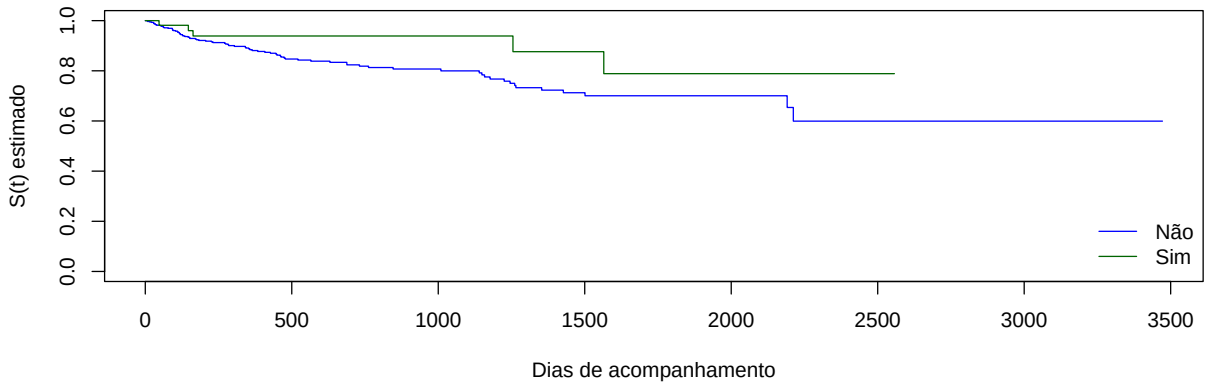
longo do tempo, ou seja, elas permanecem “saudáveis” por mais tempo (Ver Figura 3.7). Aqueles sem sinais hematológicos têm um risco maior, com a curva caindo mais rapidamente, sugerindo que a probabilidade de desenvolver tuberculose é maior entre eles.Para a variável Sinais hematológicos, o p-valor = 0.045, indicando que há evidência de violação da suposição de proporcionalidade dos riscos.

Figura 3.8: Curvas de Kaplan-Meier para herpes



A curva daqueles com herpes (verde) está acima daqueles sem (azul), indicando que pessoas com herpes têm uma maior probabilidade de não desenvolver a doença ao longo do tempo, ou seja, elas permanecem “saudáveis” por mais tempo (Ver Figura 3.8). Aqueles sem herpes têm um risco maior, com a curva caindo mais rapidamente, sugerindo que a probabilidade de desenvolver tuberculose é maior entre eles. Para a variável herpes, o p-valor = 0.72, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

Figura 3.9: Curvas de Kaplan-Meier para Pneumonia

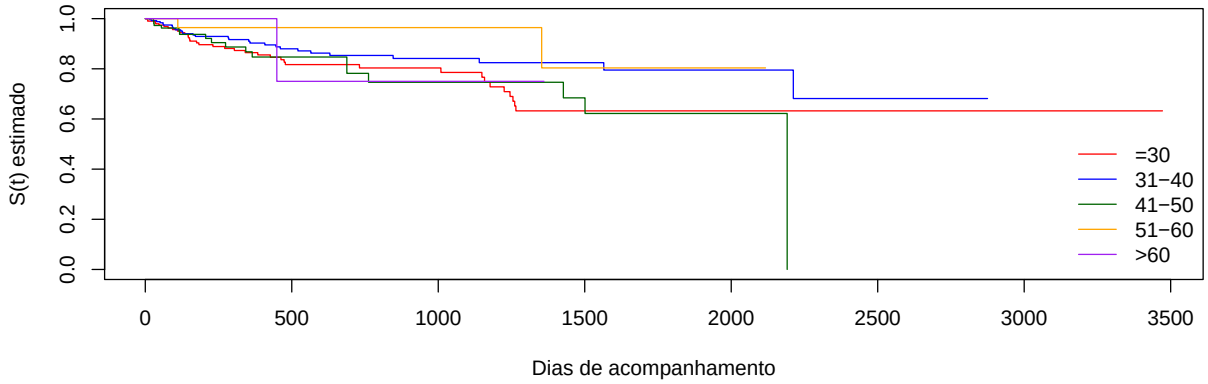


A curva daqueles com pneumonia (verde) está acima daqueles sem (azul), indicando que pessoas com pneumonia têm uma maior probabilidade de não desenvolver a doença ao longo do tempo, ou

