Análise de Sobrevivência com Redes Neurais Aplicada à Interrupção da Terapia Antirretroviral no Brasil

Arthur Pontes Motta

Orientador: Rafael Souza dos Santos

Universidade Federal do Rio de Janeiro

Instituto de Matemática

Departamento de Métodos Estatísticos

2026

Análise de Sobrevivência com Redes Neurais Aplicada à Interrupção da Terapia Antirretroviral no Brasil

Arthur Pontes Motta

Trabalho de Conclusão de Curso submetido ao Instituto de Matemática da Universidade Federal do Rio de Janeiro, como requisito necessário para obtenção do grau de Bacharel em Ciências Atuariais.

Aprovada por:

Rafael Souza dos Santos

Dr.Sc. - IM - UFRJ - Orientador.

Marina Silva Paez

Ph.D. - IM - UFRJ.

Hugo Tremonte de Carvalho

Dr.Sc. - IM - UFRJ.

Rio de Janeiro, RJ - Brasil

2026

CIP - Catalogação na Publicação

M921a

Motta, Arthur Pontes
Análise de Sobrevivência com Redes Neurais Aplicada à Interrupção da Terapia Antirretroviral no Brasil / Arthur Pontes Motta. -- Rio de Janeiro, 2026.
185 f.

Orientador: Rafael Souza dos Santos.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciências atuariais, 2026.

1. terapia antirretroviral. 2. análise de sobrevivência. 3. redes neurais artificiais. 4. aprendizado de máquina. 5. HIV/Aids. I. Santos, Rafael Souza dos, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

Agradecimentos

À minha família, devo a gratidão que não se expressa em palavras isoladas, mas no reconhecimento de que, sem o apoio moral e material que me ofereceram ao longo de toda esta jornada, nada do que aqui se registra teria sido possível.

Aos amigos que a graduação me deu, e aqui nomeio porque é preciso nomear quando se reconhece o valor de uma presença, meu imenso obrigado: Catarine Martins, Daniel Norberto, Ester Costa, Gabriel Faria, Isabela Barros, João Elias, Marcelo Máximo, Maria Clara Martinez, Maria Clara Peres, Milena Doarte, Natália Imlau (e suas caronas), Pablo Muniz, Pedro Quinet, Stefane Moraes, Victor Hugo Freire, Vitor Hugo Bellas e Yasmin Santana. Com eles compartilhei não apenas os sucessos previsíveis, mas sobretudo aqueles momentos de estresse e superação que, no entanto, são os que verdadeiramente formam o caráter. Cada um deixou em mim uma marca que transcende o âmbito acadêmico: contribuíram para o meu amadurecimento pessoal e me ensinaram, ainda que sem saber, a lidar com as adversidades que escapam ao nosso controle e desafiam nossa capacidade de previsão.

Ao meu amigo e monitor, João Pedro Costa, um agradecimento que vai além da gratidão protocolar. Sou lhe grato não apenas pela ajuda em Cálculo Numérico, ajuda essa que foi decisiva, diga-se de passagem, mas principalmente pelo voto de confiança e pela generosidade de celebrar minhas conquistas como se fossem suas. A admiração é recíproca, e a torcida também. Acredito em sua força e em seu potencial para construir um futuro brilhante; estarei sempre na primeira fila para aplaudí-lo, porque é isso que fazem aqueles que reconhecem no outro a promessa de uma grandeza ainda por vir.

Também à minha monitora de Processos Estocásticos, Karolayne Dessabato, que se esforçou enormemente, e com paciência admirável, para me fazer compreender e aproveitar a disciplina. Sem sua ajuda, eu não estaria aqui.

Aos amigos da vida afora, João Pedro Bertolino, Karolina Menezes, Luan Cidrini, Mateus Fernandes, Pedro Santos e Victor Hugo Teixeira, obrigado pela amizade leve, pelas conversas que variavam do profundo ao absurdo (e às vezes as duas coisas ao mesmo tempo), pelos conselhos sinceros e pela presença constante. Vocês tornam a vida muito mais divertida, e isso, convenhamos, não é pouca coisa.

Ao meu melhor amigo e irmão, Victor Motta, com quem estive lado a lado em cada momento, dos mais simples aos mais deprimentes. Nossa trajetória, repleta de risos, choros, reconciliações e todo o arsenal emocional que uma amizade verdadeira comporta, é o que fortalece o nosso laço. Obrigado por ser meu apoio quando o mundo insistia em desabar.

Sou grato também aos amigos que fizeram parte desta jornada e que, seguindo outros caminhos, deixaram sua marca. Algumas pessoas permanecem, outras seguem em novas direções, mas todas, sem exceção, contribuem para a construção de quem somos. Cada encontro tem seu significado, ainda que o tempo seja necessário para revelá-lo. E é justamente essa espera, esse amadurecimento da compreensão, que nos ensina a distinguir o essencial do acidental.

Aos meus professores, minha mais profunda gratidão. Aos professores Juliana Vianna e Nei Carlos, pelo compromisso decisivo, e uso aqui o termo com toda a precisão que ele merece, que foi fundamental para minha permanência e conclusão do curso, e pelo acolhimento que se tornou um guia seguro em um momento particularmente sensível da graduação. À professora Maria Eulália Vares, coordenadora do Programa de Pós-Graduação em Estatística da UFRJ, pela grande oportunidade de uma iniciação científica marcada pela genialidade, pelo bom humor e pela sensibilidade, combinação rara e muito preciosa.

Aos membros da banca examinadora, professores Marina Silva Paez e Hugo Tremonte de Carvalho, meu sincero agradecimento pela disponibilidade e pela generosidade em dedicar seu tempo à leitura e avaliação deste trabalho. Ao professor Hugo, agradeço não apenas por aceitar compor esta banca, mas também pelas aulas que ministrou ao longo do curso, pela orientação no projeto de extensão e pelo acompanhamento atento durante as orientações acadêmicas semestrais. À professora Marina, agradeço a gentileza de participar desta etapa. A presença de ambos neste momento final é uma honra que reconheço com profunda gratidão.

E, de forma especial, ao meu orientador, Rafael Souza dos Santos. Agradeço não apenas pela condução deste trabalho, mas por todo o aprendizado acumulado ao longo das diversas disciplinas que cursei sob sua orientação. Sua metodologia, sua capacidade de extrair o meu melhor (aos trancos e barrancos) e sua disponibilidade como guia foram determinantes para a minha formação. Não é exagero dizer que, sem ele, este percurso teria sido outro, provavelmente menos desafiador, mas certamente menos prolífico.

Aos demais professores, mesmo aqueles com quem tive menos contato, deixo meu

reconhecimento. Cada gesto, cada conselho e cada aula, por mais modesta que parecesse, contribuíram para a minha formação. E isso merece ser dito: muito obrigado.

Por fim, agradeço à minha gata, Jolie. Durante este percurso intenso, ela foi minha companhia silenciosa, meu descanso visual entre uma linha de código e outra. Mesmo sem entender a dimensão do processo, e talvez justamente por isso, sua presença foi um alívio em meio ao caos e uma fonte constante de afeto. Há coisas que não se explicam, apenas se agradecem.

"A desinformação vem da profusão da informação, de seu encantamento, de sua repetição em círculos, que cria um campo de percepção vazio, um espaço como que desintegrado por uma bomba de nêutrons, ou por uma bomba que absorve todo oxigênio em volta."

– Jean Baudrillard

Resumo

A terapia antirretroviral (TARV) consolidou-se como a principal estratégia de controle da epidemia de HIV/Aids, reduzindo a morbimortalidade e prevenindo a transmissão do vírus. No entanto, a interrupção do tratamento e as falhas de adesão ainda representam desafios significativos para o alcance das metas globais de eliminação do HIV como problema de saúde pública.

Este trabalho analisa os fatores associados ao tempo até a interrupção da TARV no Brasil, empregando técnicas de análise de sobrevivência integradas a métodos de aprendizado profundo. A pesquisa baseia-se em uma coorte nacional composta por 1.094.567 indivíduos acompanhados desde o início do tratamento até a ocorrência da interrupção ou censura, com horizonte temporal de até 24 anos.

Foi implementado o modelo *Nnet-Survival*, uma rede neural artificial desenvolvida para análise de sobrevivência em tempo discreto, obtendo-se métricas de desempenho preditivo bastante satisfatórias. As curvas de sobrevivência preditas revelaram desigualdades persistentes segundo características sociodemográficas e clínicas, com maior risco de interrupção entre homens, pessoas pretas e indígenas, indivíduos de menor escolaridade e faixas etárias mais jovens. A análise de importância das variáveis identificou faixa etária de 40 a 59 anos, escolaridade de 12 anos ou mais e supressão viral como os principais determinantes da retenção em cuidado.

A aplicação do modelo evidenciou a viabilidade e a relevância do uso de redes neurais artificiais na modelagem de sobrevivência em saúde pública, demonstrando sua capacidade de representar padrões complexos de risco e subsidiar o aprimoramento das estratégias de adesão e retenção em cuidado no contexto da TARV.

Palavras-Chave: terapia antirretroviral; análise de sobrevivência; redes neurais artificiais; aprendizado de máquina; HIV/Aids.

Abstract

Antiretroviral therapy (ART) has become the cornerstone strategy for controlling the HIV/AIDS epidemic, reducing morbidity and mortality while preventing viral transmission. However, treatment interruption and poor adherence remain major challenges to achieving global targets for eliminating HIV as a public health issue.

This study investigates the factors associated with time to ART interruption in Brazil, combining traditional survival analysis techniques with deep learning methods. The research is based on a nationwide cohort comprising 1,094,567 individuals monitored from the initiation of treatment until either interruption or censoring, with a follow-up period of up to 24 years.

The *Nnet-Survival* model, an artificial neural network specifically designed for discrete-time survival analysis, was implemented, achieving highly satisfactory predictive performance metrics. The predicted survival curves revealed persistent inequalities according to sociodemographic and clinical characteristics, with higher risk of treatment interruption among men, black and indigenous people, individuals with lower educational attainment, and younger age groups. Variable importance analysis identified age group of 40 to 59 years, educational level of 12 years or more, and viral suppression as the main determinants of retention in care.

The application of the model demonstrated the feasibility and relevance of using artificial neural networks for survival modeling in public health, highlighting their ability to capture complex risk patterns and support the improvement of adherence and retention strategies within the context of ART.

Keywords: antiretroviral therapy; survival analysis; artificial neural networks; machine learning; HIV/AIDS.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A epidemia de HIV/Aids representa uma das maiores crises de saúde pública da história contemporânea. Desde os primeiros casos, identificados em 1981, até os dias atuais, a síndrome da imunodeficiência adquirida infectou mais de 91 milhões de pessoas e causou aproximadamente 44 milhões de mortes em todo o mundo (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2025). A trajetória dessa epidemia, contudo, não foi linear: passou de uma doença desconhecida e invariavelmente fatal para uma condição crônica manejável, graças aos avanços científicos na compreensão do vírus e no desenvolvimento de terapias eficazes. O Brasil foi pioneiro nessa transformação ao instituir, em 1996, a política de acesso universal e gratuito à terapia antirretroviral (TARV), tornando-se referência internacional no enfrentamento de doenças infecciosas e na promoção da equidade em saúde (Galvão, 2002; Brasil. Ministério da Saúde, 2024a).

1.1 Perspectiva Histórica do HIV/Aids

Em 5 de junho de 1981, o *Centers for Disease Control and Prevention* (CDC) dos Estados Unidos publicou no *Morbidity and Mortality Weekly Report* o primeiro relato oficial daquilo que viria a ser reconhecido como a epidemia de Aids[1]: cinco homens jovens, homossexuais e previamente saudáveis, haviam sido hospitalizados em Los Angeles com pneumonia por *Pneumocystis carinii*, uma infecção oportunista extremamente rara em

[1] O Ministério da Saúde do Brasil adota a grafia "Aids", tratando o termo como substantivo próprio e não como sigla. Essa convenção é seguida nos boletins epidemiológicos, protocolos clínicos e demais publicações oficiais, bem como pela literatura científica nacional em português. Por outro lado, a literatura internacional em inglês utiliza a forma "AIDS", mantendo o padrão de sigla (*Acquired Immunodeficiency Syndrome*).

pessoas com sistema imunológico saudável (Centers for Disease Control (CDC), 1981). Dois desses pacientes já haviam morrido quando o relatório foi publicado. Nas semanas seguintes, casos adicionais de sarcoma de Kaposi e pneumocistose foram reportados em Nova Iorque e na Califórnia, estabelecendo o padrão inicial de uma síndrome caracterizada pela falência progressiva do sistema imunológico. A doença emergente foi inicialmente denominada GRID (*Gay-Related Immune Deficiency*) ou, de forma ainda mais estigmatizante, "peste gay" e "câncer gay", terminologias que refletiam e perpetuavam um profundo preconceito social (Barré-Sinoussi; Ross; Delfraissy, 2013). Em 1982, o acrônimo dos "4H", que englobava homossexuais, usuários de heroína, hemofílicos e haitianos, categorizou os chamados "grupos de risco", reforçando a percepção errônea de que a doença estaria restrita a populações marginalizadas. Em setembro de 1982 o CDC adotou oficialmente o termo AIDS (*Acquired Immune Deficiency Syndrome*), reconhecendo que a síndrome não se restringia a grupos específicos, mas poderia afetar qualquer pessoa exposta ao agente etiológico ainda desconhecido (Centers for Disease Control (CDC), 1982).

A identificação do vírus responsável pela Aids envolveu intensa competição científica entre equipes da França e dos Estados Unidos. Em 20 de maio de 1983, Luc Montagnier, Françoise Barré-Sinoussi e colaboradores do Instituto Pasteur publicaram na revista *Science* o isolamento do LAV (*Lymphadenopathy-Associated Virus*) a partir de linfonodos de pacientes com a síndrome (Barré-Sinoussi *et al.*, 1983). Em abril de 1984, Robert Gallo, do *National Institutes of Health* (NIH), anunciou a identificação do HTLV-III (*Human T-Lymphotropic Virus Type III*), que posteriormente se revelou ser o mesmo vírus isolado pela equipe francesa. Em 1986, um comitê internacional de taxonomia unificou as nomenclaturas sob o nome HIV (*Human Immunodeficiency Virus*). A controvérsia sobre a prioridade científica da descoberta estendeu-se até 1987, quando os presidentes Ronald Reagan e François Mitterrand negociaram acordo diplomático dividindo créditos e royalties das patentes de testes diagnósticos (Gallo, 2002). Em 2008, o Prêmio Nobel de Fisiologia ou Medicina foi concedido a Montagnier e Barré-Sinoussi, reconhecendo formalmente a contribuição francesa como descoberta original do vírus. O desenvolvimento do primeiro teste de anticorpos ELISA (*Enzyme-Linked Immunosorbent Assay*), aprovado pelo FDA (*Food and Drug Administration*) em março de 1985, transformou radicalmente a capacidade de detecção e triagem do suprimento sanguíneo (Alexander, 2016). Antes da disponibilização desse teste, estimava-se que uma em cada 90 transfusões sanguíneas em algumas cidades americanas estava contaminada pelo HIV. A implementação da testagem obrigatória de doações de sangue reduziu drasticamente a transmissão por via

transfusional, embora milhares de hemofílicos já tivessem sido infectados por meio de fatores de coagulação contaminados[2].

O período entre 1985 e 1995 foi caracterizado por devastação sem precedentes. A epidemia expandiu-se globalmente, atingindo de forma particularmente severa a África Subsaariana, onde se concentravam mais de dois terços das infecções mundiais. Até meados da década de 1990, o diagnóstico de HIV era amplamente considerado uma sentença de morte: mais de 40% dos pacientes diagnosticados morriam em poucos anos, frequentemente após rápida deterioração física. O pico de novas infecções ocorreu em 1996, com aproximadamente 3,4 milhões de casos.

A transformação desse cenário iniciou-se naquele mesmo ano, com a introdução da terapia antirretroviral altamente ativa (HAART, do inglês *Highly Active Antiretroviral Therapy*), que combinava múltiplas drogas de diferentes classes farmacológicas para suprimir a replicação viral (Palella *et al.*, 1998). David Ho, do Aaron Diamond AIDS Research Center, foi pioneiro da abordagem terapêutica conhecida como "*hit hard, hit early*" (atacar forte, atacar cedo). Seus estudos demonstraram que o HIV produz aproximadamente 10 bilhões de vírions por dia no organismo infectado, fundamentando a necessidade de terapia combinada agressiva para evitar o desenvolvimento de resistência medicamentosa (Ho *et al.*, 1995). Os resultados clínicos da TARV foram expressivos e imediatos nos países com acesso ao tratamento: as mortes por Aids nos Estados Unidos caíram 70% entre 1995 e 1999, as hospitalizações relacionadas à síndrome diminuíram entre 60% e 80%, e a expectativa de vida de jovens com HIV tornou-se progressivamente comparável à da população geral quando em tratamento adequado (Palella *et al.*, 1998; Samji *et al.*, 2013). A Aids deixou de ser uma doença invariavelmente fatal para se tornar uma condição crônica passível de controle, desde que o paciente tivesse acesso aos medicamentos e mantivesse adesão ao tratamento.

Contudo, enquanto os países desenvolvidos colhiam os benefícios da TARV, a mortalidade global continuou a crescer, atingindo seu ápice apenas em 2004, com 2,1 milhões de óbitos. Essa defasagem de oito anos entre a introdução do tratamento e a redução global da mortalidade reflete a profunda desigualdade no acesso aos medicamentos: a África Subsaariana, onde se concentrava a maior parte das mortes, só passou a ter acesso massivo aos antirretrovirais a partir de 2004, com iniciativas como o PEPFAR (*President's Emergency Plan for AIDS Relief*) e o Fundo Global de Combate à Aids, Tuberculose e Malária.

[2]Fatores de coagulação são proteínas do plasma sanguíneo essenciais para a coagulação. Hemofílicos necessitam de infusões regulares desses concentrados, que antes da testagem obrigatória eram derivados de *pools* de milhares de doadores.

A criação do UNAIDS (*Joint United Nations Programme on HIV/AIDS*), em janeiro de 1996, consolidou a coordenação internacional da resposta à epidemia. Nas décadas seguintes, o programa estabeleceu metas progressivamente ambiciosas para orientar as políticas nacionais: as metas 90-90-90, lançadas em 2014, propunham que, até 2020, 90% das pessoas vivendo com HIV conhecessem seu diagnóstico, 90% das diagnosticadas estivessem em tratamento e 90% das pessoas em tratamento alcançassem supressão viral. Posteriormente, as metas foram atualizadas para 95-95-95, com horizonte de cumprimento até 2025. O objetivo final declarado pelo UNAIDS é eliminar a Aids como ameaça à saúde pública até 2030, reduzindo as novas infecções anuais para menos de 200 mil casos globalmente (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2025). De acordo com as estimativas mais recentes, em 2024 aproximadamente 40,8 milhões de pessoas viviam com HIV globalmente, das quais 53% eram mulheres e meninas. Foram registradas 1,3 milhão de novas infecções, o que representa redução de 61% em relação ao pico de 1996, e 630 mil mortes relacionadas à Aids, redução de 70% em relação ao pico de 2004. Atualmente, 31,6 milhões de pessoas, correspondendo a 77% do total de infectados, têm acesso à terapia antirretroviral. Quanto ao cumprimento das metas 95-95-95, estima-se que aproximadamente 87% das pessoas vivendo com HIV conhecem seu diagnóstico[3]

O primeiro caso de Aids no Brasil foi registrado em São Paulo, em 1980, sendo formalmente classificado em 1982 (Brasil. Ministério da Saúde, 2024b). O paciente foi atendido no Hospital Emílio Ribas, que se tornaria referência nacional no tratamento da doença. Os primeiros sete casos confirmados ocorreram em São Paulo, todos em pacientes com prática homo ou bissexual, reproduzindo o perfil epidemiológico inicialmente observado nos Estados Unidos. A resposta brasileira à epidemia distinguiu-se pela mobilização precoce e vigorosa da sociedade civil organizada. Herbert de Souza, conhecido como Betinho, sociólogo hemofílico que contraiu HIV por meio de transfusões sanguíneas, fundou a Associação Brasileira Interdisciplinar de AIDS (ABIA) em 1986 e tornou-se símbolo nacional da luta contra a doença ao assumir publicamente sua condição (Hamman, 1994). Seus irmãos Henfil, cartunista consagrado, e Chico Mário, músico, também hemofílicos, morreram de complicações da Aids em 1988. O Grupo de Apoio à Prevenção à Aids de São Paulo (GAPA-SP), fundado em 1985, foi a primeira organização não governamental

[3]Os percentuais relativos às metas 95-95-95 são estimativas produzidas pelo UNAIDS por meio do *software* Spectrum, um conjunto de ferramentas de modelagem matemática que integra dados de pesquisas populacionais, registros de programas de testagem e tratamento, e modelos epidemiológicos para estimar o número total de pessoas vivendo com HIV, incluindo casos não diagnosticados (Schalkwyk *et al.*, 2024).

dedicada ao enfrentamento da Aids no Brasil, produzindo o primeiro material educativo de prevenção com o slogan "Transe numa boa". O Grupo Pela VIDDA, fundado por Herbert Daniel em 1989, foi a primeira organização brasileira criada por pessoas vivendo com HIV, desenvolvendo o conceito de "morte civil" para descrever a exclusão social sofrida pelos soropositivos antes mesmo da morte biológica.

