

Modelos Estatísticos e de Aprendizado de Máquina na Identificação de Fatores de Risco da Leishmaniose Tegumentar Americana no Brasil

Guto Leoni Santos¹, Vanderson Sampaio²

¹Universidade Federal de Pernambuco
Recife, PE – Brazil

²Instituto Todos pela Saúde (ITPS)
São Paulo, SP – Brazil

gls4@cin.ufpe.br, vandersons@gmail.com

Abstract. *Cutaneous leishmaniasis (CL) is a neglected tropical disease endemic in Brazil, with over 350,000 cases reported between 2007 and 2025. This study compares the performance of traditional statistical models (Cox Proportional Hazards) and machine learning approaches (Random Survival Forest, Extra Survival Trees) in identifying risk factors for death and treatment abandonment in CL patients. Using data from the Brazilian Notifiable Diseases Information System (SINAN), we evaluated six survival models via concordance index (C-index). For the mortality outcome, Cox PH achieved the highest C-index (0.896), while for treatment abandonment, all models showed moderate discrimination. Results suggest that traditional models remain competitive for structured epidemiological data with linear risk relationships.*

Resumo. *A leishmaniose tegumentar americana (LTA) é uma doença tropical negligenciada endêmica no Brasil, com mais de 350 mil casos notificados entre 2007 e 2025. Este estudo compara o desempenho de modelos estatísticos tradicionais (Regressão de Cox) e abordagens de aprendizado de máquina (Random Survival Forest, Extra Survival Trees) na identificação de fatores de risco para óbito e abandono de tratamento em pacientes com LTA. Utilizando dados do Sistema de Informação de Agravos de Notificação (SINAN), avaliamos seis modelos de sobrevivência por meio do índice de concordância (C-index). Para o desfecho de óbito, o modelo de Cox obteve o maior C-index (0,896), enquanto para abandono de tratamento todos os modelos apresentaram discriminação moderada. Os resultados sugerem que modelos tradicionais permanecem competitivos para dados epidemiológicos estruturados com relações de risco lineares.*

1. Introdução

A leishmaniose tegumentar americana (LTA) é uma doença infecciosa causada por protozoários do gênero *Leishmania*, transmitida pela picada de flebotomíneos infectados. A doença manifesta-se predominantemente nas formas cutânea e mucosa, podendo causar lesões desfigurantes e incapacitantes quando não tratada adequadamente [World Health Organization 2024]. Globalmente, estima-se que ocorram entre 600 mil e

1 milhão de novos casos anuais, sendo o Brasil um dos países com maior carga da doença nas Américas [Alvar et al. 2012].

No Brasil, a LTA é uma doença de notificação compulsória, com registro sistemático no Sistema de Informação de Agravos de Notificação (SINAN) [Ministério da Saúde 2024]. Entre 2007 e 2025, foram notificados mais de 351 mil casos, com predominância na região amazônica e no nordeste do país. A forma cutânea representa aproximadamente 94% dos casos, enquanto a forma mucosa, de maior gravidade, corresponde a cerca de 6% [Ministério da Saúde 2017].

Apesar dos avanços no diagnóstico e tratamento, a LTA ainda apresenta desafios significativos em termos de desfechos clínicos. A taxa de cura é elevada (aproximadamente 70% dos casos encerrados), porém o abandono de tratamento (2,7%) e o óbito (0,6%) representam desfechos desfavoráveis cujos fatores de risco ainda não são completamente compreendidos [Goto and Lindoso 2010]. A identificação desses fatores é fundamental para direcionar políticas de saúde pública e otimizar o manejo clínico dos pacientes.

Tradicionalmente, a análise de sobrevivência em epidemiologia utiliza o modelo de riscos proporcionais de Cox [Cox 1972], que assume uma relação linear entre as covariáveis e o risco. Nas últimas décadas, métodos de aprendizado de máquina (ML) como o Random Survival Forest [Ishwaran et al. 2008] e o Extra Survival Trees [Geurts et al. 2006] têm sido propostos como alternativas capazes de capturar relações não lineares e interações complexas entre variáveis, sem as restrições de premissas paramétricas.