seja, elas permanecem “saudáveis” por mais tempo (Ver Figura 3.9). Aqueles sem pneumonia têm um risco maior, com a curva caindo mais rapidamente, sugerindo que a probabilidade de desenvolver tuberculose é maior entre eles. Para a variável pneumonia, o p-valor = 0.66, indicando que não há evidência de violação da suposição de proporcionalidade dos riscos.

O teste cox.zph foi aplicado para verificar se o pressuposto de riscos proporcionais é atendido no modelo de Cox para a variável idade. A violação desse pressuposto indica que o efeito da idade sobre o risco de desenvolver tuberculose não é constante ao longo do tempo. Ou seja, a relação entre a idade e o risco de tuberculose varia durante o período de acompanhamento, o que invalida o modelo de Cox simples. Isso sugere que, para modelar adequadamente os dados, pode ser necessário ajustar o modelo por meio de estratificação.

Figura 3.10: Curvas de Kaplan-Meier para idade estratificada



O gráfico de Kaplan-Meier ilustra as curvas de sobrevivência para diferentes faixas etárias (30, 31–40, 41–50, 51–60, >60 anos) ao longo do tempo de acompanhamento (Ver Figura 3.10). Observa-se que o grupo mais jovem (30 anos, linha vermelha) apresenta a maior redução na probabilidade de sobrevivência nos primeiros 1.000 dias, indicando um risco mais elevado de desfecho adverso nesse período. Por outro lado, os grupos mais velhos (>60 anos, linha roxa, e 51–60 anos, linha verde) apresentam melhores probabilidades de sobrevivência, com declínios mais graduais ao longo do tempo. Essa tendência sugere que a idade está associada ao risco de desfecho, com indivíduos mais jovens apresentando maior vulnerabilidade inicial. Apesar disso, as curvas não violam o pressuposto de riscos proporcionais,p= 0.55, permitindo o uso do modelo de Cox para análise adicional.

Tabela 3.1: Testes logrank e de Wilcoxon para igualdade das curvas de sobrevivência.

Covariável	Logrank (valor p)	Wilcoxon (valor p)
Sexo	3 (0.083)	2.92 (0.088)
EscolarIdade	0.99 (0.911)	1.04 (0.904)
Uso de Drogas Injetáveis	0.01 (0.936)	0.07 (0.795)
Sexoual	0.38 (0.537)	0.23 (0.63)
Candidíase	16.84 (0)	17.28 (0)
Herpes Zoster	5.28 (0.022)	5.3 (0.021)

Pneumonia Pneumocystis	2.43 (0.119)	2.5 (0.114)
Sinais Hematológicos	22.83 (0)	23.87 (0)
Faixa Etária	6.61 (0.158)	6.37 (0.173)
Idade	45.5 (0.691)	44.12 (0.741)

Antes do ajuste desses modelos, será discutido um passo essencial na análise estatística: a seleção de variáveis para entrar no modelo inicial. O teste de Logrank é mais sensível às diferenças nas taxas de risco constantes ao longo do tempo, enquanto o Wilcoxon dá mais peso às diferenças iniciais. A consistência entre os dois testes reforça os resultados significativos encontrados.

Com base nos resultados apresentados na Tabela 3.1, verifica-se que as covariáveis: sexo, Candidíase, Herpes Zoster, Pneumonia, Sinais Hematológicos, Faixa Etária atenderam ao critério estabelecido e, portanto, serão incluídas na etapa de modelagem estatística. No entanto, a covariável Sinais Hematológicos não vai entrar no modelo inicial de cox pois o pressuposto de risco proporcional não foi atendido.