O marco legal definidor da resposta brasileira foi a Lei nº 9.313, de 13 de novembro de 1996, que garantiu a distribuição gratuita e universal de medicamentos antirretrovirais pelo Sistema Único de Saúde (SUS) (Brasil, 1996). O Artigo 1º da lei estabelece que os portadores do HIV e doentes de Aids receberão, gratuitamente, do Sistema Único de Saúde, toda a medicação necessária a seu tratamento. Esta política tornou o Brasil o primeiro país em desenvolvimento a implementar o programa de acesso universal à TARV, constituindo referência internacional que influenciou programas semelhantes em diversas nações de renda média e baixa (Galvão, 2002). De acordo com o Ministério da Saúde (Brasil. Ministério da Saúde, 2024a), o acesso universal e o monitoramento clínico contínuo permitiram a redução expressiva da mortalidade e da morbidade associadas à infecção, consolidando o tratamento como componente central das estratégias de prevenção e cuidado integral. Esse processo refletiu o princípio da universalidade do SUS e materializou uma política de equidade que influenciou programas semelhantes em diversos países de renda média. Para sustentar financeiramente o programa de distribuição universal, o Brasil desenvolveu a estratégia pioneira de produção nacional de medicamentos genéricos por meio de laboratórios públicos como Farmanguinhos (vinculado à Fiocruz) e o Laboratório Farmacêutico do Estado de Pernambuco (Lafepe), além de enfrentar disputas comerciais e diplomáticas relacionadas a patentes farmacêuticas (Cassier; Correa, 2003). Em maio de 2007, o presidente Luiz Inácio Lula da Silva assinou a primeira licença compulsória de medicamento na história do país, quebrando a patente do efavirenz, antirretroviral produzido pela multinacional Merck (Rodrigues; Soler, 2009). A medida reduziu o preço do medicamento de US$ 1,59 para US$ 0,45 por comprimido, com economia projetada de US$ 236,8 milhões até 2012. Essa decisão reafirmou a prevalência do direito à saúde sobre interesses comerciais e fortaleceu a posição brasileira em negociações internacionais sobre propriedade intelectual de medicamentos essenciais.

O país alcançou, em 2023, a menor taxa de mortalidade por Aids da série histórica: 3,9 óbitos por 100 mil habitantes, representando queda de 32,9% em relação a 2013 (Brasil. Ministério da Saúde, 2024b). Estima-se que aproximadamente 1 milhão de brasileiros vivam atualmente com HIV, dos quais 96% conhecem seu diagnóstico, superando a meta

de 95% estabelecida pelo UNAIDS. A introdução do dolutegravir como medicamento de primeira linha no Brasil, em 2017, representou avanço terapêutico significativo. O país foi um dos primeiros a adotar esse antirretroviral em larga escala, negociando redução de 70% no preço em relação aos valores praticados internacionalmente. Estudos conduzidos em coortes brasileiras demonstraram que o esquema baseado em dolutegravir foi significativamente mais eficaz na supressão viral quando comparado ao regime anterior com efavirenz (Silva *et al.*, 2023). Persistem, contudo, desigualdades regionais e sociodemográficas significativas na distribuição da epidemia. A maior concentração de casos encontra-se nas regiões Sudeste e Sul, enquanto que as maiores taxas de detecção são observadas nas regiões Norte e Sul. O perfil epidemiológico brasileiro apresenta predomínio masculino, com 69% dos casos de HIV ocorrendo em homens, e 52% das infecções concentradas entre homens que fazem sexo com homens (HSH). A população negra (pretos e pardos) representa 58,2% dos casos, proporção superior à sua participação na população geral, evidenciando a intersecção entre vulnerabilidade racial e exposição ao HIV (Brasil. Ministério da Saúde, 2024b).

A história da epidemia de HIV/Aids é indissociável do estigma que a acompanhou desde os primeiros casos. A associação inicial da doença com homossexualidade masculina cristalizou preconceitos que retardaram significativamente a resposta governamental em diversos países (Parker; Aggleton, 2003). Nos Estados Unidos, o governo Reagan demorou a mobilizar recursos adequados para o enfrentamento da epidemia. O presidente Reagan mencionou publicamente a Aids pela primeira vez apenas em 1987, seis anos após os primeiros casos documentados (Shilts, 2007). Segmentos do conservadorismo religioso caracterizavam a doença como "punição divina" contra comportamentos considerados imorais, contribuindo para o atraso na resposta institucional. A desinformação sobre as origens do HIV representou outro obstáculo significativo ao enfrentamento da epidemia. A chamada Operação INFEKTION, campanha de desinformação conduzida pela KGB soviética, disseminou a partir de 1983 a teoria conspiratória de que a Aids teria sido criada em laboratório militar americano como arma biológica (Boghardt, 2009). Em 1987, essa narrativa falsa havia sido publicada em mais de 80 países e traduzida para 30 idiomas, alcançando audiência global. O impacto dessa operação persiste décadas depois: uma pesquisa realizada em 2005 revelou que aproximadamente 50% dos afro-americanos acreditavam que a Aids foi deliberadamente criada em laboratório (Bogart; Thorburn, 2005). A ciência, contudo, estabeleceu de forma conclusiva a origem zoonótica do HIV. Estudos filogenéticos demonstraram que o HIV-1, responsável pela pandemia global,

originou-se do SIVcpz (vírus da imunodeficiência símia) presente em chimpanzés (*Pan troglodytes troglodytes*) no sudeste de Camarões, provavelmente na década de 1920, por meio da caça e consumo de carne de animais selvagens (Sharp; Hahn, 2011).

O caso mais devastador de negacionismo institucional na história da epidemia foi protagonizado pelo presidente Thabo Mbeki, da África do Sul, que governou entre 1999 e 2008 (Nattrass, 2007). Influenciado pelo biólogo americano Peter Duesberg, que alegava que o HIV não causa Aids, Mbeki descreveu os medicamentos antirretrovirais como "venenos" e sua Ministra da Saúde, Manto Tshabalala-Msimang, promoveu beterraba, alho e limão como alternativas ao tratamento convencional. Estima-se que as políticas negacionistas de Mbeki resultaram em aproximadamente 330 mil mortes prematuras evitáveis e 35 mil bebês nascidos com HIV que poderiam ter sido prevenidos entre 2000 e 2005 (Chigwedere *et al.*, 2008). O negacionismo relacionado ao HIV experimentou um ressurgimento preocupante após a pandemia de COVID-19, com convergência entre diferentes movimentos de ceticismo científico e ampla disseminação de desinformação em plataformas digitais (Smith, 2025). Redes sociais amplificam a desinformação sobre supostos efeitos colaterais da Profilaxia Pré-Exposição (PrEP) e promovem o abandono do tratamento antirretroviral. Um estudo conduzido em 2010 demonstrou que a crença em retórica negacionista está associada à recusa de medicação, maior frequência de hospitalizações e cargas virais detectáveis entre pessoas vivendo com HIV (Kalichman; Eaton; Cherry, 2010). O estigma persistente afeta todas as etapas da cascata do cuidado em HIV. No Brasil, apesar dos avanços nas políticas públicas, o estigma continua a impactar a adesão ao tratamento, o acesso aos serviços e a qualidade de vida de quem vive com HIV, constituindo barreira estrutural que transcende a oferta de medicamentos.

Apesar dos avanços inquestionáveis nas últimas décadas, a manutenção da adesão terapêutica ao longo do tempo permanece como um dos maiores desafios contemporâneos para o controle da epidemia. O sucesso do tratamento depende não apenas da disponibilidade dos medicamentos, mas também da continuidade do uso e do acompanhamento clínico regular. A interrupção da TARV compromete o controle virológico, favorece o surgimento de resistência medicamentosa e amplia o risco de transmissão do vírus, impactando diretamente no cumprimento das metas de saúde pública voltadas à eliminação do HIV como problema global (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2025; Costa *et al.*, 2021). No Brasil, embora os índices de acesso à TARV sejam elevados, a descontinuidade do tratamento ainda atinge proporções significativas, evidenciando desigualdades regionais, raciais e socioeconômicas que refletem a persistência

de vulnerabilidades estruturais (Cunha *et al.*, 2025; Brito; Castilho; Szwarcwald, 2001).

Paralelamente ao avanço da TARV, o Brasil tem expandido progressivamente seu arsenal de estratégias de prevenção combinada contra o HIV, incorporando intervenções biomédicas que complementam o uso tradicional de preservativos. A Profilaxia Pós-Exposição (PEP), disponível no SUS desde 1999, consiste no uso emergencial de medicamentos antirretrovirais por 28 dias após situações de risco, devendo ser iniciada preferencialmente nas primeiras horas e no máximo até 72 horas após a exposição ao vírus (Brasil. Ministério da Saúde, 2024c). Em dezembro de 2017, o Brasil incorporou ao SUS a Profilaxia Pré-Exposição (PrEP), tornando-se pioneiro na América Latina na oferta pública dessa estratégia preventiva (Brasil. Ministério da Saúde, 2025). A PrEP consiste no uso contínuo de medicamentos antirretrovirais por pessoas não infectadas em situação de maior vulnerabilidade, reduzindo em mais de 90% o risco de infecção quando utilizada adequadamente. Estudos internacionais, como o iPrEx (Grant *et al.*, 2010) e o PROUD (McCormack *et al.*, 2016), demonstraram eficácia consistente da estratégia. Em 2024, o Brasil alcançou mais de 140 mil pessoas em uso regular de PrEP, representando expansão de mais de 530% desde 2018 (Brasil. Ministério da Saúde, 2024d).

O conceito científico *Indetectável = Intransmissível* (I=I), fundamentado nos estudos PARTNER (Rodger *et al.*, 2016) e HPTN 052 (Cohen *et al.*, 2011), revolucionou a compreensão sobre transmissão e tratamento. Esses estudos demonstraram que pessoas vivendo com HIV que mantêm carga viral indetectável por meio do tratamento adequado não transmitem o vírus por via sexual. Essa evidência científica foi formalmente incorporada às políticas brasileiras de enfrentamento da epidemia, constituindo um instrumento tanto de prevenção quanto de combate ao estigma. Inovações tecnológicas recentes prometem revolucionar ainda mais o campo da prevenção. Em 2023, a Agência Nacional de Vigilância Sanitária (ANVISA) aprovou o cabotegravir injetável de longa duração (CAB-LA), primeira formulação de PrEP administrada a cada dois meses, que demonstrou eficácia superior a 95% em ensaios clínicos internacionais e maior adesão quando comparada à PrEP oral (Landovitz *et al.*, 2021; Delany-Moretlwe *et al.*, 2022). Adicionalmente, encontra-se em desenvolvimento o lenacapavir, antirretroviral injetável de administração semestral que apresentou proteção próxima a 100% em estudos com mulheres cisgênero (Bekker *et al.*, 2024), representando potencial avanço significativo no controle da epidemia. Essas tecnologias de longa duração são particularmente promissoras para superar barreiras de adesão associadas ao uso diário de medicamentos, especialmente entre populações jovens e em situação de maior vulnerabilidade social. A integração

dessas estratégias preventivas ao tratamento universal representa um esforço coordenado para alcançar as metas estabelecidas pelo Programa Conjunto das Nações Unidas sobre HIV/Aids (UNAIDS) de eliminação do HIV como problema de saúde pública até 2030.

1.2 Tempo até Interrupção da TARV

O estudo do tempo até a interrupção da TARV, entendido como o período (em anos) entre o início do tratamento e sua descontinuidade, permite compreender de maneira quantitativa a dinâmica da adesão terapêutica e os fatores que influenciam sua manutenção. Seja $T \geq 0$ a variável aleatória que representa o tempo até o evento que se deseja estudar (interrupção do tratamento), e seja $\delta \in \{0, 1\}$ o indicador de evento, tal que $\delta = 1$ denote a ocorrência da interrupção e $\delta = 0$ denote que ocorreu a censura de T, ou seja, que o tratamento não foi interrompido. A partir dessas definições, a função de sobrevivência é definida como $S(t) = P(T > t)$, enquanto a função de risco é definida como $h(t) = f(t)/S(t)$, expressando a taxa instantânea de ocorrência do evento em t, condicionada à sobrevivência até esse tempo (Kleinbaum; Klein, 2020; Colosimo; Giolo, 2024). Esse arcabouço permite estimar probabilidades condicionais de permanência em tratamento e quantificar diferenças entre subgrupos sociodemográficos e clínicos.

A análise de sobrevivência é amplamente utilizada em estudos clínicos e epidemiológicos, constituindo um modelo de referência para fenômenos em que o desfecho é temporal e sujeito à censura. Contudo, os modelos clássicos, como o de riscos proporcionais de Cox (Cox, 1972; Kalbfleisch; Prentice, 2002), pressupõem relações lineares e efeitos constantes das covariáveis ao longo do tempo, suposições frequentemente violadas em bases populacionais extensas e heterogêneas. No decorrer do trabalho, verificou-se que variáveis como raça/cor, escolaridade e faixa etária apresentam padrões de interrupção não proporcionais ao tempo de seguimento, o que indica a necessidade de abordagens mais flexíveis e adaptativas.

No estudo realizado neste trabalho, foi adotada uma arquitetura neural motivada por três características da base de dados que impuseram limitações às abordagens clássicas de análise de sobrevivência: (i) a estrutura temporal discreta dos registros administrativos do SUS, que registram tempos de acompanhamento em anos completos, tornando artificial a imposição de modelos contínuos; (ii) a elevada taxa de censura à direita (65,3% dos indivíduos), que compromete a estabilidade de estimadores baseados em verossimilhança parcial em horizontes longos (Kalbfleisch; Prentice, 2002); e (iii) as evidências de não

proporcionalidade dos riscos identificadas na análise exploratória, nas quais curvas de Kaplan-Meier estratificadas por raça/cor, escolaridade e faixa etária revelaram cruzamentos e padrões não paralelos incompatíveis com a suposição fundamental do modelo de Cox. Diante desse cenário, a adoção de uma abordagem capaz de contornar essas limitações sem impor suposições paramétricas restritivas tornou-se não apenas conveniente, mas metodologicamente necessária.

O contexto contemporâneo de disponibilidade massiva de dados em saúde, aliado ao avanço das técnicas de aprendizado de máquina, tem impulsionado o desenvolvimento de modelos capazes de representar de forma mais acurada essas relações. O aprendizado de máquina, e em especial o aprendizado profundo (*deep learning*), permite construir funções $f(\mathbf{x})$ que mapeiam covariáveis $\mathbf{x}$ em estimativas de risco ou probabilidade, sem impor uma forma funcional rígida, capturando interações de alta ordem e padrões complexos (Hastie; Tibshirani; Friedman, 2009; Goodfellow; Bengio; Courville, 2016). Entre os modelos aplicados à sobrevivência, destacam-se as redes neurais artificiais, que vêm sendo progressivamente adaptadas para lidar com dados censurados e temporais (Katzman *et al.*, 2018; Kvamme; Borgan; Scheel, 2019). Essa abordagem é particularmente promissora no contexto da TARV, em que fatores clínicos, demográficos e sociais interagem de forma dinâmica e não linear, influenciando o tempo até a interrupção do tratamento.

A modelagem de sobrevivência com redes neurais artificiais representa, portanto, uma nova fronteira metodológica para a análise de dados longitudinais de saúde. Ao incorporar técnicas de otimização e regularização próprias do aprendizado profundo, esses modelos conseguem lidar com grandes volumes de dados e capturar padrões complexos sem impor suposições paramétricas rígidas sobre a forma de $h(t)$ ou $S(t)$. Em particular, o modelo Neural Net-Survival (*Nnet-Survival*), proposto por Gensheimer e Narasimhan (2019), foi desenvolvido para dados de tempo discreto. O período de acompanhamento é dividido em K intervalos de igual tamanho, e, considerando que T_i assume valores inteiros correspondentes a esses intervalos, o modelo estima diretamente o risco condicional por intervalo $h_{ik} = P(T_i \in [t_{k-1}, t_k) \mid T_i \geq t_{k-1}, \mathbf{x}_i)$, para o i-ésimo indivíduo, e reconstrói a sobrevivência individual $S_i(t_k \mid \mathbf{x}_i) = \prod_{j=1}^{k}(1-h_{ij})$. Essa formulação oferece vantagens em termos de estabilidade numérica, compatibilidade com censura à direita e interpretabilidade das probabilidades preditas.

A aplicação do *Nnet-Survival* neste estudo justifica-se pela convergência entre as características da base analisada e as propriedades do modelo. A natureza discreta do tempo de acompanhamento (em anos completos) é intrinsecamente compatível com a

formulação do modelo, dispensando aproximações contínuas que introduziriam viés de discretização. Para a condução do estudo, foi utilizada uma base de dados com mais de um milhão de indivíduos em terapia antirretroviral no Brasil, sobre os quais estão disponíveis informações acerca da ocorrência ou não da interrupção do tratamento, do tempo até essa interrupção quando ocorreu, além de diversas covariáveis sociodemográficas e clínicas. A alta dimensionalidade resultante, com 19 variáveis preditoras após o tratamento apropriado dos dados, é adequadamente tratada pelas técnicas de regularização próprias do aprendizado profundo (*dropout*[4] e penalização L2), que controlam a complexidade do modelo e previnem sobreajuste[5]. A elevada taxa de censura (65,3%), que comprometeria a estabilidade de estimadores clássicos em horizontes longos, é naturalmente incorporada pela função de perda baseada em verossimilhança de tempo discreto, que atribui contribuições diferenciadas a eventos e censuras. Por fim, ao estimar diretamente o risco condicional por intervalo sem impor forma funcional ao preditor, o modelo dispensa tanto a suposição de linearidade dos efeitos quanto a de proporcionalidade dos riscos, ambas violadas pelos dados observados.

O presente trabalho propõe aplicar o *Nnet-Survival* para investigar os determinantes sociodemográficos e clínicos associados ao tempo até a interrupção da TARV no Brasil. A abordagem combina o rigor estatístico da análise de sobrevivência com a capacidade preditiva das redes neurais, oferecendo uma visão abrangente e moderna sobre os fatores que interferem na continuidade terapêutica. A aplicação dessa metodologia a uma base nacional de larga escala representa contribuição relevante tanto no campo da epidemiologia quanto na estatística aplicada, reforçando o potencial das técnicas de aprendizado de máquina na análise de fenômenos de saúde pública complexos.

O problema de pesquisa que orienta este estudo pode ser formulado da seguinte maneira: dado um vetor de covariáveis $\mathbf{x}$ composto por características sociodemográficas e clínicas, quais fatores influenciam o tempo T até a interrupção da TARV no Brasil e em que medida um modelo de redes neurais aplicado à análise de sobrevivência é capaz de representar adequadamente essas relações? Essa questão reflete a preocupação central do trabalho: compreender como desigualdades estruturais e condições individuais de saúde interagem para determinar a continuidade ou a interrupção do tratamento antirretroviral

[4] *Dropout* é uma técnica de regularização que, durante o treinamento, desativa aleatoriamente uma fração dos neurônios em cada iteração, reduzindo a co-adaptação entre unidades e prevenindo sobreajuste.

[5] Sobreajuste (*overfitting*) ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, capturando ruídos e particularidades da amostra em vez de padrões generalizáveis, o que resulta em desempenho inferior em dados não observados.

em nível populacional.

A justificativa deste estudo assenta-se em dois eixos principais. O primeiro é de natureza epidemiológica e social, pois compreender os fatores determinantes para ocorrência da interrupção terapêutica é fundamental para aprimorar as políticas de adesão e retenção em cuidado, contribuindo para as metas de controle do HIV/Aids estabelecidas pelo Ministério da Saúde e por organismos internacionais (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2025; Brasil. Ministério da Saúde, 2024a). O segundo eixo é metodológico: a aplicação de redes neurais à análise de sobrevivência ainda é incipiente no Brasil, especialmente em estudos populacionais baseados em registros administrativos. Assim, o trabalho busca demonstrar o potencial dessas ferramentas na modelagem de grandes coortes de pacientes, explorando sua capacidade de generalização e a robustez de suas estimativas em contextos de censura elevada. Conforme será demonstrado no Capítulo 5, o modelo implementado apresentou boa capacidade de discriminação entre indivíduos com maior e menor risco de interrupção, além de produzir estimativas de probabilidade bem calibradas ao longo do tempo, com desempenho compatível com os reportados por Katzman *et al.* (2018) e Kvamme, Borgan e Scheel (2019) em aplicações similares. Esses resultados validam empiricamente a adequação da abordagem ao problema em questão e demonstram que a arquitetura neural conseguiu explorar a flexibilidade necessária sem sacrificar a capacidade de generalização.

1.3 Relevância das Técnicas de Análise de Sobrevivência

É importante ressaltar que os métodos de análise de sobrevivência não se restringem ao contexto biomédico, apresentando aplicabilidade transversal em múltiplas disciplinas científicas e campos profissionais. O estudo do tempo até a ocorrência de eventos específicos, considerando adequadamente a presença de observações censuradas, constitui um problema analítico comum a diversas áreas do conhecimento humano. Uma das grandes vantagens dessas técnicas reside justamente em sua capacidade de lidar com informações incompletas sobre o tempo de ocorrência de eventos, situação ubíqua em estudos longitudinais nos quais nem sempre é possível acompanhar todos os indivíduos até a materialização do desfecho de interesse (Colosimo; Giolo, 2024). Ao incorporar adequadamente dados censurados na modelagem estatística, a análise de sobrevivência permite obter estimativas mais precisas, robustas e não viesadas dos parâmetros de

interesse, possibilitando inferências válidas mesmo diante de informações parcialmente observadas.