No entanto, a superioridade dos modelos de ML sobre os modelos estatísticos tradicionais não é universal e depende da natureza dos dados e do problema. Estudos comparativos em dados clínicos estruturados demonstram que modelos paramétricos e semi-paramétricos frequentemente igualam ou superam abordagens de ML [Harrell et al. 1996, Christodoulou et al. 2019]. Em dados epidemiológicos com variáveis categóricas e relações predominantemente lineares, modelos tradicionais podem apresentar desempenho equivalente ou superior aos métodos de ML.

Dado o contexto, o objetivo deste estudo é comparar o desempenho de modelos estatísticos tradicionais e de aprendizado de máquina na identificação de fatores de risco para dois desfechos da LTA: (1) óbito e (2) abandono de tratamento. São avaliados seis modelos de sobrevivência: dois estatísticos (Cox PH e Cox Elastic Net) e quatro baseados em aprendizado de máquina (Survival Tree, Extra Survival Tree, Random Survival Forest e Extra Survival Trees). A comparação é realizada por meio do índice de concordância (C-index) [Harrell et al. 1996], utilizando dados do SINAN referentes ao período de 2007 a 2025. Adicionalmente, analisamos a importância relativa das variáveis preditoras identificadas por cada modelo, contribuindo para a compreensão dos determinantes desses desfechos.

2. Materiais e Métodos

2.1. Fonte de Dados

Utilizou-se a base de dados do Sistema de Informação de Agravos de Notificação (SINAN) [Ministério da Saúde 2024], contendo registros de leishmaniose tegumentar

americana (CID-10: B55.1) notificados no Brasil entre 2007 e 2025. O dataset original compreende 351.267 registros e 75 variáveis, abrangendo dados demográficos, clínicos, laboratoriais, de tratamento e de evolução dos casos.

2.2. Critérios de Inclusão e Limpeza dos Dados

Foram aplicados os seguintes critérios de inclusão sequenciais, com rastreamento do número de registros a cada etapa:

1. Remoção de registros identificados como duplicidade (`nduplic_n = 2`);
2. Manutenção apenas de notificações individuais e negativas (`tp_not ∈ {1, 2}`), excluindo surtos e agregados;
3. Seleção de casos novos (`cla_tipo_n = 1`);
4. Restrição a casos confirmados por critério laboratorial (`critério = 1`);
5. Remoção de registros com idade ≤ 0 ou > 120 anos;
6. Conversão de valores negativos de tempo até tratamento para ausente.

Após a limpeza, o dataset resultante contém 267.912 registros.

2.3. Variáveis

2.3.1. Variáveis Preditoras

Foram selecionadas 16 variáveis preditoras organizadas em quatro grupos (Tabela 1):

Tabela 1. Variáveis preditoras utilizadas nos modelos de sobrevivência

Grupo	Variável	Tipo
Demográficas	Idade (anos)	Contínua
	Sexo	Categórica
	Gestante	Categórica
	Raça/Cor	Categórica
	Escolaridade	Categórica
Clínica	Co-infecção HIV	Categórica
Laboratoriais	Parasitológico direto	Categórica
	IRM (Montenegro)	Categórica
	Histopatologia	Categórica
Epidemiológicas	Critério de confirmação	Categórica
	Classificação epidemiol.	Categórica
	Caso autóctone	Categórica
	Doença rel. ao trabalho	Categórica
	Ano da notificação	Numérica
	UF de residência	Categórica
Temporal	Tempo até tratamento (dias)	Contínua

A variável idade foi decodificada a partir da codificação padrão do SINAN (campo `nu_idade_n`), onde o primeiro dígito indica a unidade temporal (4 = anos, 3 = meses, 2 = dias, 1 = horas). O tempo até tratamento foi derivado da diferença entre a data de início do tratamento e a data de notificação.

2.3.2. Tratamento de Dados Faltantes

Para variáveis categóricas com dados ausentes, foi criada uma categoria “Desconhecido”, preservando a informação de que o dado não foi coletado, que é uma abordagem adequada quando a ausência não é aleatória (MNAR) [Rubin 1976]. Para a variável co-infecção HIV, os registros com valor “Ignorado” foram reclassificados como ausentes antes da imputação. Para a variável contínua tempo até tratamento (15,0% de ausência), utilizou-se a mediana como valor de preenchimento. Variáveis de desfecho não foram imputadas.