3.2 Modelo de Cox

Tabela 3.2: Resultados do Modelo de Cox

Variável	Coeficiente	exp(Coef)	Erro Padrão	Z	P-valor
factor(sex)M	0.4718	1.6029	0.2414	1.954	0.0507
Faixa_idade 31-40	-0.4431	0.6420	0.2512	-1.764	0.0777
Faixa_idade 41-50	0.0310	1.0315	0.2965	0.105	0.9167
Faixa_idade 51-60	-1.0720	0.3423	0.7279	-1.473	0.1408
Faixa_idade >60	-0.6090	0.5439	1.0149	-0.600	0.5485
Candidíase	-1.4697	0.2300	0.3157	-4.656	0.0000
Herpes	-2.4475	0.0865	1.0082	-2.427	0.0152
Pneumonia	-1.2607	0.2835	0.4659	-2.706	0.0068

O primeiro modelo de regressão de Cox indicou que a idade não teve impacto significativo no risco, enquanto as comorbidades apresentaram forte associação com uma redução do risco. Com isso, ajustamos um novo modelo sem essa variável.

Tabela 3.3: Resultados do Modelo de Cox

Variável	Coeficiente	exp(Coef)	Erro Padrão	Z. valor	P-valor
Sexo(Masculino)	0.46925	1.59879	0.23919	1.962	0.04979
Candidíase	-1.49596	0.22403	0.31489	-4.751	0.00000
Herpes	-2.42199	0.08874	1.00775	-2.403	0.01625
Pneumonia	-1.25609	0.28477	0.46506	-2.701	0.00692

Os resultados do modelo de Cox indicam que o sexo masculino está associado a um risco significativamente maior de desenvolver tuberculose, com uma razão de risco (HR) de 1,60 , ou

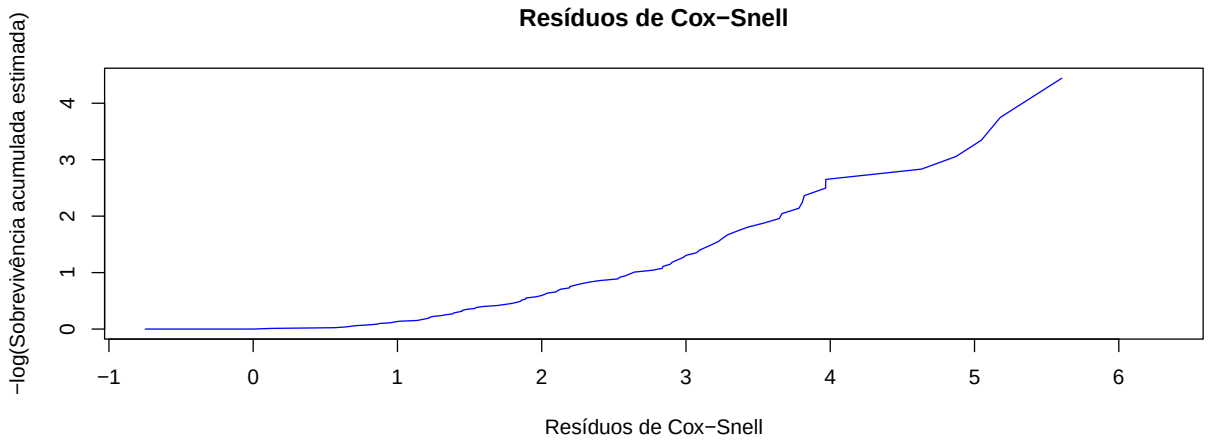
seja, homens têm aproximadamente 59,9% mais chance de desenvolver tuberculose em comparação com mulheres (p = 0.04979).

Por outro lado, a presença de algumas comorbidades parece estar associada a um menor risco de tuberculose. Indivíduos com candidíase apresentam uma redução de 77,6% no risco da doença (HR = 0,22, p < 0.0001), enquanto aqueles com herpes têm uma redução de 91,1% no risco (HR = 0,089, p = 0.01625). Da mesma forma, a pneumonia foi associada a uma redução de 71,5% no risco de tuberculose (HR = 0,285, p = 0.00692).