Na área da saúde, a análise de sobrevivência desempenha papel fundamental na compreensão da história natural de doenças, na avaliação de tratamentos e intervenções terapêuticas, e no estudo de disparidades em desfechos clínicos entre diferentes grupos populacionais. Estudos brasileiros têm empregado essas metodologias para investigar a sobrevida de pacientes oncológicos (Bustamante-Teixeira; Faerstein; Latorre, 2002), examinar disparidades raciais na sobrevivência de mulheres com câncer de mama (Nogueira *et al.*, 2018), e avaliar a efetividade da implementação de programas de saúde materno-infantil (Hartz, 1997). Tais aplicações demonstram a relevância dessas técnicas para a produção de evidências científicas que subsidiam políticas públicas de saúde e contribuem para a redução de iniquidades no acesso e nos resultados dos cuidados de saúde. Na engenharia de confiabilidade, investiga-se o tempo de operação até a falha de componentes mecânicos, eletrônicos ou estruturais, permitindo o desenvolvimento de estratégias otimizadas de manutenção preventiva e a estimação da vida útil de equipamentos industriais. Pesquisas nacionais têm aplicado modelos de sobrevivência para analisar o desempenho e a confiabilidade de aerogeradores, identificando fatores que impactam sua vida útil e subsidiam decisões sobre manutenções e substituições (Xavier, 2022). No âmbito das ciências sociais aplicadas, técnicas de sobrevivência têm sido empregadas para modelar fenômenos temporais diversos, incluindo estudos sobre os padrões de consumo e suas implicações socioambientais (Mello; Sathler, 2015), a aplicação de métodos quantitativos na pesquisa social brasileira (Bachini; Chicarino, 2018), e análises demográficas sobre envelhecimento populacional e suas consequências socioeconômicas. No contexto educacional, a metodologia tem sido particularmente útil para examinar a progressão acadêmica e os determinantes da evasão estudantil em instituições de ensino superior brasileiras, comparando trajetórias de diferentes perfis de estudantes e identificando fatores de risco para o abandono dos cursos (Junior; Ferreira, 2014; Klitzke; Carvalhaes, 2023; Silva, 2013).

Todavia, é no campo das ciências atuariais e na gestão de riscos financeiros que a análise de sobrevivência encontra algumas de suas aplicações mais consolidadas, estrategicamente relevantes e historicamente fundamentais, constituindo verdadeiramente o alicerce metodológico sobre o qual se erguem os pilares da ciência atuarial moderna. A profissão atuarial, desde suas origens históricas no século XVII com o desenvolvimento das primeiras tábuas de mortalidade demográficas (Graunt, 1662; Halley, 1693; Hald,

1990), está intrinsecamente vinculada à necessidade de quantificar matematicamente a incerteza temporal associada a eventos futuros, particularmente aqueles relacionados à longevidade humana, à morbidade e a outros riscos biométricos. A construção de tábuas de mortalidade e de sobrevivência constitui a base essencial para a operação técnica de seguradoras de vida, entidades de previdência complementar e fundos de pensão, permitindo a precificação adequada de produtos que envolvem cobertura de riscos vitais, o dimensionamento correto de reservas técnicas e provisões matemáticas, e a garantia da solvência de longo prazo dessas instituições (Macedo; Silva; Santos, 2006; Bowers *et al.*, 1997). As técnicas de graduação de tábuas, os métodos de extrapolação para idades avançadas e os modelos estocásticos de mortalidade representam refinamentos contemporâneos da análise de sobrevivência aplicada ao contexto atuarial, incorporando incertezas paramétricas, tendências temporais de melhoria da longevidade e variabilidade estocástica nas projeções de sobrevivência futura.

No mercado de seguros e resseguros, a modelagem de sobrevivência transcende a mortalidade humana, abrangendo também a análise de outros eventos seguráveis ao longo do tempo. Seguros de invalidez, por exemplo, requerem a estimação conjunta do tempo até a ocorrência de incapacidade laboral e do tempo de permanência no estado de invalidez antes da recuperação ou óbito. Seguros de doenças graves demandam análises de incidência de eventos críticos específicos (como infarto do miocárdio, acidente vascular cerebral ou diagnóstico de câncer) e da sobrevivência condicional após tais diagnósticos. A gestão do risco de longevidade, particularmente relevante para anuidades e planos de benefício definido, depende criticamente de projeções acuradas da sobrevivência em idades avançadas, nas quais pequenas diferenças nas premissas atuariais podem resultar em impactos financeiros substanciais sobre os compromissos de longo prazo das entidades. A análise de sobrevivência multivariada, que considera a dependência entre múltiplas vidas (como em seguros de sobrevivência conjugal ou pensões reversíveis), representa uma extensão metodológica particularmente relevante para produtos que envolvem benefícios contingentes à ordem de falecimento de segurados.

O setor bancário e de crédito constitui outro domínio no qual a análise de sobrevivência desempenha função estratégica fundamental. A modelagem do tempo até a inadimplência ou *default* de tomadores de empréstimos permite às instituições financeiras avaliar e precificar o risco de crédito, estabelecer limites de exposição, definir taxas de juros ajustadas ao risco e estimar perdas esperadas para fins de provisionamento e adequação de capital regulatório sob os marcos de Basileia (Dirick; Claeskens; Baesens, 2017; Narain, 2004).

Estudos sobre a evolução do sistema bancário brasileiro após a estabilização monetária e as reformas da década de 1990 têm empregado técnicas de análise de sobrevivência para examinar mudanças estruturais no setor (Camargo, 2009), analisar a dinâmica da expansão do crédito e o papel dos bancos públicos (Paula; Oreiro; Basilio, 2013), evidenciando a relevância dessas metodologias para a compreensão de fenômenos macroeconômicos e da dinâmica competitiva no mercado financeiro. A análise de sobrevivência aplicada ao risco de crédito permite incorporar covariáveis macroeconômicas variantes no tempo, características socioeconômicas dos tomadores e histórico de relacionamento com a instituição, produzindo modelos preditivos que suportam decisões de concessão de crédito e estratégias de cobrança diferenciadas conforme o perfil de risco (Bellotti; Crook, 2009).

Para além dos riscos tradicionais de morte, invalidez e crédito, a análise de sobrevivência encontra aplicações emergentes em outros contextos relevantes para atuários e gestores de risco. No estudo do cancelamento de contratos de seguros (*lapse risk*), planos de saúde ou serviços de assinatura (*churn*), modelos de sobrevivência permitem identificar determinantes do tempo até a descontinuidade da relação contratual, subsidiando estratégias de retenção de clientes, precificação dinâmica e projeções de receitas recorrentes (Milhaud; Dutang, 2018; Eling; Kiesenbauer, 2014). A persistência de apólices é particularmente relevante para a rentabilidade de produtos de seguro de vida com elevados custos de aquisição, nos quais cancelamentos precoces podem comprometer a viabilidade econômica da operação (Eling; Kiesenbauer, 2014). A modelagem do tempo até sinistros em seguros patrimoniais ou de responsabilidade civil,embora menos convencional do que em aplicações de seguros de pessoas,também pode se beneficiar de abordagens de sobrevivência quando o interesse está em caracterizar o tempo de exposição ao risco antes da materialização do sinistro. No contexto de riscos operacionais e de *compliance*, a análise de sobrevivência pode ser aplicada ao estudo do tempo até a ocorrência de eventos de perda operacional, fraudes ou violações regulatórias, permitindo a quantificação desses riscos e a alocação de capital regulatório conforme *frameworks* como Basileia e Solvência II.

Dessa forma, a análise de sobrevivência constitui um arcabouço metodológico de relevância central e transversal para os mais diversos profissionais que lidam com modelagem de fenômenos temporais e quantificação de incertezas, destacando-se particularmente sua importância para a formação e a prática profissional de atuários. O domínio dessas técnicas estatísticas é absolutamente essencial para que atuários possam desempenhar adequadamente suas funções de mensuração e gestão de riscos, precificação de produtos, dimensionamento de reservas técnicas, avaliação de solvência e assessoria técnica em

diversos segmentos do mercado segurador, previdenciário e financeiro (Colosimo; Giolo, 2024). A crescente sofisticação dos mercados, a disponibilidade de grandes volumes de dados longitudinais, o advento de técnicas computacionais avançadas e as exigências regulatórias cada vez mais rigorosas em termos de modelagem de riscos tornam imperativo que atuários mantenham-se atualizados não apenas quanto aos métodos clássicos de análise de sobrevivência, mas também quanto a desenvolvimentos contemporâneos que incorporam aprendizado de máquina, redes neurais e outras abordagens de inteligência artificial à modelagem de tempo até eventos.

Assim, o objetivo geral deste trabalho é analisar os fatores associados ao tempo até a interrupção da terapia antirretroviral no Brasil, utilizando técnicas de análise de sobrevivência baseadas em redes neurais artificiais. A partir desse objetivo central, derivam-se metas específicas que envolvem: (i) realizar uma revisão de literatura robusta sobre as técnicas de análise de sobrevivência, com ênfase nas abordagens baseadas em redes neurais artificiais e suas aplicações em estudos de saúde pública; (ii) descrever o perfil sociodemográfico e clínico da população estudada; (iii) avaliar empiricamente a distribuição dos tempos de interrupção; e (iv) implementar um modelo de sobrevivência neural capaz de estimar probabilidades condicionais de manutenção do tratamento em diferentes horizontes temporais. A combinação entre análise estatística tradicional e modelagem preditiva visa produzir resultados que sejam, ao mesmo tempo, interpretáveis e operacionalmente úteis para a formulação de estratégias de saúde pública.

A estrutura deste trabalho está organizada em seis capítulos. O Capítulo 2 apresenta os fundamentos da análise de sobrevivência, incluindo conceitos básicos, funções fundamentais, estimadores não paramétricos clássicos (Kaplan-Meier e Nelson-Aalen) e modelos de regressão paramétricos e semiparamétricos. O Capítulo 3 introduz os conceitos de aprendizado de máquina e redes neurais artificiais aplicados à análise de sobrevivência, culminando na apresentação do modelo *Nnet-Survival* e das métricas de avaliação utilizadas neste estudo. O Capítulo 4 introduz em detalhes o conjunto de dados utilizado no estudo e realiza a análise exploratória deles, caracterizando a coorte de tratamento e descrevendo as variáveis utilizadas na modelagem. O Capítulo 5 expõe os procedimentos de implementação do modelo *Nnet-Survival*, o processo de treinamento, as métricas de desempenho e os resultados obtidos. Por fim, o Capítulo 6 apresenta a discussão geral dos resultados e as considerações finais, destacando as implicações epidemiológicas e metodológicas dos achados e apontando caminhos para futuras pesquisas.

Em síntese, esta introdução situa o presente estudo no cruzamento entre estatística

aplicada, epidemiologia e ciência de dados. A trajetória histórica da epidemia de HIV/Aids, desde os primeiros casos desconhecidos e estigmatizados em 1981 até a consolidação da TARV como tratamento transformador, demonstra como o conhecimento científico, aliado a políticas públicas fundamentadas em direitos humanos, pode alterar radicalmente o prognóstico de uma doença. A proposta de integrar o arcabouço formal da análise de sobrevivência à flexibilidade das redes neurais artificiais busca ampliar as possibilidades analíticas e oferecer uma compreensão mais profunda dos determinantes sociais e clínicos que condicionam a adesão terapêutica e a continuidade da TARV no Brasil, ao mesmo tempo em que demonstra a aplicabilidade de técnicas computacionais avançadas a problemas de análise de sobrevivência, contribuindo para a formação de profissionais capazes de empregar ferramentas modernas na modelagem de riscos e fenômenos temporais em diversos contextos aplicados.

Capítulo 2

Fundamentos da Análise de Sobrevivência

A compreensão dos mecanismos que regem o tempo até a interrupção da terapia antirretroviral requer a aplicação de métodos estatísticos especializados capazes de lidar com a natureza temporal dos dados e com a presença de censura. A análise de sobrevivência fornece o arcabouço matemático e estatístico necessário para descrever o comportamento de variáveis aleatórias que descrevem o tempo até a ocorrência de eventos de interesse, permitindo a estimação de probabilidades de sobrevivência e a identificação de fatores associados ao risco. Este capítulo apresenta os fundamentos teóricos da análise de sobrevivência, introduzindo conceitos, notações e métodos que fundamentam a metodologia adotada neste estudo.

2.1 Conceitos Básicos e Notação

A análise de sobrevivência é o ramo da estatística que estuda o tempo até a ocorrência de um evento aleatório. Seja $T \geq 0$ uma variável aleatória contínua representando o tempo até a ocorrência desse evento (neste caso, a interrupção da terapia antirretroviral). O valor observado para o indivíduo i é denotado por t_i. Nem todos os indivíduos experimentam o evento durante o acompanhamento: quando o evento não é observado, o tempo é *censurado à direita*. Para cada observação, define-se um indicador de evento

$$\delta_i = \begin{cases} 1, & \text{se o evento foi observado;} \\ 0, & \text{se o evento não foi observado durante o acompanhamento (censura).} \end{cases}$$

As propriedades probabilísticas de T são descritas por quatro funções principais que se inter-relacionam matematicamente:

- A **função de distribuição acumulada**, $F(t) = P(T \leq t)$, representa a probabilidade do evento de interesse ocorrer até o tempo t.
- A **função densidade de probabilidade**, $f(t) = \frac{d}{dt}F(t)$, expressa a taxa de ocorrência do evento de interesse ao longo do tempo.
- A **função de sobrevivência**, $S(t) = P(T > t) = 1 - F(t)$, representa a probabilidade do evento ainda não ter ocorrido até o tempo t. Por definição, $S(0) = 1$ e $\lim_{t\to\infty} S(t) = 0$.
- A **função de risco** ou *hazard function*, $h(t)$, quantifica a taxa instantânea de ocorrência do evento no tempo t, condicionada à sobrevivência até t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

O **risco acumulado** $H(t)$ é dado por

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

de modo que $S(t) = \exp[-H(t)]$. Essas relações são fundamentais e permeiam toda a análise de sobrevivência, conectando diferentes representações do mesmo fenômeno probabilístico. Mais detalhes sobre a teoria de análise de sobrevivência podem ser encontrados em Colosimo e Giolo (2024) e Kleinbaum e Klein (2020).

2.2 Estimadores Não Paramétricos

Na ausência de suposições sobre a forma paramétrica de $S(t)$ ou $h(t)$, os estimadores não paramétricos constituem ferramentas essenciais para a descrição empírica dos dados de sobrevivência. Esses métodos são especialmente valiosos em análises exploratórias e quando não há justificativa teórica para assumir uma distribuição específica para os tempos até ocorrência dos eventos de interesse.

2.2.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier (Kaplan; Meier, 1958), também conhecido como estimador produto-limite, é o método não paramétrico mais amplamente utilizado para estimar a função de sobrevivência na presença de censura à direita. Seja $t_1 < t_2 < \cdots < t_m$ os m tempos distintos em que pelo menos um evento foi observado. Para cada tempo t_j, define-se:

- d_j: número de eventos ocorridos em t_j;
- n_j: número de indivíduos em risco imediatamente antes de t_j (aqueles que não sofreram o evento nem foram censurados antes de t_j).

O estimador de Kaplan-Meier é então definido como:

$$\hat{S}_{KM}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

A quantidade $\left(1 - \frac{d_j}{n_j}\right)$ está associada a probabilidade condicional de sobreviver ao tempo t_j, dado que o indivíduo estava em risco em t_j. O estimador de Kaplan-Meier é uma função em escada (*step function*), com descontinuidades (saltos) nos tempos em que algum evento é observado.

A variância do estimador pode ser estimada pela fórmula de Greenwood (Greenwood, 1926):

$$\widehat{\text{Var}}[\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)},$$

permitindo a construção de intervalos de confiança para $S(t)$ e a realização de testes de hipóteses sobre diferenças entre curvas de sobrevivência.

O estimador de Kaplan-Meier é consistente, assintoticamente não viesado e apresenta propriedades de máxima verossimilhança sob o modelo não paramétrico (Fleming; Harrington, 2005). Sua principal vantagem reside na simplicidade conceitual e na capacidade de incorporar adequadamente observações censuradas sem impor restrições paramétricas sobre a forma de $S(t)$. Por essas razões, o estimador de Kaplan-Meier é rotineiramente empregado em análises descritivas de dados de sobrevivência, inclusive para a construção de curvas de sobrevivência estratificadas por subgrupos (como sexo, raça/cor ou faixa etária) e para a avaliação visual de diferenças entre grupos antes da modelagem formal.

2.2.2 Estimador de Nelson-Aalen

Enquanto o estimador de Kaplan-Meier foca na função de sobrevivência $S(t)$, o estimador de Nelson-Aalen (Nelson, 1969; Aalen, 1978) fornece uma estimativa não paramétrica para o risco acumulado $H(t)$. Utilizando a mesma notação introduzida nas Seções 2.1 e 2.2.1, o estimador de Nelson-Aalen é definido como:

$$\hat{H}_{NA}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j},$$

em que cada termo $\frac{d_j}{n_j}$ representa o incremento estimado no risco acumulado no tempo t_j.

A partir do estimador de Nelson-Aalen, pode-se obter uma estimativa alternativa para a função de sobrevivência por meio da relação $S(t) = \exp[-H(t)]$:

$$\hat{S}_{NA}(t) = \exp\left[-\hat{H}_{NA}(t)\right] = \exp\left[-\sum_{t_j \leq t} \frac{d_j}{n_j}\right].$$

Embora o estimador de Kaplan-Meier e o estimador derivado de Nelson-Aalen para $S(t)$ sejam assintoticamente equivalentes, eles diferem em amostras finitas. O estimador de Nelson-Aalen tende a apresentar menor viés em situações com muitos eventos e poucos indivíduos em risco, e é frequentemente preferido quando o interesse primário reside no risco acumulado $H(t)$ em vez de diretamente em $S(t)$ (Fleming; Harrington, 2005).

A variância do estimador de Nelson-Aalen pode ser estimada por:

$$\widehat{\text{Var}}[\hat{H}_{NA}(t)] = \sum_{t_j \leq t} \frac{d_j}{n_j^2},$$

permitindo inferências sobre $H(t)$ e, por extensão, sobre $S(t)$.

Ambos os estimadores, Kaplan-Meier e Nelson-Aalen, desempenham papéis complementares na análise de sobrevivência. O estimador de Kaplan-Meier é mais intuitivo e diretamente interpretável em termos de probabilidades de sobrevivência, sendo amplamente utilizado em apresentações gráficas e análises descritivas. O estimador de Nelson-Aalen, por sua vez, é conceitualmente mais próximo de modelos de processos de contagem e é preferido em contextos onde o foco reside na taxa de ocorrência acumulada de eventos ou quando se deseja maior estabilidade numérica em situações de amostragem esparsa (Andersen *et al.*, 1993).

2.3 Modelos Paramétricos e Semiparamétricos

2.3.1 Modelos Paramétricos

Nos modelos paramétricos, assume-se que a variável aleatória T segue uma distribuição de probabilidade conhecida, caracterizada por um número finito de parâmetros. A especificação paramétrica de $S(t)$, $h(t)$ e $f(t)$ permite a estimação por métodos de máxima verossimilhança, conferindo eficiência estatística às inferências quando as suposições distributivas são válidas. Esses modelos são particularmente úteis quando há evidências teóricas ou empíricas sobre a forma do risco ao longo do tempo, permitindo a incorporação de conhecimento prévio sobre o fenômeno estudado. A seguir, apresentam-se três das distribuições paramétricas mais utilizadas em análise de sobrevivência: exponencial, Weibull e log-normal.

Distribuição Exponencial

A distribuição exponencial constitui o modelo paramétrico mais simples, sendo caracterizada por uma função de risco constante ao longo do tempo. Formalmente, se $T \sim \text{Exp}(\lambda)$ com $\lambda > 0$, as funções fundamentais são dadas por:

$$h(t) = \lambda, \qquad S(t) = e^{-\lambda t}, \qquad f(t) = \lambda e^{-\lambda t}, \quad t \geq 0,$$

onde λ é o parâmetro de taxa (ou intensidade) de falha. A constância da função de risco implica que a taxa instantânea de ocorrência do evento é independente do tempo já decorrido, refletindo a propriedade de ausência de memória (*memoryless property*): para quaisquer $s, t \geq 0$, tem-se $P(T > s + t \mid T > s) = P(T > t)$. Essa propriedade torna o modelo exponencial matematicamente tratável, sendo frequentemente utilizado como primeira aproximação em análises de sobrevivência, especialmente em contextos nos quais não há razões teóricas para supor variação temporal no risco.