2.3.3. Desfechos

Foram definidos dois desfechos para análise de sobrevivência, ambos derivados da variável de evolução do caso e do tempo até encerramento (diferença entre a data de encerramento e a data de notificação):

- **Óbito:** evento definido como evolução para óbito por LTA ou óbito por outras causas. Demais evoluções (cura, abandono, transferência) foram tratadas como censura. Prevalência do evento: 0,6% (1.288 eventos em 202.472 registros).
- **Abandono de tratamento:** evento definido como abandono do tratamento. Demais evoluções foram tratadas como censura. Prevalência do evento: 3,5% (7.042 eventos em 202.472 registros).

2.4. Modelos de Sobrevivência

Foram avaliados seis modelos de sobrevivência, organizados em duas categorias: modelos estatísticos tradicionais e modelos baseados em aprendizado de máquina. Todos foram implementados por meio da biblioteca *scikit-survival* [Pölsterl 2020].

2.4.1. Modelos Estatísticos

1. **Cox PH:** Modelo semiparamétrico de riscos proporcionais de Cox [Cox 1972], com regularização $\alpha = 0,01$. Estima a razão de risco (*hazard ratio*) de cada covariável assumindo que o efeito é multiplicativo e constante ao longo do tempo (proporcionalidade dos riscos). É o modelo padrão em análise de sobrevivência clínica.
2. **Cox Elastic Net:** Extensão do modelo de Cox com penalização Elastic Net [Simon et al. 2011], que combina regularização L1 (Lasso) e L2 (Ridge) com $l_1\text{-ratio} = 0,5$. A penalização L1 promove esparsidade nos coeficientes, permitindo seleção automática de variáveis, enquanto a L2 estabiliza a estimação em presença de multicolinearidade.

2.4.2. Modelos de Aprendizado de Máquina

3. **Survival Tree:** Árvore de sobrevivência única, com profundidade máxima de dez níveis. Particiona recursivamente o espaço de covariáveis buscando maximizar a

diferença nas curvas de sobrevivência entre os nós filhos. Modelo interpretável, porém sujeito a alta variância.

4. **Extra Survival Tree:** Variante da árvore de sobrevivência que utiliza pontos de corte aleatórios [Geurts et al. 2006] em vez de buscar o ponto ótimo, reduzindo o viés de seleção e o tempo computacional.
5. **Random Survival Forest (RSF):** Ensemble de 100 árvores de sobrevivência [Ishwaran et al. 2008], cada uma treinada em uma amostra *bootstrap* com seleção aleatória de $\sqrt{p}$ covariáveis por nó (onde p é o número total de covariáveis). A agregação reduz a variância das árvores individuais.
6. **Extra Survival Trees:** Ensemble de 100 árvores com pontos de corte aleatórios [Geurts et al. 2006], combinando a aleatorização do Extra Tree com a agregação do ensemble. Tende a ser mais rápido que o RSF por não buscar pontos de corte ótimos.

Todos os modelos foram treinados com o mesmo conjunto de dados para garantir comparabilidade. As variáveis categóricas foram codificadas via *one-hot encoding* com remoção da primeira categoria (*drop-first*).

2.5. Avaliação e Comparação

O dataset foi dividido em treino (80%) e teste (20%), com estratificação pelo evento de interesse. A métrica principal de avaliação foi o índice de concordância (C-index) de Harrell [Harrell et al. 1996], calculado no conjunto de teste. O C-index mede a capacidade discriminativa do modelo: um valor de 0,5 indica predição aleatória, enquanto 1,0 indica discriminação perfeita.

Para análise da importância das variáveis, utilizou-se: (a) o valor absoluto dos coeficientes para os modelos de Cox; (b) a importância baseada em impureza (*feature importance*) para as árvores de decisão; e (c) importância por permutação (*permutation importance*) para os modelos de ensemble (RSF e Extra Survival Trees).

Adicionalmente, foram construídas curvas de Kaplan-Meier [Kaplan and Meier 1958] estratificadas por forma clínica, droga inicial, co-infecção HIV, faixa etária e raça/cor, com teste de log-rank para comparação entre grupos. O código-fonte utilizado neste estudo está disponível publicamente¹. Os dados do SINAN são públicos e podem ser obtidos por meio do portal DATASUS².