Tabela 3.4: Análise de resíduos de Schoenfeld

	chisq	df	p
factor(sex)	0.0727196	1	0.7874175
factor(candida)	0.8215195	1	0.3647363
factor(herpes)	0.0053108	1	0.9419056
factor(pneumo)	0.1062111	1	0.7444996
GLOBAL	1.1041761	4	0.8936091

Figura 3.11



Foi feita a análise de resíduos de Schoenfeld,para avaliar a suposição de riscos proporcionais. Ao nível de significância de 5%, nenhuma variável apresenta violação do pressuposto. Também foi feito um gráfico utilizando os resíduos de Cox-Snell, que avalia a qualidade global do modelo. Nota-se que os resíduos seguem uma distribuição exponencial, então pode-se dizer que o modelo está bemajustado.

3.3 Modelo Paramétrico

Usando um modelo paramétrico, estamos modelando o tempo de falha para uma distribuição específica. O primeiro passo é entender qual distribuição melhor descreve esse tempo, para assim

se ter o modelo mais adequado de interpretação. Estará sendo avaliado os modelos Exponencial, Weibull e o Log-normal.

Analisando a tabela, vemos os resultados dos testes da razão de verossimilhança para as hipóteses de que: i) o modelo de Weibull é adequado, ii) o modelo lognormal é adequado e iii) o modelo exponencial é adequado, sendo realizados utilizando-se o modelo gama generalizado. Pelos valores, vemos que o modelo Log-normal é o mais adequado.

Comparação	TRV	GL	Valor-p
Weibull	6.397541	2	0.0408124
Lognormal	2.737214	2	0.2544612
Exponencial	7.417784	2	0.0245047

Variável	Estimativa	Erro.Padrão	z.valor	p.valor
Intercepto	5.076	0.317	16.03	8.32e-58
Sexo (Masculino)	-0.611	0.278	-2.19	2.83e-02
Idade 31-40	0.685	0.286	2.39	1.67e-02
Idade 41-50	0.522	0.361	1.44	1.49e-01
Idade 51-60	0.628	0.686	0.92	3.60e-01
Idade >60	1.413	1.298	1.09	2.76e-01
Candidíase	3.769	0.365	10.33	5.35e-25
Herpes	4.557	0.750	6.08	1.23e-09
Pneumocistose	3.569	0.471	7.58	3.36e-14
Doença Hematológica	4.026	0.381	10.56	4.40e-26
Log(Scale)	0.441	0.078	5.65	1.57e-08

Já sabendo que o tempo de falha melhor segue uma distribuição lognormal,precisa se saber qual o melhor modelo a se construir com essa distribuição. Para isso, segue-se com o método de backward selection. O modelo final ficaram as variáveis sexo, Candidíase, Herpes, Pneumocistos e Doença Hematológica

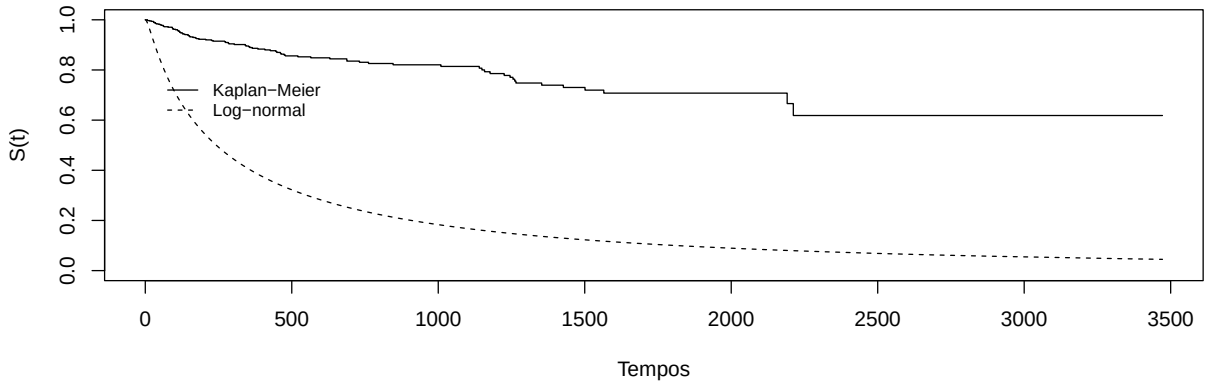
Variável	Estimativa	OR	Erro.Padrão	z.valor	p.valor
Intercepto	5.491	242.500	0.274	20.04	< 2e-16
Sexo (Masculino)	-0.600	0.549	0.273	-2.20	0.028
Candidíase	3.816	45.422	0.369	10.34	< 2e-16
Herpes	4.555	95.107	0.767	5.94	2.9e-09
Pneumocistose	3.555	34.988	0.469	7.58	3.3e-14
Doença Hematológica	3.943	51.573	0.370	10.65	< 2e-16
Log(Scale)	0.453	1.573	0.078	5.80	6.6e-09