Distribuição Weibull

A distribuição Weibull generaliza a distribuição exponencial ao introduzir um parâmetro de forma que permite modelar diferentes padrões de evolução temporal do risco. Se $T \sim \text{Weibull}(\lambda, \gamma)$ com $\lambda > 0$ e $\gamma > 0$, as funções fundamentais são:

$$h(t) = \lambda\gamma t^{\gamma-1}, \qquad S(t) = e^{-\lambda t^{\gamma}}, \qquad f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}}, \quad t \geq 0,$$

onde λ é o parâmetro de escala e γ é o parâmetro de forma. O comportamento qualitativo da função de risco depende criticamente de γ:

- **Risco decrescente** ($\gamma < 1$): a função de risco é estritamente decrescente, apropriada para modelar fenômenos em que a taxa de falha diminui ao longo do tempo, como a mortalidade neonatal, em que o risco de óbito é elevado nos primeiros dias de vida e decresce rapidamente à medida que o recém-nascido se adapta ao ambiente extrauterino, ou a falha precoce de componentes eletrônicos, em que unidades defeituosas tendem a falhar no início da operação;

- **Risco constante** ($\gamma = 1$): a distribuição reduz-se ao caso exponencial, com $h(t) = \lambda$, sendo adequada para eventos que ocorrem de forma aleatória e independente do tempo já decorrido, como falhas em sistemas causadas por choques externos imprevisíveis ou a ocorrência de acidentes de trânsito em condições estáveis de tráfego;

- **Risco crescente** ($\gamma > 1$): a função de risco é estritamente crescente, adequada para processos de envelhecimento ou deterioração progressiva, como o desgaste mecânico de equipamentos industriais, em que a probabilidade de falha aumenta com o tempo de uso, ou a incidência de doenças crônico-degenerativas, cuja taxa cresce com o avanço da idade.

A versatilidade da distribuição Weibull, aliada à sua tratabilidade analítica e facilidade de interpretação, a torna uma escolha natural quando se deseja acomodar diferentes formas de função de risco sem impor restrições paramétricas excessivamente rígidas.

Distribuição Log-normal

A distribuição log-normal assume que o logaritmo natural do tempo de sobrevivência segue uma distribuição normal. Formalmente, T tem distribuição log-normal, denotada $T \sim \text{LN}(\mu, \sigma^2)$, se $\log(T) \sim N(\mu, \sigma^2)$, onde $\mu \in \mathbb{R}$ e $\sigma > 0$ são, respectivamente, a média e o desvio padrão da distribuição normal de $\log(T)$. As funções de sobrevivência e risco são expressas em termos da função de distribuição acumulada $\Phi(\cdot)$ e da função densidade $\phi(\cdot)$ da distribuição normal padrão:

$$h(t) = \frac{\phi\left(\frac{\log(t)-\mu}{\sigma}\right)}{\sigma t\left[1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)\right]}, \qquad S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right), \qquad t > 0.$$

Uma característica distintiva da distribuição log-normal é o comportamento não monótono de sua função de risco: $h(t)$ parte de zero quando $t \to 0^+$, cresce até atingir um valor máximo único em algum instante $t^* > 0$, e então decresce gradualmente, tendendo a zero quando $t \to \infty$. Este padrão, conhecido na literatura de confiabilidade como formato de "sino invertido" ou *upside-down bathtub* (Sweet, 1990), contrasta com os riscos monótonos das distribuições exponencial e Weibull.

Embora $h(t) \to 0$ assintoticamente, a velocidade de decaimento é suficientemente lenta para assegurar que a distribuição log-normal constitui uma distribuição de sobrevivência própria, isto é, que $S(t) \to 0$ conforme t fica arbitrariamente grande, garantindo que todo indivíduo eventualmente experimenta o evento. Para compreender esta propriedade aparentemente paradoxal, é necessário examinar o comportamento assintótico da função de risco acumulada

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

Uma distribuição de sobrevivência própria requer que $S(t) \to 0$, ou equivalentemente, que $H(t) \to \infty$, conforme t fica arbitrariamente grande. A questão central é: como pode uma função que tende a zero ter integral divergente?

A resposta reside na *velocidade* do decaimento. Definindo $z = (\log t - \mu)/\sigma$ e $\Psi(z) = 1 - \Phi(z)$, a função de risco pode ser reescrita como

$$h(t) = \frac{1}{\sigma t} \times \frac{\phi(z)}{\Psi(z)}.$$

Conforme t fica arbitrariamente grande, tem-se $z \to \infty$, e aplica-se a aproximação assintótica da razão de Mills para a cauda superior da distribuição normal padrão (Feller, 1968): $\Psi(z) \to \phi(z)/z$ conforme z fica arbitrariamente grande. Substituindo esta aproximação, obtém-se o comportamento assintótico

$$h(t) \sim \frac{1}{\sigma t} \times \frac{\phi(z)}{\phi(z)/z} = \frac{z}{\sigma t} = \frac{\log t - \mu}{\sigma^2 t}, \quad \text{conforme} \quad t \quad \text{fica arbitrariamente grande.}$$

O decaimento de $h(t)$ é, portanto, de ordem $\mathcal{O}\left(\frac{\log t}{t}\right)$, que é mais lento que qualquer decaimento polinomial da forma $1/t^{1+\epsilon}$ com $\epsilon > 0$.

Para verificar que a integral diverge, integra-se a aproximação assintótica, utilizando o método da substituição:

$$\int_1^t \frac{\log u}{u}\, du = \left[\frac{(\log u)^2}{2}\right]_1^t = \frac{(\log t)^2}{2} \to \infty \quad \text{conforme} \quad t \quad \text{fica arbitrariamente grande.}$$

Consequentemente, o risco acumulado apresenta o comportamento assintótico

$$H(t) \sim \frac{(\log t)^2}{2\sigma^2} \to \infty,$$

assegurando que $S(t) = \exp\{-H(t)\} \to 0$.

Este resultado ilustra um princípio fundamental da análise assintótica: a integrabilidade de uma função não depende apenas de seu limite, mas da velocidade com que este limite é atingido. Para contrastar, considere a função $\exp(-t)$, que também tende a zero quando $t \to \infty$, porém sua integral converge:

$$\int_0^\infty \exp(-t)\, dt = 1 < \infty.$$

Seu decaimento exponencial é "rápido demais" para produzir uma distribuição de sobrevivência própria. Em contraste, o decaimento logarítmico-polinomial $\mathcal{O}((\log t)/t)$ da log-normal situa-se precisamente no limiar crítico: lento o suficiente para garantir que $H(t) \to \infty$ conforme t fica arbitrariamente grande, mas ainda assim convergindo a zero assintoticamente.

Esse padrão de risco não monótono é particularmente adequado para modelar fenômenos em que a taxa de falha aumenta em uma fase inicial, atinge um pico e, posteriormente, diminui ao longo do tempo. Exemplos incluem processos de desgaste com manutenção corretiva, mortalidade infantil seguida de estabilização na idade adulta, e eventos biomédicos caracterizados por períodos críticos iniciais de adaptação (*burn-in*) seguidos de taxas decrescentes de falha.

A escolha entre esses modelos paramétricos deve ser guiada pela natureza do fenômeno sob estudo, pela disponibilidade de informação prévia sobre a forma esperada de $h(t)$ e por testes de adequação de ajuste aos dados observados. Quando não há justificativa teórica suficiente para impor uma forma paramétrica específica, abordagens semiparamétricas ou não paramétricas tornam-se mais apropriadas. O modelo de riscos proporcionais de Cox (que será introduzido na Seção 2.3.2) oferece flexibilidade ao não especificar a forma da função de risco basal, enquanto os estimadores de Kaplan-Meier e Nelson-Aalen (discutidos na Seção 2.2) fornecem estimativas não paramétricas das funções de sobrevivência e risco acumulado, respectivamente.

2.3.2 Modelo Semiparamétrico de Cox

O modelo de riscos proporcionais de Cox (Cox, 1972) representa uma das contribuições mais influentes à análise de sobrevivência, combinando flexibilidade na especificação da função de risco de base com a capacidade de modelar os efeitos de covariáveis sobre o tempo até o evento. No modelo de Cox, o risco de um indivíduo i no tempo t, quando

temos acesso a p covariáveis, é expresso como:

$$h_i(t \mid \mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

onde $h_0(t)$ é a função de risco de base (*baseline hazard*), não especificada parametricamente, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ é o vetor de covariáveis do indivíduo i, e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ é o vetor de coeficientes de regressão a ser estimado. A ausência de forma funcional para $h_0(t)$ confere ao modelo caráter semiparamétrico, concentrando a inferência nos efeitos relativos das covariáveis sem necessidade de especificar a forma temporal do risco.

O pressuposto fundamental do modelo de Cox é a proporcionalidade dos riscos (*proportional hazards assumption*), que estabelece que a razão de riscos entre dois indivíduos com diferentes valores de covariáveis permanece constante ao longo do tempo. Formalmente, para quaisquer dois vetores de covariáveis $\mathbf{x}_1$ e $\mathbf{x}_2$, a razão de riscos (*hazard ratio*) é:

$$\frac{h(t \mid \mathbf{x}_1)}{h(t \mid \mathbf{x}_2)} = \frac{h_0(t) \exp(\mathbf{x}_1^\top \boldsymbol{\beta})}{h_0(t) \exp(\mathbf{x}_2^\top \boldsymbol{\beta})} = \exp[(\mathbf{x}_1 - \mathbf{x}_2)^\top \boldsymbol{\beta}],$$

que é independente de t. Essa propriedade implica que o efeito multiplicativo de uma covariável sobre o risco não varia com o tempo, facilitando a interpretação. Por exemplo, se x_j é uma variável numérica, $\exp(\beta_j)$ representa o aumento proporcional no risco associado ao aumento de uma unidade na covariável x_j, mantendo as demais constantes.

A estimação de $\boldsymbol{\beta}$ é realizada por meio da verossimilhança parcial de Cox (Cox, 1972), que elimina a necessidade de especificar $h_0(t)$, tornando o modelo extremamente flexível e amplamente aplicável. A verossimilhança parcial baseia-se na ordenação relativa dos tempos de evento, condicionando nos conjuntos de risco em cada tempo de falha observado, e possui propriedades assintóticas similares às da verossimilhança completa. Devido à sua elegância matemática, robustez prática e facilidade de interpretação, o modelo de Cox dominou a literatura de análise de sobrevivência nas últimas décadas, sendo o método de referência em inúmeras aplicações biomédicas, epidemiológicas e industriais.

2.4 Verossimilhança e Estimação sob Censura

Com n observações independentes $(t_i, \delta_i, \mathbf{x}_i)$, onde $\mathbf{x}_i$ é o vetor de covariáveis associadas ao indivíduo i, a função de verossimilhança geral para dados de sobrevivência é

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_i \mid \mathbf{x}_i, \boldsymbol{\theta})]^{\delta_i} [S(t_i \mid \mathbf{x}_i, \boldsymbol{\theta})]^{1-\delta_i},$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros do modelo. Essa formulação combina, em um único produto, a contribuição dos indivíduos que sofreram o evento ($\delta_i = 1$), cuja contribuição é dada pela densidade $f(t_i)$, e daqueles censurados ($\delta_i = 0$), cuja contribuição é dada pela probabilidade de sobrevivência $S(t_i)$. Essa estrutura assegura consistência estatística na estimação de $\boldsymbol{\theta}$, pois utiliza toda a informação disponível, tanto de eventos observados quanto de observações censuradas.

2.5 Limitações dos Modelos Clássicos e Motivação para Abordagens Flexíveis

Embora amplamente utilizado, o modelo de Cox e os modelos paramétricos tradicionais enfrentam limitações em contextos de grande heterogeneidade e alta dimensionalidade, características comuns em bases de dados administrativas de saúde pública.

(i) **Linearidade dos efeitos**: O modelo de Cox assume que o efeito das covariáveis sobre o log-risco é linear, o que pode ser inadequado quando existem relações não lineares ou interações complexas entre variáveis.

(ii) **Proporcionalidade de riscos**: A suposição de riscos proporcionais pode ser violada quando os efeitos das covariáveis variam ao longo do tempo. Extensões como modelos de Cox estratificados ou com termos de interação tempo-covariável existem, mas aumentam a complexidade e podem ser insuficientes em situações de não proporcionalidade severa.

(iii) **Tempo discreto**: Bases administrativas, como as do Sistema Único de Saúde (SUS), frequentemente registram tempos em unidades discretas (anos completos, meses), contexto no qual modelos contínuos podem não ser ideais. Modelos de tempo discreto, embora menos explorados na literatura clássica, são mais apropriados para essa estrutura de dados.

(iv) **Alta dimensionalidade**: Em bases de dados com grande número de covariáveis e possíveis interações entre elas, abordagens paramétricas e semiparamétricas tradicionais podem enfrentar dificuldades de estimação, problemas de multicolinearidade e *overfitting*, além de limitações computacionais significativas.

Essas limitações motivam o uso de abordagens não paramétricas mais flexíveis e de técnicas de aprendizado de máquina, que serão apresentadas no Capítulo 3. Tais métodos oferecem maior capacidade para representar $h(t)$ e $S(t)$ de forma adaptativa, capturando padrões complexos sem impor restrições estruturais rígidas.

O modelo *Nnet-Survival*, que será detalhado no Capítulo 3, foi concebido precisamente para contornar essas limitações. Ao estimar diretamente o risco condicional por intervalo de tempo através de uma rede neural, o modelo dispensa a suposição de linearidade dos efeitos, permitindo que relações não lineares e interações de alta ordem entre covariáveis sejam capturadas automaticamente durante o processo de treinamento. A formulação em tempo discreto é intrinsecamente compatível com a granularidade dos registros administrativos do SUS, evitando aproximações contínuas artificiais que introduziriam viés de discretização. Por não depender da proporcionalidade de riscos, a arquitetura neural pode acomodar efeitos que variam ao longo do tempo, característica identificada na análise exploratória do Capítulo 4 para variáveis como raça/cor, escolaridade e faixa etária, cujas curvas de Kaplan-Meier estratificadas apresentaram cruzamentos e padrões não paralelos. Finalmente, as técnicas de regularização próprias do aprendizado profundo, especificamente *dropout* e penalização L2, controlam a complexidade efetiva da rede neural, que, por sua arquitetura com múltiplas camadas e unidades ocultas, possui um número de parâmetros internos muito superior ao de covariáveis originais, prevenindo sobreajuste mesmo em um modelo com alta capacidade expressiva, conforme será demonstrado empiricamente no Capítulo 5.

Este capítulo apresentou os fundamentos teóricos da análise de sobrevivência, desde os conceitos básicos e funções fundamentais até os estimadores não paramétricos clássicos (Kaplan-Meier e Nelson-Aalen) e os modelos de regressão paramétricos e semiparamétricos. Esses métodos constituem a base sobre a qual se desenvolvem abordagens mais flexíveis baseadas em aprendizado de máquina. No Capítulo 3, serão introduzidos os conceitos de aprendizado de máquina e redes neurais artificiais aplicados à análise de sobrevivência, culminando na apresentação do modelo *Nnet-Survival* que será implementado neste estudo.

Capítulo 3

Aprendizado de Máquina Aplicado à Análise de Sobrevivência

O avanço das técnicas de aprendizado de máquina nas últimas décadas revolucionou a modelagem estatística em diversas áreas, incluindo a análise de sobrevivência. Enquanto o Capítulo 2 apresentou os fundamentos clássicos da análise de sobrevivência, este capítulo introduz os conceitos de aprendizado de máquina e, em particular, de redes neurais artificiais, explorando como essas ferramentas podem ser adaptadas para modelar dados de sobrevivência. A integração entre a estrutura probabilística formal da análise de sobrevivência e a flexibilidade computacional das redes neurais constitui o alicerce metodológico deste estudo.

O aprendizado de máquina (*machine learning*) abrange algoritmos capazes de ajustar funções complexas a partir dos dados, aprendendo padrões sem a necessidade de especificação explícita de regras. Formalmente, busca-se uma função $f : \mathbb{R}^p \to \mathbb{R}$ tal que $\hat{y} = f(\mathbf{x})$ aproxime uma relação desconhecida entre entradas $\mathbf{x} = (x_1, \ldots, x_p)$ e uma saída y. O processo de ajuste consiste em minimizar uma função de perda $\mathcal{L}(y, \hat{y})$, que mede a discrepância entre valores previstos e observados (Hastie; Tibshirani; Friedman, 2009; Goodfellow; Bengio; Courville, 2016).

Para dados de sobrevivência, a função de perda precisa incorporar adequadamente a censura. Funções baseadas na verossimilhança de tempo contínuo ou discreto, conforme apresentado na Seção 2.4, asseguram que tanto eventos quanto censuras contribuam de forma apropriada para a estimação dos parâmetros do modelo.

3.1 Redes Neurais Artificiais: Estrutura e Funcionamento

Uma rede neural artificial (RNA) é um modelo computacional inspirado na estrutura do cérebro humano, composto por camadas de neurônios artificiais interconectados. Cada neurônio aplica uma transformação não linear sobre uma combinação linear das entradas. Seja $\mathbf{x} \in \mathbb{R}^p$ o vetor de entrada (covariáveis), então a ativação a (saída) de um neurônio é dada por

$$a = \varphi(\mathbf{w}^\top \mathbf{x} + b),$$

em que $\mathbf{w}$ são os pesos associados a cada covariável (refletindo o grau de importância de cada variável para a predição), b é o viés (*bias*), que permite deslocar a função de ativação, e $\varphi(\cdot)$ é a função de ativação, responsável por introduzir não-linearidade no modelo.

3.1.1 Funções de Ativação

As funções de ativação mais comumente utilizadas incluem:

- **ReLU** (*Rectified Linear Unit*): $\varphi(z) = \max(0, z)$, que retorna zero para valores negativos e o próprio valor para entradas não-negativas. É computacionalmente eficiente e amplamente utilizada em redes profundas por mitigar o problema de desvanecimento do gradiente, ou seja, a redução excessiva da magnitude dos gradientes durante a retropropagação em redes com muitas camadas, o que dificulta o aprendizado das camadas iniciais;
- **Sigmoide**: $\varphi(z) = \frac{1}{1+e^{-z}}$, que mapeia qualquer valor real para o intervalo $(0, 1)$, sendo útil para representar probabilidades;
- **Tangente hiperbólica**: $\varphi(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$, que mapeia valores reais para o intervalo $(-1, 1)$, centralizando a saída em torno de zero.

3.1.2 Arquitetura em Camadas

Em notação matricial, para uma rede com múltiplas camadas, a ativação da camada ℓ é dada por:

$$a^{(\ell)} = \varphi(W^{(\ell)} a^{(\ell-1)} + b^{(\ell)}),$$

onde $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ é a matriz de pesos que conecta a camada $\ell - 1$ (com $n_{\ell-1}$ neurônios) à camada ℓ (com n_ℓ neurônios), $b^{(\ell)} \in \mathbb{R}^{n_\ell}$ é o vetor de vieses da camada ℓ, e $a^{(\ell)} \in \mathbb{R}^{n_\ell}$ representa o vetor de ativações (saídas) dos neurônios na camada ℓ, com $a^{(0)} = \mathbf{x}_i \in \mathbb{R}^p$, o vetor de entrada (covariáveis do indivíduo i). O conjunto de parâmetros da rede é $\Theta = \{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L}$, onde L é o número total de camadas ocultas. As camadas ocultas (*hidden layers*), correspondentes a $\ell \in \{1, 2, \ldots, L\}$, são as camadas intermediárias situadas entre a camada de entrada $a^{(0)}$ e a camada de saída final. Diferentemente das camadas de entrada e saída, que são diretamente observáveis (covariáveis e predições, respectivamente), as camadas ocultas realizam transformações internas dos dados, extraindo representações hierárquicas e progressivamente mais abstratas que capturam padrões não lineares e interações complexas entre as covariáveis.

A arquitetura de uma rede neural *feedforward* (alimentação adiante), na qual a informação flui exclusivamente da camada de entrada em direção à camada de saída, sem conexões recorrentes, com L camadas ocultas é ilustrada na Figura 3.1.

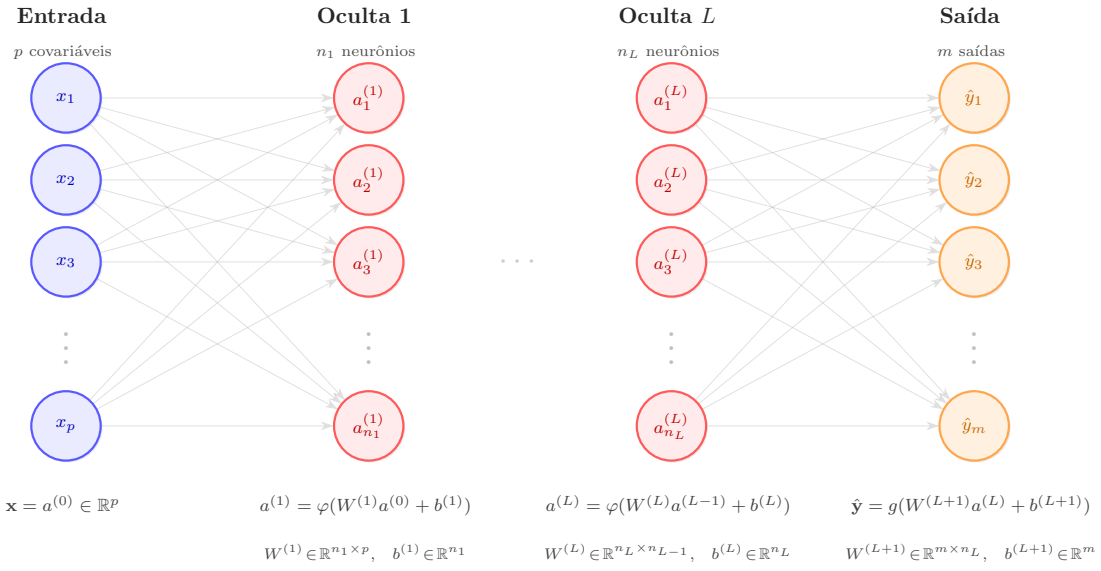

Figura 3.1 – Representação esquemática de uma rede neural *feedforward* com L camadas ocultas. A camada de entrada recebe o vetor $\mathbf{x} \in \mathbb{R}^p$, cada camada oculta ℓ aplica a transformação $a^{(\ell)} = \varphi(W^{(\ell)} a^{(\ell-1)} + b^{(\ell)})$ com $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, e a camada de saída produz $\hat{\mathbf{y}} \in \mathbb{R}^m$ mediante uma função de ativação $g(\cdot)$. As reticências horizontais indicam camadas ocultas intermediárias omitidas.