3. Resultados

3.1. Caracterização da Amostra

Após a aplicação dos critérios de limpeza, o dataset preparado contém 267.912 registros. Para a análise de sobrevivência, foram utilizados 202.472 registros com informações válidas de tempo até encerramento e evolução do caso. O perfil predominante dos pacientes é do sexo masculino (72,9%), raça/cor parda (63,3%), com mediana de idade de 33 anos. A forma cutânea representa 93,8% dos casos e o antimonial pentavalente foi a droga inicial em 78,1% dos tratamentos.

¹<https://github.com/GutoL/lta-survival-models>

²<https://datasus.saude.gov.br/transferencia-de-arquivos/>

3.2. Análise Exploratória com Curvas de Kaplan-Meier

As curvas de Kaplan-Meier estratificadas, apresentadas nas Figuras 1 e 2, avaliaram a proporção de cura ao longo do tempo por forma clínica e por co-infecção HIV. Para ambas as estratificações, o teste de log-rank não revelou diferença estatisticamente significativa entre os grupos: forma cutânea vs. mucosa ($p = 0,889$) e HIV+ vs. HIV– ($p = 0,126$). Este resultado, aparentemente contraintuitivo dada a relevância clínica reconhecida dessas variáveis, é discutido adiante (Seção 4.3) à luz da definição do desfecho e do mecanismo de censura empregado.

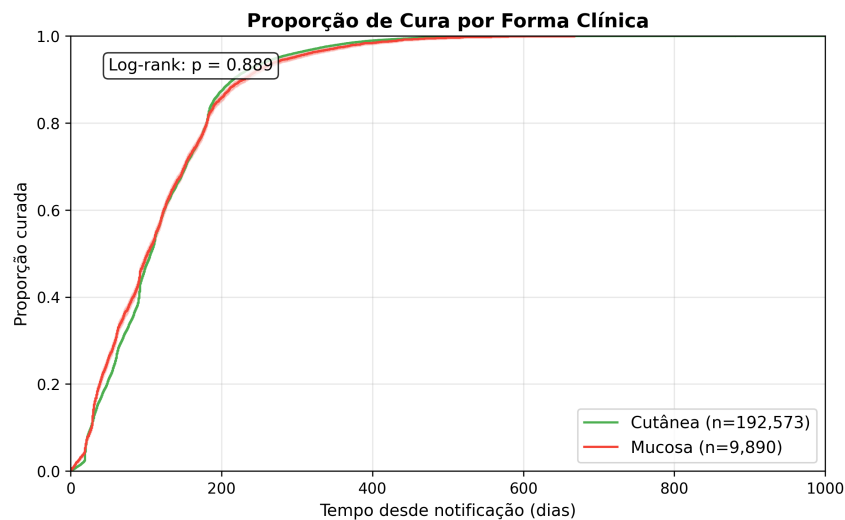


Figura 1. Curva de Kaplan-Meier da proporção de cura estratificada por forma clínica (cutânea vs. mucosa). Log-rank: $p = 0,889$.

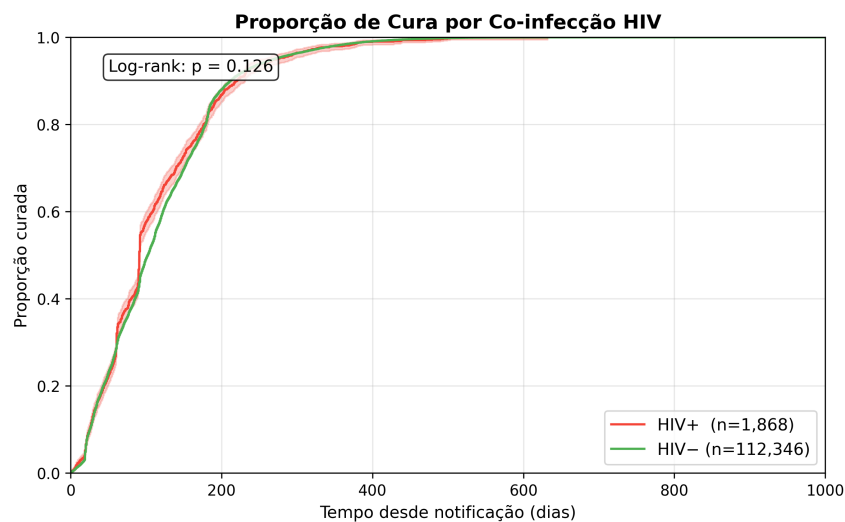


Figura 2. Curva de Kaplan-Meier da proporção de cura estratificada por co-infecção HIV. Log-rank: $p = 0,126$.