O modelo de sobrevivência Lognormal indicam que diversas condições clínicas estão fortemente associadas ao risco de desenvolver tuberculose. As variáveis candidíase, herpes, pneumocistose e doença hematológica apresentaram razões de chance elevadas, sugerindo que indivíduos com essas condições possuem um risco significativamente maior de desenvolver a doença. Dentre elas, a

presença de herpes se destacou com a maior razão de chance (OR = 95.107), evidenciando um impacto expressivo na progressão para tuberculose.

Além disso, o sexo masculino apresentou um efeito protetor moderado, com uma razão de chance inferior a 1 (OR = 0.549), indicando que homens possuem um risco ligeiramente menor em comparação com as mulheres. Essa diferença pode estar relacionada a fatores biológicos ou comportamentais que influenciam a suscetibilidade à tuberculose.

Figura 3.12: Adequação do modelo paramétrico



Contudo, analisando o Figura 3.12 , que mostra as curvas de sobrevivência estimadas por Kaplan-Meier e pelo modelo log-normal, nota-se que o modelo não é adequado para a análise desses dados.

4 Conclusão

Embora o modelo Lognormal tenha se mostrado o mais adequado entre as distribuições testadas (Weibull, Lognormal e Exponencial), a comparação com as curvas de Kaplan-Meier revelou que o ajuste do modelo não foi ideal, o que sugere que, apesar de suas forças, ele não é totalmente adequado para este conjunto de dados.

Ao comparar o desempenho dos modelos, o Modelo de Cox se destacou por sua robustez e bom ajuste aos dados. Ele forneceu resultados significativos. O Modelo de Cox é particularmente vantajoso em estudos de sobrevivência, pois não exige suposições rígidas sobre a distribuição dos dados e é capaz de lidar com as variáveis explicativas de forma eficiente, mostrando a relação entre as covariáveis e o tempo de sobrevivência.

Por outro lado, o Modelo Paramétrico Lognormal, embora útil para detalhar o tempo de falha e modelar distribuições específicas de risco, não apresentou o mesmo nível de adequação quando comparado ao Modelo de Cox. A comparação com as curvas de Kaplan-Meier indicou que o modelo Lognormal não conseguiu capturar adequadamente os padrões de sobrevivência observados nos dados. Isso sugere que, apesar de ser uma ferramenta importante, o modelo paramétrico Lognormal não deve ser priorizado neste contexto, uma vez que não representa de forma ideal a dinâmica de sobrevivência dos dados.

Em conclusão, o Modelo de Cox revelou-se mais adequado para este estudo, com uma interpretação mais clara e resultados mais confiáveis. O Modelo Lognormal, por sua vez, pode ser descartado como a melhor opção, dado o seu desempenho inferior em relação às curvas de Kaplan-Meier. Portanto, recomenda-se o uso do Modelo de Cox para uma análise mais precisa e robusta do tempo de sobrevivência e dos fatores associados.