Fonte: Elaboração própria.

A arquitetura da rede, incluindo a profundidade L e o número de neurônios em cada camada, constitui um conjunto de hiperparâmetros que devem ser especificados a priori e não são estimados durante o treinamento. Sua escolha envolve um compromisso (*trade-off*) entre capacidade representacional e risco de sobreajuste: redes mais profundas e largas podem, em princípio, aproximar funções mais complexas devido ao aumento da capacidade do modelo, mas também apresentam maior propensão ao sobreajuste e maior custo computacional. Na prática, a arquitetura pode ser determinada por validação cruzada ou mediante a divisão do conjunto de dados em subconjuntos de treino, validação e teste, embora a busca conjunta pela profundidade e pelo número de neurônios em cada camada torne o espaço de configurações candidatas muito grande, elevando substancialmente o custo computacional. Por essa razão, é comum que parte da escolha arquitetural se baseie em heurísticas, experiência prévia e práticas consolidadas na literatura, restringindo a busca sistemática a um subconjunto reduzido de configurações plausíveis (Bergstra; Bengio, 2012). A arquitetura específica adotada neste estudo é detalhada na Subseção 5.2.2.

3.1.3 Treinamento via Retropropagação

O treinamento visa minimizar uma função de perda global $\mathcal{L}(\Theta)$, tipicamente pela retropropagação do erro (*backpropagation*) combinada com algoritmos de otimização baseados em gradiente. O gradiente $\nabla_\Theta \mathcal{L}$ é calculado iterativamente, atualizando os pesos e vieses até apresentar indícios fortes de convergência. A retropropagação é, essencialmente, uma forma eficiente de aplicar a regra da cadeia do cálculo diferencial: em vez de calcular independentemente a derivada de $\mathcal{L}$ em relação a cada parâmetro da rede, o algoritmo propaga os erros da camada de saída em direção à camada de entrada, reaproveitando resultados intermediários já computados nas camadas posteriores para obter os gradientes das camadas anteriores, evitando cálculos redundantes e tornando viável o treinamento de redes com milhares de parâmetros. O algoritmo de otimização e os hiperparâmetros de treinamento adotados neste estudo são descritos na Subseção 5.2.1.

3.2 Modelos de Sobrevivência Baseados em Redes Neurais

A integração entre redes neurais e análise de sobrevivência tem sido objeto de intensa pesquisa nas últimas décadas, resultando em diversas arquiteturas que adaptam a flexibilidade das redes neurais às especificidades dos dados de sobrevivência.

3.2.1 DeepSurv

O modelo *DeepSurv* (Katzman *et al.*, 2018) substitui o preditor linear do modelo de Cox por uma função não linear $\psi(\mathbf{x}; \Theta)$ parametrizada por uma rede neural, mantendo a estrutura da verossimilhança parcial de Cox descrita na Seção 2.3.2. Assim, o risco individual torna-se:

$$h_i(t) = h_0(t) \exp[\psi(\mathbf{x}_i; \Theta)].$$

Essa abordagem preserva a interpretabilidade relativa do modelo de Cox (razões de risco entre indivíduos) enquanto permite capturar relações não lineares complexas entre as covariáveis e o log-risco.

3.2.2 Cox-Time

O modelo *Cox-Time* (Kvamme; Borgan; Scheel, 2019) vai além e relaxa a suposição de proporcionalidade de riscos descrita na Seção 2.3.2, permitindo que o efeito das covariáveis varie ao longo do tempo. Isso é alcançado ao condicionar a função $\psi(\mathbf{x}; \Theta)$ também ao tempo t, resultando em $\psi(\mathbf{x}, t; \Theta)$, de modo que a função de representação possa variar temporalmente e capturar padrões de risco não proporcionais ao longo do seguimento, tornando o modelo mais flexível e adequado para situações em que a suposição de riscos proporcionais não se sustenta.

3.2.3 Nnet-Survival: Modelagem em Tempo Discreto

O modelo *Nnet-Survival* (Gensheimer; Narasimhan, 2019), que será adotado neste estudo, distingue-se por adotar explicitamente uma formulação em tempo discreto. Essa escolha é particularmente adequada para bases de dados administrativas nas quais os tempos são registrados em unidades discretas (como anos completos), evitando a imposição de uma estrutura contínua inadequada aos dados observados.

O período de acompanhamento é dividido em J intervalos de igual tamanho $[t_0, t_1)$, $[t_1, t_2), \ldots, [t_{J-1}, t_J)$, tal que $t_{k+1} - t_k = t_k - t_{k-1}$ para todo $k = 1, \ldots, J-1$. Denota-se por t_k o extremo direito do k-ésimo intervalo, de modo que t_0 é o início do acompanhamento e t_J é o fim do último intervalo. Para cada indivíduo i, define-se k_i como o índice do intervalo em que o indivíduo experimentou o evento ($\delta_i = 1$) ou foi censurado ($\delta_i = 0$), ou seja, k_i é tal que o tempo observado t_i pertence ao intervalo $[t_{k_i-1}, t_{k_i})$. A rede estima as probabilidades condicionais de falha em cada intervalo:

$$h_{ik} = P(T_i \in [t_{k-1}, t_k) \mid T_i \geq t_{k-1}, \mathbf{x}_i), \quad k = 1, \ldots, J,$$

e a função de sobrevivência discreta associada ao k-ésimo intervalo é reconstruída por:

$$S_i(t_k \mid \mathbf{x}_i) = \prod_{j=1}^{k} (1 - h_{ij}),$$

representando a probabilidade de que o indivíduo i sobreviva além do tempo t_k.

A contribuição do indivíduo i para a verossimilhança pode ser expressa como:

- Se $\delta_i = 1$ (evento observado no intervalo k_i): a contribuição é a probabilidade de sobreviver até t_{k_i-1} e falhar em $[t_{k_i-1}, t_{k_i})$:

$$L_i = h_{ik_i} \prod_{j=1}^{k_i-1} (1 - h_{ij}).$$

- Se $\delta_i = 0$ (censurado em k_i): a contribuição é a probabilidade de sobreviver até pelo menos t_{k_i}:

$$L_i = \prod_{j=1}^{k_i} (1 - h_{ij}).$$

Tomando o logaritmo e combinando ambos os casos, a log-verossimilhança para o indivíduo i é:

$$\log L_i = \delta_i \log h_{ik_i} + \sum_{j=1}^{k_i-\delta_i} \log(1 - h_{ij}).$$

Note que quando $\delta_i = 1$, o somatório vai até $k_i - 1$, e quando $\delta_i = 0$, vai até k_i. Então, considerando uma amostra de n observações independentes, a log-verossimilhança negativa total a ser minimizada é:

$$\mathcal{L}(\Theta) = -\sum_{i=1}^{n} \left[\delta_i \log h_{ik_i} + \sum_{j=1}^{k_i-\delta_i} \log(1 - h_{ij}) \right], \tag{3.1}$$

que incorpora naturalmente a censura ($\delta_i = 0$) e os eventos observados ($\delta_i = 1$).

Essa formulação apresenta diversas vantagens: (i) evita integrais, simplificando a implementação computacional; (ii) oferece boa estabilidade numérica durante o treinamento; (iii) é eficiente para grandes bases de dados com alta taxa de censura; e (iv) é conceitualmente alinhada com a natureza discreta dos registros administrativos de saúde. A implementação específica dessa função de perda, incluindo as adaptações numéricas adotadas, é detalhada na Subseção 5.2.1.

3.3 Métricas de Avaliação para Modelos de Sobrevivência

A avaliação do desempenho de modelos de sobrevivência requer métricas específicas que considerem adequadamente a presença de censura e a natureza temporal das predições. Diferentemente de problemas de classificação ou regressão tradicionais, nos quais a métrica de acurácia ou o erro quadrático médio podem ser diretamente aplicados, a análise de sobrevivência demanda medidas que incorporem tanto a ordenação dos tempos de evento quanto a calibração probabilística das estimativas. Nesta seção, apresentam-se as principais métricas utilizadas neste estudo: o índice de concordância (discriminação), o escore de Brier integrado (discriminação e calibração conjuntas) e a importância por permutação de variáveis (Harrell *et al.*, 1982; Graf *et al.*, 1999; Steyerberg *et al.*, 2010).

3.3.1 Índice de Concordância (C-index)

O índice de concordância, comumente denominado *C-index* ou estatística C de Harrell, é uma generalização da área sob a curva ROC (*AUC-ROC*) para dados de sobrevivência com censura (Harrell *et al.*, 1982; Pencina; D'Agostino; Steyerberg, 2011). A conexão é direta: em um problema de classificação binária, a AUC-ROC corresponde à probabilidade de que um indivíduo positivo receba um escore de risco maior que um indivíduo negativo; analogamente, o *C-index* corresponde à probabilidade de que, dado um par de indivíduos comparáveis, aquele que experimentou o evento primeiro tenha recebido maior escore de risco pelo modelo. A diferença fundamental é que o *C-index* lida com a censura, restringindo a avaliação aos pares para os quais a ordenação temporal é observável. Essa métrica quantifica a capacidade discriminatória do modelo, isto é, sua habilidade de ordenar corretamente os indivíduos conforme seus riscos relativos.

Considere dois indivíduos i e j com tempos observados t_i e t_j e indicadores de evento δ_i e δ_j. O par (i, j) é comparável se pudermos determinar inequivocamente qual indivíduo experimentou o evento primeiro, o que ocorre quando $t_i < t_j$ e $\delta_i = 1$, ou vice-versa. O par é concordante se o modelo atribui maior risco ao indivíduo que falhou primeiro, ou seja, se $\hat{r}_i > \hat{r}_j$ quando $t_i < t_j$ e $\delta_i = 1$.

O *C-index* é então calculado como:

$$C = \frac{\sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(t_i < t_j) \times \delta_i \times \mathbb{I}(\hat{r}_i > \hat{r}_j)}{\sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(t_i < t_j) \times \delta_i},$$

onde $\mathbb{I}(\cdot)$ denota a função indicadora, $\hat{r}_i$ é o escore de risco predito para o indivíduo i, e o denominador representa o número total de pares comparáveis. Um valor de $C = 0{,}5$ indica capacidade discriminatória equivalente ao acaso, enquanto $C = 1{,}0$ representa discriminação perfeita. Valores no intervalo de 0,7 a 0,8 são considerados bons em aplicações clínicas, e valores acima de 0,8 são considerados excelentes (Pencina; D'Agostino; Steyerberg, 2011).

O *C-index* é particularmente útil por ser não paramétrico, não assumir proporcionalidade de riscos e ser robusto a transformações monotônicas crescentes dos escores de risco. Entretanto, essa métrica avalia exclusivamente a capacidade de ordenação, sem verificar se as probabilidades de sobrevivência preditas estão bem calibradas (Uno *et al.*, 2011). Um modelo pode ordenar corretamente os indivíduos mas produzir estimativas de probabilidade sistematicamente enviesadas, limitando sua utilidade para decisões que dependem de probabilidades absolutas (Steyerberg *et al.*, 2010). Por essa razão, o *C-index* deve ser complementado por métricas de calibração, como o escore de Brier integrado.

3.3.2 Escore de Brier Integrado (Integrated Brier Score)

Enquanto o *C-index* avalia discriminação, o escore de Brier mede a acurácia probabilística das predições de sobrevivência, considerando simultaneamente discriminação e calibração (Graf *et al.*, 1999; Gerds; Schumacher, 2006). Dado o modelo em tempo discreto com J intervalos, conforme definido na Subseção 3.2.3, o escore de Brier no tempo t_k quantifica o erro quadrático médio entre as probabilidades de sobrevivência preditas e o status observado. Utilizando a notação estabelecida na Subseção 3.2.3, onde t_k denota o extremo direito do k-ésimo intervalo e k_i é o índice do intervalo associado ao indivíduo i, o escore de Brier é definido como:

$$\mathrm{BS}(t_k) = \frac{1}{n} \sum_{i=1}^{n} w_i(k) \left[\hat{S}_i(t_k) - \mathbb{I}(t_{k_i} > t_k) \right]^2, \tag{3.2}$$

onde:

- $\hat{S}_i(t_k) = \prod_{j=1}^{k}(1 - h_{ij})$ é a probabilidade de sobrevivência predita pelo modelo para o indivíduo i até o final do k-ésimo intervalo;
- $\mathbb{I}(t_{k_i} > t_k)$ é o indicador de que o indivíduo i sobreviveu além do k-ésimo intervalo;
- $w_i(k)$ são pesos de ponderação pela censura inversa da probabilidade (*inverse probability of censoring weighting*, IPCW), necessários para corrigir o viés introduzido pela censura (Graf *et al.*, 1999).

Pesos IPCW

Os pesos IPCW são calculados como:

$$w_i(k) = \begin{cases} \frac{1}{\hat{G}(t_{k_i})}, & \text{se } k_i \leq k \text{ e } \delta_i = 1, \\ \frac{1}{\hat{G}(t_k)}, & \text{se } k_i > k, \\ 0, & \text{se } k_i \leq k \text{ e } \delta_i = 0, \end{cases} \tag{3.3}$$

onde $\hat{G}(t_m)$ é a estimativa da probabilidade de não ocorrer censura até o tempo t_m, obtida a partir do estimador de Kaplan-Meier. Formalmente, $\hat{G}(t_m) = P(T_c > t_m)$, em que T_c denota o tempo de censura. Os pesos IPCW podem ser interpretados da seguinte forma:

- **Eventos observados** ($k_i \leq k$, $\delta_i = 1$): O peso $1/\hat{G}(t_{k_i})$ compensa a perda de indivíduos similares censurados antes do intervalo k_i. Quanto menor $\hat{G}(t_{k_i})$, maior o peso atribuído, compensando a maior escassez de informação sobre as falhas ocorridas no tempo t_{k_i}.
- **Indivíduos sob risco** ($k_i > k$): O peso $1/\hat{G}(t_k)$ ajusta pela probabilidade de não ser censurado até o final do k-ésimo intervalo, garantindo que indivíduos ainda em acompanhamento representem adequadamente a população original.
- **Censurados antes ou durante o k-ésimo intervalo** ($k_i \leq k$, $\delta_i = 0$): Recebem peso zero e não contribuem para $\text{BS}(t_k)$, pois não há informação sobre seu status verdadeiro no tempo t_k, já que ele está além do tempo de censura. Este terceiro caso garante que apenas observações com informação verificável influenciem a avaliação da qualidade do ajuste.

Para ilustrar, considere uma coorte hipotética com acompanhamento discretizado em intervalos anuais. Suponha que o indivíduo i experimentou o evento no 5º intervalo ($k_i = 5$, correspondendo a $t_5 = 5$ anos de seguimento) e que a probabilidade estimada de não ser censurado até esse tempo é $\hat{G}(t_5) = 0{,}80$, ou seja, 80% dos indivíduos da coorte ainda não haviam sido censurados até o 5º ano. Pelo primeiro caso da Equação (3.3), esse indivíduo recebe peso $w_i = 1/0{,}80 = 1{,}25$. A interpretação é que, para cada indivíduo que efetivamente experimentou o evento em t_5, havia outros indivíduos com perfil semelhante que foram censurados antes de t_5 e, portanto, não puderam contribuir para a avaliação nesse tempo. O peso de 1,25 compensa essa perda informacional, atribuindo 25% a mais de importância à observação do indivíduo i. Caso a censura fosse mais intensa, por exemplo $\hat{G}(t_5) = 0{,}50$ (metade dos indivíduos já censurados), o peso seria $w_i = 1/0{,}50 = 2{,}00$, duplicando a contribuição desse indivíduo para refletir a maior escassez de informação disponível nesse horizonte.

Truncamento de Pesos IPCW

Em cenários de alta censura, os pesos IPCW podem tornar-se instáveis em horizontes temporais longos, quando $\hat{G}(t_k)$ se aproxima de zero. Para prevenir instabilidade numérica, Kvamme, Borgan e Scheel (2019) recomendam o truncamento dos pesos em um limite superior $w_{\max}$. Formalmente, o peso truncado é:

$$w_i^{\text{trunc}}(k) = \min\left(w_i(k), w_{\max}\right).$$

A escolha do valor de $w_{\max}$ adotado neste estudo é justificada na Subseção 5.4.1.

Horizonte Confiável de Avaliação

Graf *et al.* (1999) argumentam que a avaliação do escore de Brier deve ser restrita a horizontes temporais onde a censura não seja excessivamente pesada, recomendando o uso da mediana do tempo de seguimento como critério de referência. Um critério operacional comum é adotar o limiar $\hat{G}(t_k) > 0{,}10$, garantindo que pelo menos 10% dos indivíduos permaneçam livres de censura até o tempo t_k. Resultados para horizontes além desse limiar devem ser interpretados com cautela devido à instabilidade crescente dos pesos IPCW. A operacionalização desse critério no presente estudo é detalhada na Subseção 5.4.1.

Escore de Brier Integrado

Para avaliar o desempenho preditivo ao longo de um intervalo temporal específico, utiliza-se o escore de Brier integrado (*integrated Brier score*, IBS), definido como (Gerds; Schumacher, 2006):

$$\mathrm{IBS}(t_a, t_b) = \frac{1}{t_b - t_a} \int_{t_a}^{t_b} \mathrm{BS}(t)\, dt \approx \frac{1}{n_{\mathrm{pts}}} \sum_{t=t_a}^{t_b} \mathrm{BS}(t), \tag{3.4}$$

onde t_a e t_b definem o intervalo temporal de interesse, $n_{\mathrm{pts}} = t_b - t_a$ é o número de pontos temporais discretos nesse intervalo, e a integral contínua é aproximada por uma soma discreta.

Quando se deseja avaliar o desempenho ao longo de todo o horizonte confiável, define-se $t_a = t_0$ (início do acompanhamento) e $t_b = t_K$ (fim do horizonte confiável), obtendo:

$$\mathrm{IBS}(t_0, t_K) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{BS}(t_k).$$

O IBS fornece uma medida global da acurácia preditiva no intervalo especificado: valores menores indicam melhor desempenho. A flexibilidade dessa formulação permite avaliar o modelo em diferentes horizontes temporais, como curto, médio e longo prazo, mediante a escolha adequada dos limites t_a e t_b.

3.3.3 Avaliação de Calibração via Escore de Brier Integrado

Calibração refere-se à concordância entre as probabilidades preditas pelo modelo e as frequências observadas dos eventos (Houwelingen, 2000; Steyerberg *et al.*, 2010). Um modelo bem calibrado produz predições que, quando agregadas, refletem com precisão as taxas empíricas de sobrevivência. Por exemplo, se o modelo prediz que 30% dos indivíduos sobreviverão além de 5 anos, então aproximadamente 30% devem de fato sobreviver além dos 5 anos.

Enquanto o *C-index* (Subseção 3.3.1) mede exclusivamente discriminação, a calibração avalia a acurácia das probabilidades absolutas preditas. Modelos com boa discriminação mas má calibração podem ordenar corretamente os riscos, porém com magnitudes incorretas, limitando sua utilidade clínica (Steyerberg *et al.*, 2010). Portanto, discriminação e calibração são características complementares e essenciais.

Ao final da Subseção 3.3.2, o IBS foi apresentado como métrica que incorpora simultaneamente discriminação e calibração, penalizando tanto a má ordenação quanto o desvio

das probabilidades preditas (Graf *et al.*, 1999; Gerds; Schumacher, 2006). Valores baixos de IBS indicam bom desempenho em ambas as características. O valor de referência para predições aleatórias é aproximadamente 0,25 quando os pesos IPCW são uniformes (cenário sem censura ou com censura ignorada); na presença de censura e pesos IPCW não uniformes, esse valor de referência pode variar ligeiramente dependendo da estrutura de censura dos dados.