3.3. Análise de Sobrevivência: Óbito

Para o desfecho de óbito (LTA e outras causas), entre os 202.472 registros analisados, foram identificados 1.288 eventos (0,6%) e 201.184 censurados (99,4%). Os

resultados dos seis modelos são apresentados na Tabela 2.

Tabela 2. Índice de concordância (C-index) dos modelos para o desfecho de óbito

Tipo	Modelo	C-index	Tempo (s)
Estatístico	Cox PH	0,8965	18,1
	Cox Elastic Net	0,8913	0,4
Árvore única	Survival Tree	0,7956	1,6
	Extra Survival Tree	0,8314	0,6
Ensemble	Random Survival Forest	0,8946	7,1
	Extra Survival Trees	0,8945	4,9

O modelo de Cox PH obteve o maior C-index (0,8965), seguido de perto pelo Random Survival Forest (0,8946) e Extra Survival Trees (0,8945). A diferença entre o melhor modelo estatístico e os melhores modelos de ensemble é inferior a 0,002, indicando desempenho discriminativo praticamente equivalente. O Cox Elastic Net apresentou C-index ligeiramente inferior (0,8913), possivelmente devido à penalização excessiva de variáveis relevantes. As árvores únicas apresentaram desempenho inferior aos demais modelos, com a Survival Tree obtendo o menor C-index (0,7956).

3.4. Análise de Sobrevivência: Abandono

Para o desfecho de abandono de tratamento, foram identificados 7.042 eventos (3,5%) entre 195.430 censurados (96,5%). Os resultados são apresentados na Tabela 3.

Tabela 3. Índice de concordância (C-index) dos modelos para o desfecho de abandono de tratamento

Tipo	Modelo	C-index	Tempo (s)
Estatístico	Cox PH	0,6599	7,5
	Cox Elastic Net	0,5117	0,4
Árvore única	Survival Tree	0,6171	1,8
	Extra Survival Tree	0,6225	0,6
Ensemble	Random Survival Forest	0,6697	7,6
	Extra Survival Trees	0,6663	4,9

Para o abandono, os modelos de ensemble superaram os modelos estatísticos, com o Random Survival Forest obtendo o maior C-index (0,6697), seguido pelo Extra Survival Trees (0,6663) e Cox PH (0,6599). No entanto, todos os modelos apresentaram discriminação moderada (C-index entre 0,5 e 0,7), sugerindo que as variáveis disponíveis no SINAN têm capacidade limitada para prever o abandono de tratamento. O Cox Elastic Net apresentou desempenho próximo ao aleatório (0,5117), indicando que a regularização eliminou variáveis que, embora com efeito individual fraco, contribuem coletivamente para a predição.

3.5. Comparação entre Desfechos

A capacidade discriminativa dos modelos foi substancialmente maior para o desfecho de óbito (C-index entre 0,80 e 0,90) do que para o abandono (C-index entre 0,51 e 0,67). Essa diferença sugere que o óbito está mais fortemente associado a fatores clínicos e demográficos registrados no SINAN, enquanto o abandono é influenciado por determinantes sociais e de acesso ao serviço de saúde que não são capturados pelo sistema de notificação.

Para o óbito, os modelos estatísticos (Cox PH) e de ensemble (RSF, Extra Survival Trees) apresentaram desempenho equivalente, indicando que a relação entre as co-variáveis e o risco de óbito é predominantemente linear e proporcional. Para o abandono, os modelos de ensemble apresentaram vantagem marginal, sugerindo a presença de interações não lineares entre as variáveis preditoras.