Recalibração Pós-Treinamento

Modelos de redes neurais, especialmente aqueles que utilizam *dropout*, podem apresentar descalibração mesmo com boa discriminação (Guo *et al.*, 2017). O *temperature scaling* é um método de recalibração que ajusta a confiança das predições mediante:

$$\hat{h}_k^{\text{recal}}(\mathbf{x}) = \sigma\left(\frac{z_k(\mathbf{x})}{\tau}\right),$$

onde $\sigma(\cdot)$ é a função sigmoide, $z_k(\mathbf{x})$ é o *logit* (saída pré-ativação da rede) para o intervalo k, e $\tau > 0$ é um parâmetro de calibração que é otimizado no conjunto de validação. Este método preserva a ordenação (mantendo o *C-index*) enquanto corrige a calibração das probabilidades absolutas. A decisão sobre a aplicação de recalibração neste estudo é justificada na Subseção 5.4.1.

3.3.4 Importância por Permutação de Variáveis

A importância por permutação (*permutation importance*) é uma técnica modelo-agnóstica[1] que quantifica a contribuição de cada variável preditora (Breiman, 2001; Fisher; Rudin; Dominici, 2019). Diferentemente de métodos específicos (coeficientes em regressão, pesos em redes neurais), aplica-se a qualquer modelo preditivo, incluindo algoritmos de *black-box*[2].

Divisão dos Dados

A avaliação não viesada requer dividir os dados de modo aleatório em três conjuntos (Hastie; Tibshirani; Friedman, 2009):

[1] Uma técnica modelo-agnóstica é aquela que pode ser aplicada a qualquer modelo preditivo, independentemente de sua estrutura interna ou algoritmo de treinamento.

[2] Um modelo *black-box* (caixa-preta) é aquele cuja estrutura interna é suficientemente complexa para que a relação entre entradas e saídas não seja diretamente interpretável pelo analista, como redes neurais profundas, *random forests* com muitas árvores ou métodos de *ensemble*.

- Conjunto de treino: Para ajustar parâmetros (60–80% da amostra).
- Conjunto de validação: Para monitoramento, seleção de hiperparâmetros e *early stopping*[3] (10–20% da amostra).
- Conjunto de teste: Para avaliação final, completamente isolado durante o desenvolvimento (10–20% da amostra).

Essa divisão previne sobreajuste e garante estimativas não viesadas de generalização (Goodfellow; Bengio; Courville, 2016). Em análise de sobrevivência, deve-se preservar proporções similares de eventos e censuras em cada subconjunto, o que pode ser realizado por amostragem estratificada: primeiro separam-se os dados entre censurados e não censurados, e então extraem-se aleatoriamente os mesmos percentuais de cada grupo para compor os conjuntos de treino, validação e teste (Hastie; Tibshirani; Friedman, 2009). As proporções específicas adotadas neste estudo são detalhadas na Seção 5.3.

Procedimento de Cálculo da Importância por Permutação

O procedimento de cálculo da importância por permutação é formalizado no Algoritmo 1. A permutação deve ser realizada exclusivamente no conjunto de teste para avaliar a contribuição à generalização do modelo já treinado.

Algoritmo 1: Importância por permutação de variáveis

Input: Modelo treinado f, conjunto de teste $\mathcal{D}_{\text{teste}}$, variáveis $\{x_1, \ldots, x_p\}$

Output: Vetor de importâncias $\mathbf{I} = (I_1, \ldots, I_p)$

1 Avaliar f em $\mathcal{D}_{\text{teste}}$ usando o *C-index*, obtendo o desempenho de referência M_0;
2 **para** $j = 1, \ldots, p$ **faça**
3 Permutar aleatoriamente os valores de x_j em $\mathcal{D}_{\text{teste}}$, quebrando sua associação com o desfecho;
4 Avaliar f no conjunto com x_j permutada, obtendo M_j;
5 $I_j \leftarrow M_0 - M_j$;
6 **fim**
7 **(Opcional)** Repetir os passos 1–5 com diferentes permutações para obter intervalos de confiança;
8 **retorna I**

[3] *Early stopping* é uma técnica de regularização que interrompe o treinamento quando o desempenho no conjunto de validação começa a deteriorar, prevenindo sobreajuste.

Interpretação e Propriedades

Uma importância I_j elevada indica que x_j é altamente informativa, indicando que desconsiderá-la no modelo causa prejuízo substancial. Importâncias próximas de zero sugerem baixa contribuição ou ruído (Strobl *et al.*, 2008).

Vantagens: (i) intuitiva; (ii) modelo-agnóstica; (iii) captura efeitos não lineares e interações; (iv) reflete contribuição real à generalização.

Limitações: Em presença de forte correlação entre variáveis, pode gerar combinações irreais e estimativas enviesadas (Hooker; Mentch; Zhou, 2021). Extensões condicionais mitigam esse problema (Strobl *et al.*, 2008).

No contexto de sobrevivência, a importância por permutação identifica características com maior influência sobre o tempo até o evento, orientando decisões clínicas e futuros estudos. Em redes neurais, onde pesos individuais são de difícil interpretação, fornece uma janela interpretável. Embora o procedimento descrito utilize o *C-index* como métrica de referência, outras métricas de desempenho podem ser empregadas dependendo do objetivo da análise.

Por fim, este capítulo apresentou os fundamentos de aprendizado de máquina, a estrutura e funcionamento de redes neurais artificiais, e sua aplicação à análise de sobrevivência por meio de modelos como *DeepSurv*, *Cox-Time* e, particularmente, o *Nnet-Survival*, que será implementado no estudo realizado neste trabalho. A integração entre a tradição estatística da análise de sobrevivência, apresentada no Capítulo 2, e as inovações do aprendizado profundo, ampliam as possibilidades analíticas e oferecem ferramentas poderosas para modelar fenômenos complexos em saúde pública. O Capítulo 5 aplicará esses conceitos aos dados reais da coorte nacional de pacientes em terapia antirretroviral.

Capítulo 4

Análise Exploratória e Qualidade dos Dados

4.1 Introdução

A análise exploratória de dados representa uma etapa fundamental em estudos quantitativos que envolvem métodos de Análise de Sobrevivência, sobretudo quando aplicados a grandes bases provenientes de registros administrativos na área da saúde. Trata-se de um processo anterior à modelagem estatística propriamente dita, cujo objetivo central é realizar uma avaliação minuciosa da consistência e da qualidade das informações disponíveis, bem como oferecer uma visão geral sobre a forma como os dados se distribuem e fornecer direcionamentos essenciais para a etapa de modelagem (Kleinbaum; Klein, 2020; Colosimo; Giolo, 2024).

No contexto deste trabalho, a análise exploratória assumiu papel decisivo na definição de uma coorte de estudo metodologicamente robusta, voltada à investigação do tempo até a interrupção da terapia antirretroviral (TARV) em pessoas vivendo com HIV/Aids no Brasil. Este recorte analítico ganha especial relevância no campo da saúde pública, dada a reconhecida influência de determinantes sociais da saúde, como escolaridade, raça/cor e localização geográfica, sobre os desfechos associados à adesão e continuidade do tratamento antirretroviral (Cunha *et al.*, 2025; Brasil. Ministério da Saúde, 2024e).

A integração dos diferentes sistemas que compõem o Painel Integrado de Monitoramento do Cuidado do HIV e da Aids (PIMC) possibilita a realização de estudos epidemiológicos em larga escala, oferecendo uma perspectiva detalhada e longitudinal dos cuidados prestados às pessoas vivendo com HIV/Aids no Brasil. Contudo, a qualidade dos indicadores gerados

depende diretamente das características das bases de origem, incluindo a frequência dos eventos registrados, o tamanho populacional das unidades geográficas analisadas e a completude das informações disponíveis. Por essa razão, recomenda-se cautela na interpretação dos resultados, sobretudo em análises voltadas a subgrupos com pequeno número de casos ou baixa cobertura informacional (Brasil. Ministério da Saúde, 2024e).

Diante da complexidade inerente ao PIMC, que reúne milhões de registros individuais ao longo de extensos períodos de acompanhamento, a realização de uma análise exploratória rigorosa mostrou-se indispensável. A heterogeneidade sociodemográfica e clínica da população analisada, aliada às múltiplas formas de vulnerabilidade que caracterizam o contexto brasileiro da epidemia de HIV/Aids, demandaram estratégias analíticas capazes de captar tais heterogeneidades e mitigar potenciais fontes de vieses (Melo *et al.*, 2019; Guerra; Seidl, 2010; Costa *et al.*, 2021).

4.2 Fonte de Dados

Os dados utilizados nesta pesquisa foram obtidos a partir do Painel Integrado de Monitoramento do Cuidado do HIV e da Aids (PIMC), uma iniciativa desenvolvida pelo Departamento de HIV, Aids, Tuberculose, Hepatites Virais e Infecções Sexualmente Transmissíveis (Dathi), vinculado à Secretaria de Vigilância em Saúde e Ambiente do Ministério da Saúde do Brasil. O PIMC tem como principal objetivo consolidar informações clínicas, laboratoriais, administrativas e epidemiológicas relativas às pessoas vivendo com HIV/Aids atendidas no Sistema Único de Saúde (SUS), permitindo o acompanhamento longitudinal desses indivíduos em âmbito nacional. Todos os dados utilizados neste estudo são de livre acesso e podem ser obtidos em https://www.gov.br/aids/pt-br/indicadores-epidemiologicos/painel-de-monitoramento/painel-integrado-de-monitoramento-do-cuidado-do-hiv.

Para a construção de seus indicadores, o painel integra registros oriundos de diferentes sistemas nacionais de informação em saúde, incluindo o Sistema de Informação de Agravos de Notificação (Sinan), responsável pelas notificações compulsórias de casos de HIV e aids; o Sistema de Controle Logístico de Medicamentos (Siclom), que documenta os processos de dispensação de antirretrovirais; o Sistema de Controle de Exames Laboratoriais (Siscel), que reúne resultados de exames de carga viral e de contagem de linfócitos T CD4+; e o Sistema de Informações sobre Mortalidade (SIM), que registra óbitos em território nacional. Adicionalmente, o painel incorpora informações populacionais provenientes

dos censos demográficos conduzidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE), bem como dados de outros sistemas de monitoramento mantidos pelo próprio Dathi (Brasil. Ministério da Saúde, 2024e).

Essa estrutura integrada de dados permite a realização de análises epidemiológicas de grande escala, com diferentes níveis de desagregação geográfica e temporal, abrangendo desde o nível municipal até o nacional, e fornecendo subsídios relevantes para a gestão e avaliação das políticas públicas de enfrentamento da epidemia de HIV/Aids no Brasil.

Neste estudo, a construção da base analítica teve como ponto de partida duas fontes principais derivadas do PIMC: as bases denominadas PRIM e ULT, ambas convertidas para o formato *Apache Parquet*, visando otimizar o desempenho nas etapas de processamento analítico (conforme detalhado na Seção 4.3).

A base PRIM contém uma linha por indivíduo e reúne informações sobre o primeiro evento relacionado ao cuidado com o HIV/Aids, incluindo a data de início da terapia antirretroviral (TARV) e o primeiro exame de contagem de linfócitos T CD4+. Já a base ULT oferece um acompanhamento longitudinal, organizando dados anuais para cada pessoa que apresentou ao menos um evento clínico no respectivo ano, contemplando o status de tratamento, resultados laboratoriais e demais informações pertinentes ao seguimento clínico.

A Tabela 4.1 apresenta a descrição detalhada das principais variáveis utilizadas no estudo, especificando sua origem (PRIM ou ULT) e o conteúdo de cada uma.

Tabela 4.1 – Descrição das principais variáveis utilizadas no estudo, extraídas das bases PRIM e ULT do PIMC.

`Variável`	**Base**	**Descrição**
`cod_unificado`	PRIM e ULT	Identificador único de cada indivíduo, utilizado para integração entre as bases e construção da coorte longitudinal.
`sexo_cat`	PRIM	Sexo de nascimento, categorizado como Homem, Mulher ou Não informado. Destaca-se que essa variável não contempla identidade de gênero, o que representa uma limitação para análises específicas de populações transgênero (Guerra; Seidl, 2010).

continua na próxima página

(continuação da Tabela 4.1)

Variável	Base	Descrição
raca_cat2	PRIM	Raça/cor autodeclarada no momento do cadastro: Indígena, Preta, Parda, Branca/Amarela ou Não informado.
escol_cat	PRIM	Escolaridade declarada no cadastro, categorizada como: 0–7 anos, 8–11 anos, 12 anos ou mais, ou Não informado.
idade_vinc_cat	PRIM	Faixa etária no primeiro registro: Menos de 13 anos, 13–17 anos, 18–24 anos, 25–39 anos, 40–59 anos, 60 anos ou mais, ou Não informado.
cd4_cat200	PRIM	Resultado da primeira contagem de linfócitos T CD4+: ≥ 200 células/mm^3, < 200 células/mm^3 ou Sem resultado.
cd4_cat350	PRIM	Classificação alternativa da primeira contagem de linfócitos T CD4+, com ponto de corte em 350 células/mm^3.
dias_diag_tarv_cat	PRIM	Intervalo entre o primeiro exame laboratorial (carga viral ou linfócitos T CD4+) e o início da TARV: No mesmo dia, 1–6 dias, 7–30 dias, 1–6 meses, Mais de 6 meses, Não iniciou TARV, Iniciou TARV antes dos exames no SUS ou Não há exames disponíveis.
status_ano	ULT	Situação anual da pessoa quanto ao tratamento antirretroviral: TARV regular, Interrupção (evento de interesse) ou Gap de tratamento.
deteccao_cv50	ULT	Situação da carga viral no ano de referência, considerando o ponto de corte de 50 cópias/mL: CV indetectável, CV detectável ou Sem dados disponíveis.

continua na próxima página

(continuação da Tabela 4.1)

Variável	Base	Descrição
uf_inst_completo	ULT	Unidade Federativa da instituição responsável pelo último exame clínico (carga viral ou CD4) realizado até o final de cada ano.

Fonte: Elaboração própria com base em Brasil. Ministério da Saúde (2024e).

A escolha metodológica por utilizar a variável deteccao_cv50, que adota o limiar de 50 cópias/mL como critério de detecção de carga viral, fundamentou-se em recomendações nacionais e internacionais (Brasil. Ministério da Saúde, 2024a; HHS Panel on Antiretroviral Guidelines for Adults and Adolescents, 2024). Esse ponto de corte é amplamente reconhecido como o parâmetro mais sensível e específico para definição de supressão virológica sustentada e como padrão-ouro para avaliação do sucesso terapêutico na TARV.

4.3 Sobre os Recursos Computacionais

Todas as análises foram executadas em um computador pessoal equipado com processador AMD Ryzen 7 5700U (1,80 GHz, 8 núcleos / 16 *threads*), 12 GB de memória RAM DDR4 e GPU integrada AMD Radeon Graphics (512 MB de VRAM compartilhada), sob sistema operacional Windows 11 de 64 bits. O treinamento da rede neural foi realizado integralmente em CPU, uma vez que a GPU integrada não oferece suporte ao *framework* CUDA utilizado pelo PyTorch para aceleração por GPU. O treinamento completo (50 *epochs*) foi concluído em tempo computacionalmente viável para o *hardware* utilizado.

As análises estatísticas e as visualizações gráficas deste estudo foram conduzidas utilizando dois ambientes computacionais complementares. As etapas de manipulação de dados, análise exploratória e elaboração das figuras do Capítulo 4 foram realizadas no *software* R, versão 4.5.0 (R Core Team, 2024), em ambiente RStudio (Posit Science, 2023). As análises de modelagem de sobrevivência apresentadas no Capítulo 5 foram implementadas em Python, versão 3.11, utilizando Jupyter Notebook (Kluyver *et al.*, 2016) como ambiente de desenvolvimento interativo. Os códigos completos estão disponíveis nos Apêndices B e C, respectivamente.

A documentação das análises em R foi organizada por meio de arquivos no formato Quarto (.qmd), que possibilita a integração entre código, resultados e texto descritivo

em um único documento, facilitando tanto a rastreabilidade quanto a reprodução dos procedimentos analíticos adotados (Posit Science, 2023).

Em relação ao armazenamento e à leitura dos dados, optou-se pela utilização do formato Apache Parquet (.parquet), uma solução de armazenamento colunar que apresenta importantes vantagens técnicas no contexto de grandes volumes de informação. Entre os principais benefícios desse formato destacam-se a capacidade de compressão eficiente, a leitura seletiva de colunas específicas e a integração direta com R por meio do pacote `arrow` (Apache Software Foundation, 2024; MotherDuck, 2023). Tais características mostraram-se particularmente relevantes para o presente estudo, considerando-se o processamento de mais de um milhão de registros individuais.

A adoção do formato Parquet resultou em uma redução substancial no tamanho dos arquivos de dados utilizados. No caso da base PRIM, a conversão do formato CSV para Parquet (Apêndice A) proporcionou uma diminuição aproximada de 93% no volume de armazenamento, reduzindo o tamanho original de 265 megabytes (MB) para 18 MB. De forma ainda mais expressiva, a base ULT apresentou uma redução de cerca de 94%, passando de 2,2 gigabytes (GB) para 127 MB após a conversão. Esses ganhos em termos de compactação não apenas otimizaram a gestão do espaço em disco, mas também reduziram significativamente os tempos de leitura e escrita durante as etapas de pré-processamento e análise estatística. Tal eficiência foi fundamental para viabilizar o processamento das análises realizadas neste estudo, sobretudo diante da elevada dimensão e complexidade da base de dados analisada.

Além das vantagens operacionais, a adoção do formato Parquet também se alinha às boas práticas de ciência de dados reprodutível, uma vez que se trata de um padrão aberto, de fácil portabilidade e compatível com diferentes plataformas analíticas.

A integração entre os ambientes R e Python foi viabilizada pelo pacote `pyreadr` (Fajardo, 2018), que permitiu a leitura direta dos arquivos `.RData` gerados após o pré-processamento descrito neste capítulo, assegurando a continuidade do fluxo analítico sem perda de informação ou necessidade de conversões intermediárias. O formato `.RData` utiliza serialização nativa do R com compressão *gzip*, resultando em arquivos compactos e de leitura rápida. Ao ser carregado pelo `pyreadr`, o arquivo é descomprimido e convertido diretamente para um *DataFrame* do `pandas`, de modo que, uma vez na memória, o formato de origem não influencia o tempo de processamento das etapas subsequentes de modelagem. O fluxo completo de dados seguiu, portanto, a seguinte sequência: as bases brutas em CSV foram convertidas para o formato Parquet, permitindo leitura eficiente

das bases PRIM e ULT no R; em seguida, realizou-se o pré-processamento e a imputação no R, com exportação da base resultante em formato `.RData`; por fim, o arquivo foi lido no Python via `pyreadr` para o treinamento do modelo *Nnet-Survival* (Capítulo 5).

Durante a elaboração das representações gráficas deste trabalho, buscou-se garantir a acessibilidade visual das figuras, assegurando a legibilidade adequada tanto em meios digitais quanto em materiais impressos em escala de cinza. A escolha das paletas de cores considerou, de forma prioritária, a inclusão de pessoas com diferentes tipos de deficiência na percepção cromática, bem como a necessidade de minimizar distorções visuais que pudessem comprometer a interpretação dos resultados.

Para as variáveis de natureza contínua, optou-se pela utilização de paletas perceptualmente uniformes desenvolvidas por Fabio Crameri, implementadas no R por meio do pacote `scico` (Crameri, 2023). Essas paletas foram projetadas com base em princípios de uniformidade perceptual e de ordenação intuitiva de cores, de modo a assegurar que variações iguais nos dados fossem representadas por diferenças visuais igualmente perceptíveis. Conforme destacado por Crameri, Shephard e Heron (2020), o uso de esquemas de cores não científicos, como os tradicionalmente conhecidos como *rainbow* ou *jet*, pode introduzir distorções artificiais, mascarar padrões reais e excluir leitores com deficiência na visão de cores.

No caso das variáveis categóricas e discretas, foi adotada a paleta *Normal, 12 colors*, proposta por Tsitsulin (2019). Essa paleta foi desenvolvida por meio de um processo de otimização computacional, utilizando métricas de distância perceptual em espaços de cor, com o objetivo de maximizar o contraste entre categorias distintas. Além disso, foi concebida para manter a distinção entre cores, mesmo em condições de impressão monocromática e sob simulações de diferentes tipos de daltonismo. Tal abordagem busca minimizar o risco de sobreposição visual entre categorias e garantir que os resultados sejam interpretáveis por um público amplo e diverso.

A combinação dessas estratégias visou reduzir potenciais vieses perceptuais e promover a ampla acessibilidade das visualizações apresentadas ao longo deste estudo, em conformidade com as recomendações metodológicas atuais para comunicação visual em pesquisas científicas.

4.4 Metodologia de Pré-processamento

O pré-processamento dos dados representou uma etapa central na construção da base analítica deste estudo, sendo fundamental para garantir a consistência, a integridade e a adequação metodológica dos registros utilizados nas análises subsequentes. Considerando o caráter longitudinal e a natureza administrativa das informações provenientes do PIMC (Brasil. Ministério da Saúde, 2024e), foi necessário adotar um conjunto estruturado de procedimentos de verificação e transformação dos dados.

A primeira etapa envolveu a sanitização da base, com o objetivo de tratar problemas relacionados à codificação de caracteres, inconsistências na formatação e duplicidades nas categorias de variáveis qualitativas. Foram realizadas a padronização de nomenclaturas, a correção de categorias com grafias divergentes e a conversão de campos textuais em fatores categóricos, de acordo com os requisitos das técnicas estatísticas a serem aplicadas nas etapas posteriores.

Na sequência, procedeu-se à integração das bases PRIM e ULT, ambas derivadas do PIMC. Essa integração teve como variável-chave o campo `cod_unificado`, que corresponde ao identificador exclusivo de cada indivíduo no sistema. A junção das informações possibilitou a construção de uma base longitudinal, reunindo dados sociodemográficos, informações clínicas iniciais e o histórico anual de acompanhamento de cada participante.

A definição da variável resposta, correspondente ao tempo até o evento (*time-to-event*), representada por T_i, seguiu os pressupostos clássicos da Análise de Sobrevivência (Kleinbaum; Klein, 2020; Colosimo; Giolo, 2024). O tempo de seguimento foi calculado em anos completos, considerando a diferença entre o ano da primeira dispensação de terapia antirretroviral (TARV), obtido a partir da base PRIM, e o ano de ocorrência do evento de interesse ou da censura, conforme registrado na variável `status_ano`, proveniente da base ULT. Dessa forma, a variável T_i assumiu valores inteiros não negativos[1], refletindo o tempo de exposição ao risco de interrupção da TARV para o indivíduo i.

Indivíduos que não apresentaram o evento de interrupção até o final do período de acompanhamento foram considerados censurados à direita ($\delta_i = 0$), em conformidade com a definição operacional da variável `status_ano`. Essa abordagem segue as recomendações metodológicas vigentes para o tratamento de dados censurados em estudos de tempo até evento (Kalbfleisch; Prentice, 2002).

[1]Valores iguais a zero ($T_i = 0$) ocorrem quando o paciente apresenta o evento de interrupção ou é censurado ainda no primeiro ano de início da TARV, ou seja, quando não completa ao menos um ano de tratamento.

Uma etapa particularmente sensível deste processo foi a identificação e o tratamento de inconsistências temporais. Foram observados registros em que o *tempo de seguimento* apresentava valores negativos, ou seja, casos em que o ano de ocorrência do evento de interesse (interrupção da TARV) ou da censura era anterior ao ano de início da TARV, resultando em $T_i < 0$. Essa situação representa uma violação dos pressupostos fundamentais da Análise de Sobrevivência, especialmente no que se refere à definição temporal do risco de ocorrência do evento (Kleinbaum; Klein, 2020; Kalbfleisch; Prentice, 2002). É importante destacar que tais casos não configuram um exemplo de censura à esquerda, pois esse tipo de censura ocorre quando o evento já aconteceu antes do início da janela de observação e pode ser tratado por métodos estatísticos específicos. Porém, as situações identificadas neste estudo refletem erros de codificação ou inconsistências nas datas registradas nas bases administrativas, sem respaldo lógico-temporal. A exclusão desses registros da população de análise foi, portanto, a estratégia metodologicamente adequada e recomendada na literatura, com o objetivo de assegurar a coerência lógica da variável tempo e minimizar a introdução de vieses nas estimativas de risco (Chalita; Colosimo; Demétrio, 2002).

Por fim, realizou-se uma análise exploratória detalhada da estrutura de dados faltantes nas covariáveis selecionadas para os modelos de sobrevivência. As estratégias adotadas para o tratamento desses dados ausentes, incluindo os procedimentos de imputação, serão descritas na Seção 4.6.

4.5 Definição da Coorte

A definição da coorte de análise constituiu uma etapa fundamental para assegurar a validade das estimativas de sobrevivência obtidas neste estudo. O ponto de partida incluiu todos os indivíduos inicialmente registrados nas bases PRIM e ULT, ambas oriundas do PIMC (Brasil. Ministério da Saúde, 2024e). No total, o conjunto de dados inicial contabilizava 1.290.118 registros individuais.

Após a exclusão das observações com inconsistências cronológicas ($T_i < 0$), conforme explicado na Seção 4.4, a base analítica passou a contar com 1.094.567 indivíduos elegíveis.

É importante destacar que, após a etapa de imputação dos dados ausentes nas covariáveis de interesse, não houve exclusões adicionais. Todos os indivíduos cuja trajetória cronológica se manteve válida permaneceram na base final de análise. O tratamento dos dados faltantes foi realizado por meio de imputação não paramétrica, utilizando o

algoritmo `missForest`, conforme será detalhado na Seção 4.6.

Outro aspecto relevante refere-se à distribuição geográfica da coorte final. Mesmo após as exclusões por inconsistências temporais, nenhuma Unidade Federativa (UF) foi totalmente eliminada da População Válida. Esse resultado garantiu a preservação da representatividade territorial da amostra, condição essencial para estudos de base populacional que buscam avaliar desigualdades regionais nos desfechos analisados (Cunha *et al.*, 2025; Brasil. Ministério da Saúde, 2024e).

O fluxo de formação da População Válida, com os quantitativos absolutos de indivíduos em cada etapa de processamento, está sintetizado na Tabela 4.2.

Tabela 4.2 – Fluxo de formação da População Válida após as etapas de exclusão (N = número de indivíduos).

Etapa	***N***
Total inicial nas bases PRIM e ULT	1.290.118
Após exclusão por inconsistência cronológica ($T_i < 0$)	1.094.567
População Válida final (após imputação)	1.094.567

Fonte: Elaboração própria.

Esse processo de definição da coorte teve como objetivo minimizar potenciais fontes de viés de seleção e garantir que as análises de sobrevivência subsequentes reflitam, de maneira válida e representativa, a dinâmica da interrupção da TARV entre pessoas vivendo com HIV/Aids no Brasil.

4.6 Tratamento de Inconsistências e Dados Ausentes

A utilização de registros administrativos provenientes de sistemas nacionais de saúde apresenta, por sua própria natureza, desafios inerentes relacionados à qualidade dos dados. Entre os problemas mais frequentemente observados destacam-se as inconsistências cronológicas e a ocorrência de dados faltantes em variáveis essenciais para a análise. Essas limitações são amplamente reconhecidas na literatura que aborda o uso de bases secundárias em pesquisas epidemiológicas e de vigilância em saúde pública (Brasil. Ministério da Saúde, 2022).

No contexto deste estudo, a identificação e o tratamento dessas questões configuraram

um desafio relevante, com impactos diretos sobre a qualidade e a robustez da base analítica empregada na modelagem de sobrevivência. As estratégias adotadas para lidar com as inconsistências e os dados ausentes, detalhadas no decorrer dessa seção, buscaram minimizar potenciais fontes de viés e assegurar a validade das análises estatísticas subsequentes.

4.6.1 Dados Ausentes nas Covariáveis

Além das inconsistências cronológicas previamente abordadas, a base de dados analisada apresentou proporções expressivas de dados ausentes em diversas covariáveis consideradas essenciais para as análises. Esse fenômeno é amplamente reconhecido em estudos que utilizam grandes bases de dados administrativas, como as integradas no PIMC, sendo resultado de limitações inerentes aos processos de registro e consolidação das informações em sistemas nacionais de saúde (Brasil. Ministério da Saúde, 2024e).

As proporções de ausência variaram de forma significativa entre as diferentes variáveis. A variável *raça/cor declarada*, por exemplo, apresentava 10,9% de registros faltantes antes da aplicação de qualquer método de imputação. Já a variável *escolaridade* apresentava um percentual ainda mais elevado de ausência, alcançando 31,8%. Em relação à variável *CD4 inicial*, observou-se uma taxa de 9,8% de dados faltantes. Por outro lado, as variáveis *sexo de nascimento*, *faixa etária ao início do tratamento* e *carga viral com corte de 50 cópias/mL* não apresentaram registros ausentes.

A literatura especializada destaca que a falta de informações em variáveis sociodemográficas, como escolaridade e raça/cor, pode refletir desigualdades estruturais no acesso ao diagnóstico, bem como fragilidades nos processos de coleta e registro de dados nos serviços de saúde. Tal cenário pode acarretar riscos relevantes de viés de informação, especialmente quando os dados ausentes não são tratados de forma metodologicamente adequada (Cunha *et al.*, 2025; Guerra; Seidl, 2010).

Considerando esse contexto, optou-se pela adoção de métodos de imputação não paramétricos, reconhecidos por sua robustez frente a padrões complexos de ausência e por sua capacidade de lidar com conjuntos de variáveis de diferentes naturezas (categóricas e numéricas). A estratégia utilizada será detalhada na Subseção 4.6.2.

4.6.2 Estratégia de Imputação: o Algoritmo `missForest`

Considerando as limitações práticas e metodológicas dos métodos tradicionais de imputação múltipla baseados em modelos paramétricos, como o *Multivariate Imputation by Chained Equations* (MICE), que apresentaram instabilidade computacional quando aplicados ao presente conjunto de dados, optou-se pela utilização do algoritmo `missForest`, implementado no R por meio do pacote homônimo (Martins, 2017; Stekhoven, 2022).

O `missForest` é um método de imputação não paramétrico que emprega a técnica de *Random Forests* para predizer os valores ausentes com base em informações disponíveis nas demais variáveis observadas (Stekhoven; Bühlmann, 2012). Entre suas principais vantagens, destacam-se a capacidade de modelar relações complexas, não lineares e com múltiplas interações, além da robustez para lidar simultaneamente com variáveis contínuas e categóricas (Breiman, 2001).

O processo de imputação por meio do `missForest` segue uma abordagem iterativa. Inicialmente, os valores faltantes são preenchidos por imputações simples, como médias para variáveis contínuas e modas para variáveis categóricas. Em seguida, o algoritmo ajusta sucessivos modelos de *Random Forest* para cada variável com dados ausentes, utilizando as demais variáveis como preditoras. A cada ciclo, os valores imputados são atualizados com base nas novas previsões, e o procedimento se repete até atingir um critério de convergência, geralmente baseado na redução do erro fora-da-amostra[2] (Stekhoven; Bühlmann, 2012; Albu *et al.*, 2024).

Além de sua eficiência computacional, evidências de estudos comparativos indicam que o `missForest` apresenta desempenho superior a outros métodos de imputação, tanto no que se refere ao menor erro médio quanto à preservação da estrutura de correlação entre as variáveis e ao desempenho preditivo satisfatório dos modelos subsequentes (Waljee *et al.*, 2013; Bianchi, 2022). Essas características tornam o método especialmente adequado para contextos de alta dimensionalidade e grandes volumes de dados, como o deste estudo.

Entretanto, é importante reconhecer as limitações dessa abordagem. Por se tratar de uma técnica de imputação única (*single imputation*), o `missForest` não permite a quantificação direta da variabilidade introduzida pelo processo de imputação, ao contrário dos métodos baseados em imputação múltipla (*Multiple Imputation*) (Azur *et al.*, 2011; Martins, 2017). Essa limitação foi devidamente considerada na interpretação dos resultados, e futuras extensões metodológicas poderão explorar abordagens híbridas,

[2] O erro fora-da-amostra (*out-of-bag error*) é calculado usando observações que ficaram de fora do processo de reamostragem (*bootstrap*) em cada árvore, funcionando como validação interna do modelo.

incluindo combinações entre imputação múltipla e algoritmos baseados em *Random Forest*, conforme sugerido por estudos recentes (Albu *et al.*, 2024).

Durante o processo de imputação, foi dada uma atenção especial à variável com maior percentual de ausência, notadamente *escolaridade*. As métricas de erro de imputação fornecidas pelo `missForest` foram cuidadosamente avaliadas, a fim de garantir que os valores estimados apresentassem coerência com a distribuição original dos dados e com os perfis epidemiológicos historicamente documentados para a população brasileira vivendo com HIV/Aids (Cunha *et al.*, 2025; Menezes *et al.*, 2018).

4.6.3 Base Pós-Imputação

Ao término do processo de imputação, obteve-se uma base de dados completa em relação às variáveis selecionadas para os modelos de Análise de Sobrevivência, totalizando 1.094.567 indivíduos, agora sem apresentar informações faltantes nas covariáveis de interesse. Esse procedimento permitiu maximizar o aproveitamento amostral, reduzindo perdas decorrentes da exclusão de casos com informações incompletas e, consequentemente, ampliando o poder estatístico das análises subsequentes. A Tabela 4.3 apresenta as proporções de dados ausentes antes da aplicação do algoritmo `missForest`. Após a imputação, não houve mais registros com dados faltantes.

Tabela 4.3 – Resumo das proporções de dados ausentes antes da imputação via `missForest`.

Variável	**% Ausentes (Pré-Imputação)**
Sexo de nascimento	0,0%
Raça/cor declarada	10,9%
Escolaridade	31,8%
Faixa etária ao início da TARV	0,0%
CD4 inicial	9,8%
Carga viral ($<$ 50 cópias/mL)	0,0%

Fonte: Elaboração própria.

A opção metodológica pela realização de uma imputação única (*single imputation*) foi motivada por limitações computacionais, particularmente relevantes diante do elevado volume de registros processados, e pela necessidade de garantir a rastreabilidade e a

reprodutibilidade do *pipeline*[3] analítico.

Adicionalmente, foi realizada uma avaliação comparativa das distribuições marginais das variáveis antes e após a imputação, com o intuito de verificar a manutenção da plausibilidade epidemiológica dos dados estimados (Stekhoven; Bühlmann, 2012; Waljee *et al.*, 2013). Essa análise evidenciou que as distribuições obtidas após a imputação permaneceram coerentes com os padrões historicamente descritos em coortes nacionais de pessoas vivendo com HIV/Aids (Cunha *et al.*, 2025; Menezes *et al.*, 2018).

4.7 Caracterização da População Válida

4.7.1 Perfil Sociodemográfico da Coorte

A caracterização sociodemográfica da População Válida revelou um perfil predominantemente composto por indivíduos do sexo masculino ao nascimento, com baixos níveis de escolaridade e concentração majoritária nos adultos jovens (18 a 30 anos).

No que se refere ao sexo de nascimento, observou-se que 67,21% dos indivíduos da População Válida eram classificados como homens e 32,79% como mulheres. É importante ressaltar que a variável utilizada para essa classificação (`sexo_cat`) corresponde ao sexo biológico registrado no primeiro vínculo com os serviços de saúde, não contemplando a identidade de gênero. Consequentemente, pessoas transgênero estão classificadas de acordo com o sexo atribuído ao nascer, o que representa uma limitação relevante para análises de gênero mais aprofundadas. Essa restrição é reconhecida em documentos oficiais, que apontam a necessidade de avanços na coleta de dados sobre identidade de gênero nos sistemas nacionais de vigilância em HIV/Aids (Brasil. Ministério da Saúde, 2024e).

Em relação à raça/cor declarada, identificou-se que 43,35% da coorte era composta por indivíduos autodeclarados como sendo pessoas *brancas/amarelas*, seguidos por 42,67% de pessoas *pardas*, 13,14% de pessoas *pretas* e 0,84% de pessoas *indígenas*. Essa configuração é compatível com o perfil racial dos usuários do Sistema Único de Saúde (SUS) em âmbito nacional (Azevedo, 2022). No entanto, é importante destacar que a literatura aponta barreiras estruturais adicionais enfrentadas por pessoas negras e indígenas no acesso ao diagnóstico, início oportuno da TARV e manutenção da adesão ao tratamento (Brasil. Ministério da Saúde, 2022; Brito; Castilho; Szwarcwald, 2001). Entre os fatores citados, incluem-se discriminação institucional, desigualdades territoriais na oferta de serviços

[3] Fluxo sequencial e automatizado de etapas de pré-processamento, análise e geração de resultados, utilizado para assegurar padronização e reprodutibilidade dos procedimentos analíticos.

especializados e restrições no acesso a recursos sociais que favorecem a continuidade do tratamento.

No que diz respeito à escolaridade, observou-se um predomínio de indivíduos com o ensino fundamental incompleto (39,12%), seguidos por aqueles com ensino fundamental completo ou médio incompleto (31,53%) e, por fim, por indivíduos com ensino médio completo ou superior (29,35%). Essa distribuição reflete o perfil educacional da população atendida pelo SUS e está em consonância com estudos anteriores que descrevem características semelhantes em coortes nacionais de pessoas vivendo com HIV/Aids (Cunha *et al.*, 2025; Menezes *et al.*, 2018). Na literatura, níveis educacionais mais baixos têm sido associados à menor compreensão sobre a importância da adesão terapêutica, à maior vulnerabilidade socioeconômica e à exposição a contextos de vida adversos, fatores que podem comprometer o engajamento no cuidado ao HIV/Aids (Costa *et al.*, 2021).

Quanto à faixa etária ao início da TARV, verificou-se que quase metade da coorte (49,16%) iniciou o tratamento entre 25 e 39 anos de idade. As demais proporções foram: 27,51% entre 40 e 59 anos, 17,04% entre 18 e 24 anos, 3,18% acima de 60 anos, 1,67% até 13 anos e apenas 1,44% entre 13 e 17 anos. Em particular, o número reduzido de adolescentes em tratamento (faixa de 13 a 17 anos) merece atenção, dado que a literatura aponta a adolescência como um período crítico para a adesão à TARV. Estudos qualitativos indicam que adolescentes enfrentam desafios específicos relacionados ao uso contínuo da medicação, à aceitação do diagnóstico e à obtenção de suporte familiar e institucional, fatores que podem dificultar a adesão ou resultar em maiores taxas de interrupção do tratamento (Guerra; Seidl, 2010).

A Figura 4.1 apresenta a distribuição das variáveis sociodemográficas da coorte, enquanto a Tabela 4.4 resume os valores absolutos e percentuais correspondentes a cada categoria.

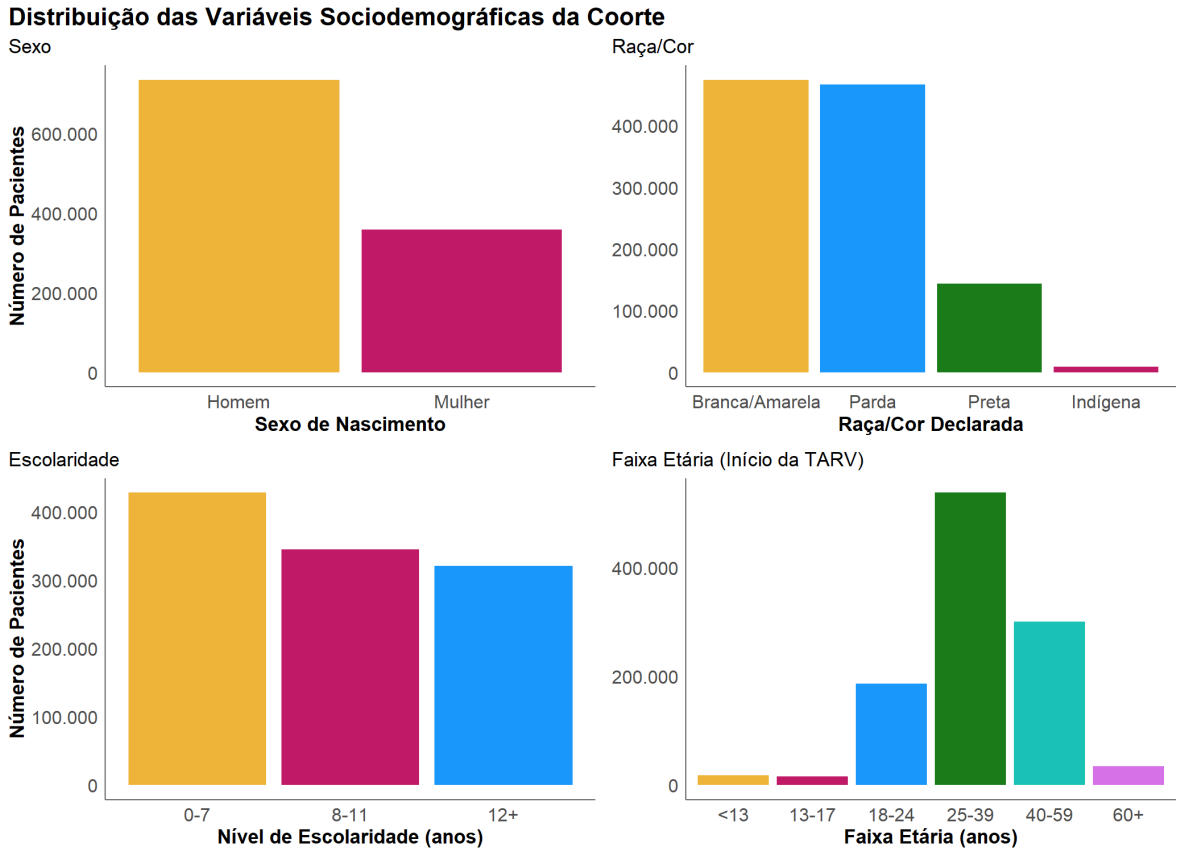

Figura 4.1 – Distribuição das variáveis sociodemográficas da coorte segundo sexo de nascimento, raça/cor, escolaridade e faixa etária ao início da TARV.

Fonte: Elaboração própria.

Tabela 4.4 – Distribuição da População Válida segundo variáveis sociodemográficas.