4. Discussão

Os resultados deste estudo demonstram que a escolha entre modelos estatísticos tradicionais e de aprendizado de máquina para análise de sobrevivência em LTA depende fundamentalmente do desfecho investigado. Para o óbito, o modelo de Cox PH apresentou o melhor desempenho discriminativo (C-index = 0,8965), enquanto para o abandono de tratamento, os modelos de ensemble (RSF e Extra Survival Trees) obtiveram vantagem marginal sobre os modelos lineares.

4.1. Desempenho dos Modelos para Óbito

O elevado C-index obtido por todos os modelos para o desfecho de óbito (0,80–0,90) indica que as variáveis disponíveis no SINAN possuem forte capacidade preditiva para mortalidade em pacientes com LTA. A superioridade do Cox PH sobre os modelos de ML, ainda que por margem estreita, sugere que a relação entre os fatores de risco e o óbito é predominantemente linear e proporcional, que é a premissa do modelo de Cox [Cox 1972].

Este achado é consistente com a literatura em análise de sobrevivência aplicada a dados clínicos estruturados, onde modelos paramétricos e semiparamétricos frequentemente igualam ou superam abordagens de ML [Harrell et al. 1996]. A vantagem dos modelos lineares nesse contexto pode ser atribuída à natureza das variáveis preditoras: predominantemente categóricas, com efeitos aditivos sobre o risco, e sem interações complexas de alta ordem.

As árvores únicas (Survival Tree e Extra Survival Tree) apresentaram desempenho inferior aos demais modelos, o que é esperado dado que árvores individuais tendem a sofrer de alta variância e são sensíveis à estrutura específica dos dados de treino. Os modelos de ensemble (RSF e Extra Survival Trees) mitigam essa limitação pela agregação de múltiplas árvores, alcançando desempenho próximo ao do Cox PH.

4.2. Desempenho dos Modelos para Abandono

A discriminação moderada obtida por todos os modelos para o abandono de tratamento (C-index entre 0,51 e 0,67) revela uma limitação importante: as variáveis registradas no SINAN são insuficientes para prever adequadamente esse desfecho. O abandono de tratamento é um fenômeno multifatorial, influenciado por determinantes sociais,

econômicos e de acesso ao serviço de saúde, como distância até a unidade de tratamento, renda familiar, suporte social e efeitos adversos da medicação, que não são capturados pelo sistema de notificação [Ministério da Saúde 2017].

A ligeira superioridade dos modelos de ensemble (RSF: 0,6697; Extra Survival Trees: 0,6663) sobre o Cox PH (0,6599) para esse desfecho sugere a presença de interações não lineares entre as variáveis preditoras. É plausível que combinações específicas de fatores, como idade avançada em determinadas regiões geográficas, ou co-infecção HIV associada a baixa escolaridade contribuam para o risco de abandono de forma que modelos lineares não capturam completamente.

O desempenho próximo ao aleatório do Cox Elastic Net (0,5117) para o abandono merece atenção. A regularização Elastic Net penaliza coeficientes pequenos, podendo eliminar variáveis que individualmente têm efeito fraco mas que coletivamente contribuem para a predição. Esse resultado reforça que, para desfechos com sinais fracos e distribuídos entre múltiplas variáveis, a regularização agressiva pode ser contraproducente.

4.3. Curvas de Kaplan-Meier e o Viés do Desfecho

Um achado aparentemente paradoxal deste estudo merece atenção: embora a forma clínica e a co-infecção HIV sejam reconhecidamente associadas a pior prognóstico em LTA [Goto and Lindoso 2010], as curvas de Kaplan-Meier estratificadas por essas variáveis não revelaram diferenças estatisticamente significativas na proporção de cura ao longo do tempo (log-rank $p = 0,889$ e $p = 0,126$, respectivamente).

Essa ausência de diferença tem explicação metodológica. A análise de Kaplan-Meier utilizada neste estudo modela especificamente o tempo até a cura, tratando como censurados todos os pacientes com desfechos distintos (óbito, abandono, transferência e mudança de diagnóstico). Assim, pacientes com piores prognósticos, que tendem a evoluir para óbito ou abandono, são removidos da contagem de cura e não contribuem para a curva como “não-cura”, mas sim como censurados. Em consequência, a curva de cura captura apenas o comportamento dos pacientes que de fato evoluem para cura, perdendo o sinal dos casos mais graves.

Esse viés explica por que os modelos preditivos de óbito identificam HIV e forma clínica como relevantes (C-index elevado), enquanto as curvas KM não detectam diferença na cura: são análises de desfechos diferentes. Uma abordagem mais informativa para investigar o impacto dessas variáveis no prognóstico completo seria a análise de riscos competitivos (*competing risks*) [Ishwaran et al. 2008], que modela simultaneamente os múltiplos desfechos possíveis (cura, óbito e abandono) preservando a informação sobre a distribuição de eventos entre os grupos.

4.4. Limitações

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. Primeiro, a base do SINAN está sujeita a subnotificação e incompletude dos registros, particularmente para variáveis como co-infecção HIV (44,3% de dados ausentes após reclassificação dos ignorados) e evolução do caso (22,7% ausente). Segundo, a baixa prevalência do evento de óbito (0,6%) pode inflar artificialmente o C-index, uma vez que a maioria dos pares comparados envolve pacientes censurados. Embora o C-index seja uma métrica baseada em ordenação de pares e não dependa diretamente da prevalência

do evento (diferentemente de métricas como acurácia ou F1-score), em cenários com poucos eventos a variância da estimativa aumenta e a interpretação deve ser cautelosa [Harrell et al. 1996]. Ainda assim, o C-index permanece a métrica padrão em análise de sobrevivência por sua capacidade de avaliar discriminação sem exigir a definição de um ponto de corte temporal. Terceiro, não foram avaliadas as premissas do modelo de Cox (proporcionalidade dos riscos) por meio do teste de Schoenfeld, o que poderia revelar violações que favoreçam os modelos de ML. Por fim, a ausência de variáveis socioeconômicas e de acesso ao serviço de saúde limita a capacidade preditiva dos modelos para o desfecho de abandono.

5. Conclusão

Este estudo comparou seis modelos de sobrevivência, dois estatísticos tradicionais e quatro baseados em aprendizado de máquina, na identificação de fatores de risco para óbito e abandono de tratamento em pacientes com leishmaniose tegumentar americana no Brasil, utilizando dados do SINAN (2007–2025).

Os principais achados são: (1) para o desfecho de óbito, o modelo de Cox PH apresentou o melhor desempenho (C-index = 0,8965), demonstrando que modelos estatísticos tradicionais permanecem competitivos quando a relação entre covariáveis e risco é predominantemente linear; (2) para o abandono de tratamento, os modelos de ensemble (RSF e Extra Survival Trees) superaram marginalmente o Cox PH, sugerindo a presença de interações não lineares; e (3) a capacidade preditiva para abandono foi substancialmente inferior à do óbito, indicando que os determinantes do abandono não são adequadamente capturados pelas variáveis disponíveis no SINAN.

Trabalhos futuros devem considerar a incorporação de variáveis socioeconômicas e de acesso ao serviço de saúde para melhorar a predição do abandono, a validação das premissas do modelo de Cox por meio do teste de Schoenfeld, a aplicação de modelos de riscos competitivos para tratar simultaneamente os múltiplos desfechos possíveis da LTA, e a análise de importância das variáveis preditoras por meio de técnicas como importância por permutação e valores SHAP, permitindo identificar quais fatores de risco são mais relevantes para cada desfecho segundo os diferentes modelos.

Referências

- Alvar, J., Vélez, I. D., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., and den Boer, M. (2012). Leishmaniasis worldwide and global estimates of its incidence. *PLoS ONE*, 7(5):e35671.
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Calster, B. V. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

- Goto, H. and Lindoso, J. A. L. (2010). Current diagnosis and treatment of cutaneous and mucocutaneous leishmaniasis. *Expert Review of Anti-infective Therapy*, 8(4):419–433.
- Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.
- Ministério da Saúde (2017). Manual de vigilância da leishmaniose tegumentar. *Secretaria de Vigilância em Saúde*.
- Ministério da Saúde (2024). Sistema de informação de agravos de notificação – SINAN. Disponível em: <https://portalsinan.saude.gov.br/>.
- Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.
- World Health Organization (2024). Leishmaniasis: Fact sheet. *WHO Media Centre*. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/leishmaniasis>.