Variável	Categoria	Proporção
Sexo de nascimento	Homem	67,21%
	Mulher	32,79%
Raça/cor declarada	Branca/Amarela	43,35%
	Parda	42,67%
	Preta	13,14%
	Indígena	0,84%

continua na próxima página

(continuação da Tabela 4.4)

Variável	Categoria	Proporção
Escolaridade	0–7 anos	39,12%
	8–11 anos	31,53%
	12 anos ou mais	29,35%
Faixa etária ao início da TARV	13 anos ou menos	1,67%
	13–17 anos	1,44%
	18–24 anos	17,04%
	25–39 anos	49,16%
	40–59 anos	27,51%
	60 anos ou mais	3,18%

Fonte: Elaboração própria.

A análise conjunta da distribuição por sexo e faixa etária pode ser melhor visualizada na Figura 4.2, que apresenta a pirâmide etária da coorte estratificada por sexo de nascimento. Observa-se um predomínio masculino nas faixas etárias de 25 a 39 anos, que concentram o maior número absoluto de pacientes. Nas faixas etárias mais avançadas (40 a 59 anos e 60 anos ou mais), a distribuição entre os sexos tende a ser mais equilibrada, enquanto entre os mais jovens (abaixo de 18 anos), há uma proporção bem menor de pacientes, com equilíbrio entre os sexos.

Essa estrutura etária reforça o caráter predominantemente adulto-jovem da epidemia de HIV no Brasil e destaca a importância de considerar o fator idade nas análises de sobrevivência, uma vez que o risco de interrupção da TARV pode variar substancialmente entre as diferentes faixas etárias e entre os sexos.

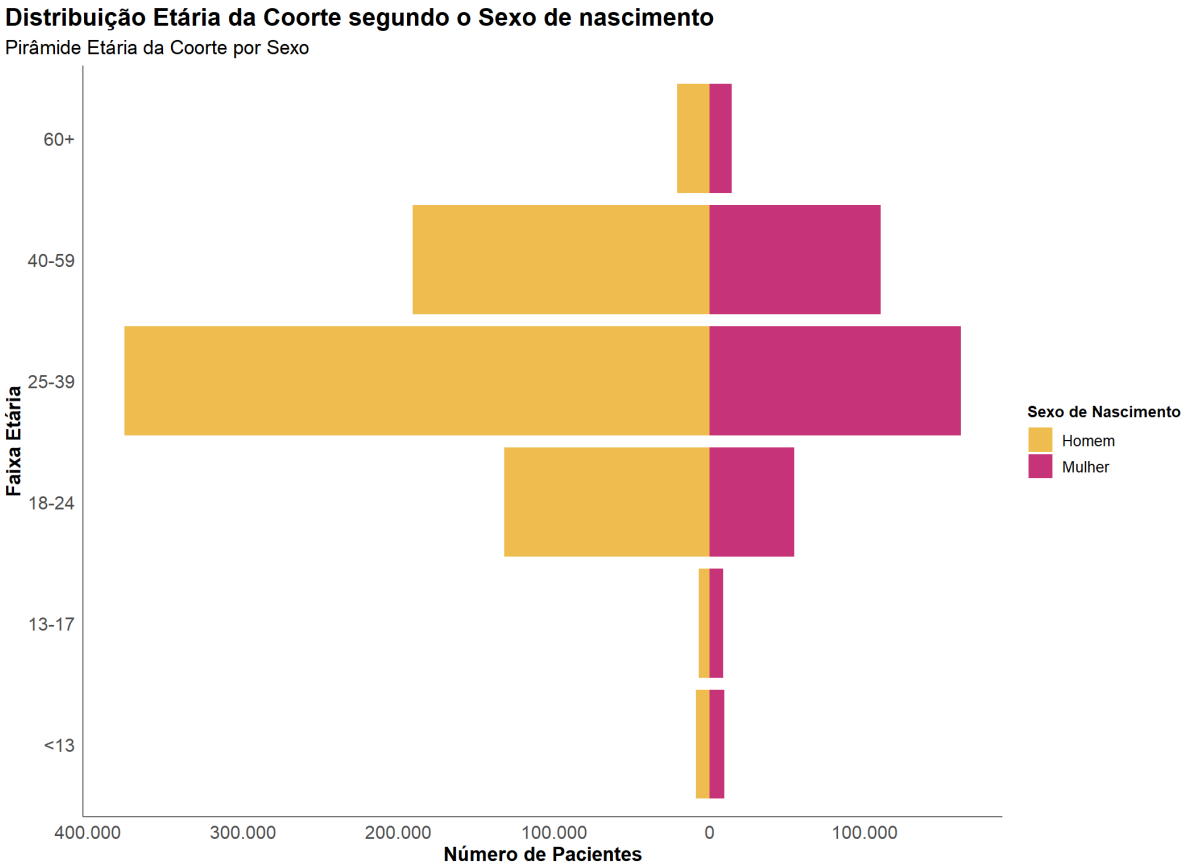

Figura 4.2 – Pirâmide etária da População Válida segundo sexo de nascimento e faixa etária ao início da TARV.

Fonte: Elaboração própria.

4.7.2 Perfil Clínico Basal

A caracterização clínica da coorte, a partir das variáveis disponíveis no início da TARV, revelou aspectos relevantes do estado de saúde dos indivíduos ao ingressarem no tratamento, refletindo padrões historicamente descritos nas populações atendidas pelo Sistema Único de Saúde (SUS) (Brasil. Ministério da Saúde, 2024a; Melo *et al.*, 2019).

No que se refere à contagem de linfócitos T CD4+, observou-se que 21,54% dos pacientes apresentavam valores inferiores a 200 células/mm^3 na primeira medição disponível, sugerindo um diagnóstico tardio e a presença de imunossupressão avançada. Em contrapartida, 78,30% dos indivíduos apresentavam contagens iguais ou superiores a 200 células/mm^3, indicando início do tratamento em estágios menos avançados da infecção. Vale destacar que apenas 0,16% dos participantes estavam classificados como "Sem resultado de CD4", valor considerado válido na base.

Ao considerar o ponto de corte de 350 células/mm^3, verificou-se que 39,35% da coorte apresentava valores abaixo desse limiar, reforçando a evidência de que uma parcela expressiva dos pacientes iniciou a TARV em condições clínicas avançadas. Já 60,49% apresentavam contagens iguais ou superiores a 350 células/mm^3. Este achado está em consonância com estudos anteriores, que apontam para o atraso no diagnóstico e no início do tratamento entre usuários do SUS (Cunha *et al.*, 2025; Menezes *et al.*, 2018). Além disso, é amplamente reconhecido que baixos níveis de CD4 estão associados ao aumento do risco de morbidade e mortalidade relacionadas ao HIV (HHS Panel on Antiretroviral Guidelines for Adults and Adolescents, 2024).

Quanto à carga viral inicial, medida pela variável `deteccao_cv50`, 25,39% dos indivíduos apresentaram carga viral indetectável (< 50 cópias/mL) na primeira medição disponível na base, o que reflete pacientes que já haviam atingido supressão virológica após início prévio da TARV. Apenas 6,40% tinham carga viral detectável, enquanto 68,21% não possuíam dados de carga viral disponíveis para o momento inicial. A opção metodológica por adotar o ponto de corte de 50 cópias/mL fundamenta-se em sua elevada especificidade para a detecção de supressão virológica sustentada, sendo este o critério amplamente reconhecido como padrão-ouro nas diretrizes clínicas nacionais e internacionais para o monitoramento da resposta terapêutica à TARV (Brasil. Ministério da Saúde, 2024a; HHS Panel on Antiretroviral Guidelines for Adults and Adolescents, 2024).

A Figura 4.3 apresenta a distribuição gráfica das variáveis clínicas de linha de base, enquanto a Tabela 4.5 sintetiza os valores percentuais correspondentes.

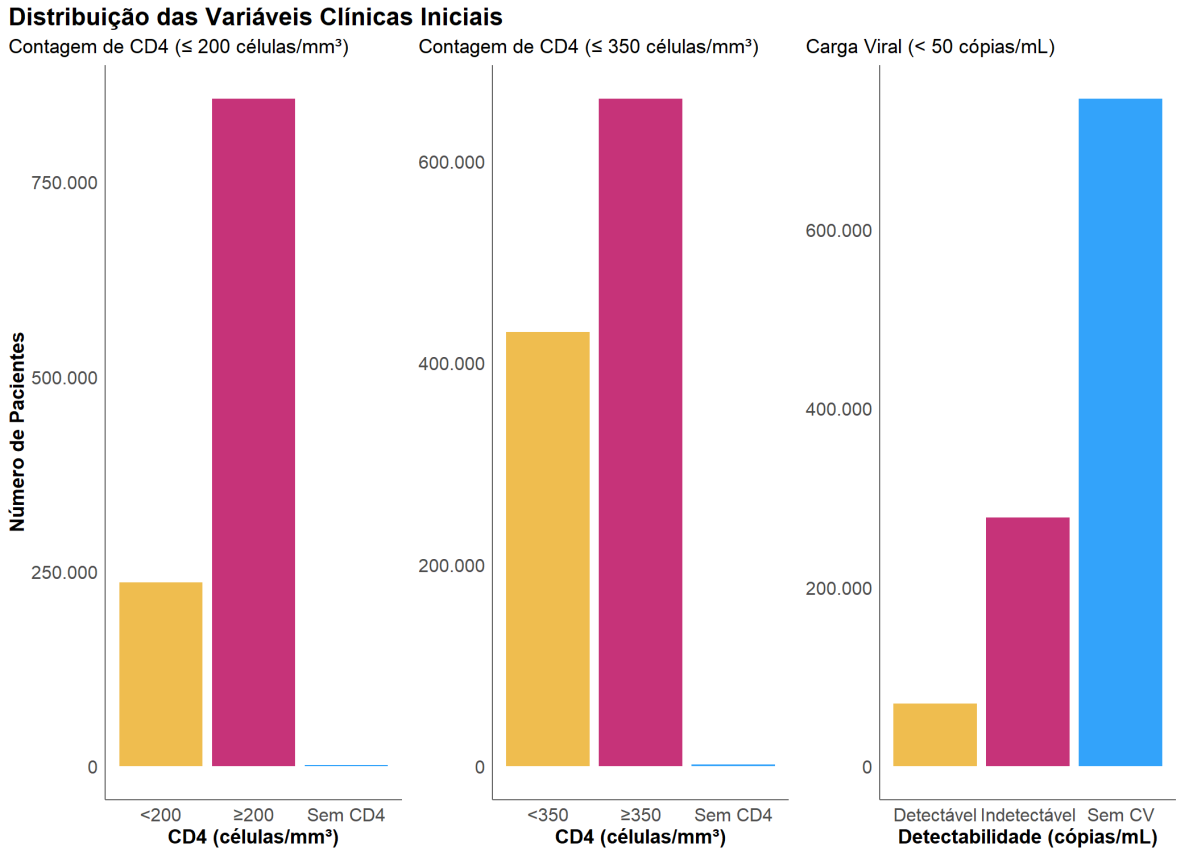

Figura 4.3 – Distribuição das variáveis clínicas basais da coorte segundo CD4 inicial e carga viral (< 50 cópias/mL).

Fonte: Elaboração própria.

Tabela 4.5 – Distribuição da População Válida segundo variáveis clínicas de linha de base.

Variável	**Proporção**
CD4 inicial (< 200 células/mm^3)	21,54%
CD4 inicial (≥ 200 células/mm^3)	78,30%
Sem resultado de CD4 (corte 200 células/mm^3)	0,16%
CD4 inicial (< 350 células/mm^3)	39,35%
CD4 inicial (≥ 350 células/mm^3)	60,49%

continua na próxima página

(continuação da Tabela 4.5)

Variável	Proporção
Sem resultado de CD4 (corte 350 células/mm^3)	0,16%
Carga viral indetectável ($<$ 50 cópias/mL)	25,39%
Carga viral detectável ($\geq$ 50 cópias/mL)	6,40%
Sem dados de CV	68,21%

Fonte: Elaboração própria.

4.7.3 Variação Geográfica na Qualidade dos Registros

A análise da distribuição geográfica da População Válida evidenciou importantes variações na qualidade e na consistência dos registros entre as diferentes Unidades Federativas (UFs) brasileiras. Embora nenhuma UF tenha sido completamente excluída da análise após a limpeza dos dados, observou-se uma heterogeneidade expressiva nas proporções de exclusão de indivíduos em razão de inconsistências cronológicas.

Conforme apresentado na Tabela 4.6, os estados com os maiores percentuais de exclusão foram o Rio Grande do Sul (17,7%), São Paulo (16,8%) e Mato Grosso do Sul (15,8%). Essas variações podem refletir, ao menos em parte, diferenças regionais na organização dos serviços de saúde, na infraestrutura de vigilância epidemiológica e na qualidade dos processos de registro nos sistemas oficiais (Grangeiro; Escuder; Castilho, 2010; Cunha; Cruz; Pedroso, 2022). Estudos anteriores destacam que, apesar de concentrarem grande número de casos de HIV/Aids, estados das regiões Sudeste e Sul frequentemente enfrentam desafios relacionados à atualização e à integridade dos registros administrativos, especialmente em bases de acompanhamento longitudinal.

Importante salientar que os padrões de exclusão observados não ocorreram de forma aleatória entre as UFs. Essa constatação reforça a necessidade de cautela na interpretação de comparações inter-regionais relacionadas aos tempos até a interrupção da TARV, uma vez que diferenças na qualidade dos registros e nos fluxos assistenciais podem impactar as análises de forma diferenciada entre os estados (Grangeiro; Escuder; Castilho, 2010; Cunha; Cruz; Pedroso, 2022).

Tabela 4.6 – Número de indivíduos (N) por UF antes e após a limpeza de dados para elegibilidade na coorte (dispostos em ordem alfabética).

UF	***N* Inicial**	***N* Válido Final**	***N* Excluídos**	**% Excluídos**
Acre	2.647	2.270	377	14,2%
Alagoas	14.908	12.702	2.206	14,8%
Amapá	5.090	4.458	632	12,4%
Amazonas	34.506	30.523	3.983	11,5%
Bahia	55.445	46.742	8.703	15,7%
Ceará	40.627	35.863	4.764	11,7%
Distrito Federal	22.102	19.244	2.858	12,9%
Espírito Santo	24.928	21.276	3.652	14,7%
Goiás	32.363	28.315	4.048	12,5%
Maranhão	29.121	25.524	3.597	12,4%
Mato Grosso	21.386	18.472	2.914	13,6%
Mato Grosso do Sul	17.263	14.530	2.733	15,8%
Minas Gerais	89.028	76.522	12.506	14,0%
Pará	51.056	43.303	7.753	15,2%
Paraíba	14.488	12.325	2.163	14,9%
Paraná	67.493	57.723	9.770	14,5%
Pernambuco	51.907	44.957	6.950	13,4%
Piauí	14.163	12.032	2.131	15,0%
Rio de Janeiro	161.873	136.798	25.075	15,5%
Rio Grande do Norte	15.529	13.456	2.073	13,3%

continua na próxima página

(continuação da Tabela 4.6)

UF	*N* Inicial	*N* Válido Final	*N* Excluídos	% Excluídos
Rio Grande do Sul	123.083	101.320	21.763	17,7%
Rondônia	9.875	8.374	1.501	15,2%
Roraima	5.192	4.527	665	12,8%
Santa Catarina	69.985	59.719	10.266	14,7%
São Paulo	300.106	249.747	50.359	16,8%
Sergipe	9.604	8.365	1.239	12,9%
Tocantins	6.350	5.480	870	13,7%
Total	**1.290.118**	**1.094.567**	**195.551**	**15,2%**

Fonte: Elaboração própria.

Além das discrepâncias nas taxas de exclusão, a análise descritiva da distribuição dos indivíduos por UF (Figura 4.4) revelou que os estados de São Paulo, Rio de Janeiro e Rio Grande do Sul concentraram, conjuntamente, um grande número absoluto de indivíduos na População Válida. Esse padrão reflete, em parte, o maior tamanho populacional dessas unidades federativas e a maior oferta de serviços especializados. Assim, não se pode inferir, a partir desses números absolutos, uma maior incidência ou prevalência da infecção nessas regiões sem considerar as respectivas populações de base (Grangeiro; Escuder; Castilho, 2010; Cunha; Cruz; Pedroso, 2022).

No que se refere ao tempo médio de permanência em TARV, calculado com base na variável *time_to_event*, observou-se uma variação geográfica considerável. As UFs com maior tempo médio de seguimento foram, em geral, aquelas com maior volume absoluto de pacientes, como São Paulo e Minas Gerais. Essa relação pode refletir, indiretamente, o maior tamanho populacional desses estados, sua maior densidade de serviços especializados e, possivelmente, melhores condições estruturais e de desenvolvimento socioeconômico que favorecem a retenção em cuidado. Por outro lado, estados com menor número de indivíduos, como Roraima e Amapá, apresentaram tempos médios ligeiramente inferiores.

A variável *time_to_event* considerou o número de anos completos transcorridos entre a primeira dispensação de TARV, registrada na base PRIM, e o ano de ocorrência do

evento de interesse (interrupção do tratamento) ou da censura, conforme os registros da base ULT. Esse procedimento seguiu os pressupostos metodológicos clássicos da Análise de Sobrevivência para dados agrupados por tempo discreto (Colosimo; Giolo, 2024; Kleinbaum; Klein, 2020).

As disparidades regionais observadas podem refletir não apenas diferenças nos perfis epidemiológicos das populações atendidas, mas também desigualdades estruturais no acesso aos serviços, na qualidade da assistência oferecida e na capacidade de retenção dos usuários no cuidado. Evidências recentes sugerem que fatores socioeconômicos, desigualdades raciais e disparidades regionais exercem influência direta sobre os desfechos de saúde relacionados ao HIV no Brasil (Grangeiro; Escuder; Castilho, 2010; Cunha; Cruz; Pedroso, 2022).

A Figura 4.4 apresenta, em painel duplo, a distribuição proporcional da coorte por UF e o tempo médio de seguimento em TARV por estado.

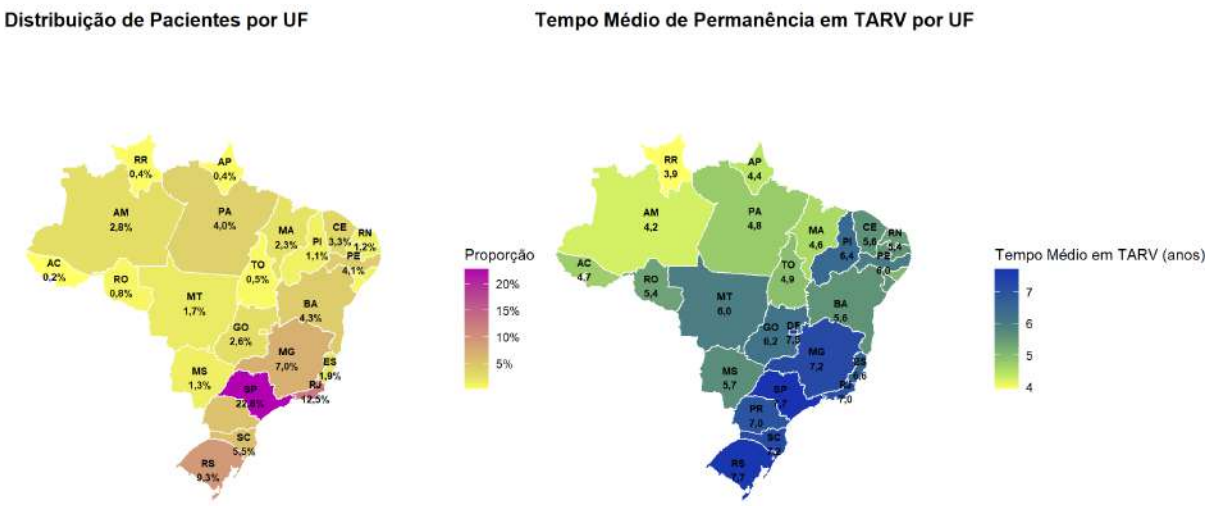

Figura 4.4 – Distribuição geográfica da População Válida segundo UF de atendimento. À esquerda: proporção de indivíduos por UF. À direita: tempo médio de permanência em TARV (em anos) por UF.

Fonte: Elaboração própria.

4.7.4 Considerações sobre os Dados Pós-Imputação

A aplicação do algoritmo `missForest` para imputação dos dados ausentes resultou em uma base de dados completa em relação às covariáveis utilizadas nas análises subsequentes,

com percentual final de dados ausentes reduzido a zero.

Esse aprimoramento na base de dados apresenta implicações metodológicas relevantes para a qualidade estatística das análises que serão conduzidas no Capítulo 5. Por um lado, a imputação possibilitou a maximização do aproveitamento amostral, evitando perdas adicionais de informação e reduzindo o risco de viés de seleção que poderia decorrer de análises restritas apenas aos casos completos. Por outro lado, é fundamental reconhecer as limitações inerentes à adoção de uma abordagem de imputação única (*single imputation*), especialmente quanto à impossibilidade de quantificar formalmente a variabilidade imputacional. Essa limitação pode impactar a precisão das estimativas de variância nos modelos inferenciais subsequentes (Azur *et al.*, 2011; Martins, 2017).

Adicionalmente, foi realizada uma análise comparativa das distribuições marginais das covariáveis antes e após o processo de imputação, com o objetivo de verificar a manutenção da plausibilidade epidemiológica dos dados. Essa comparação revelou que as frequências relativas das categorias das variáveis permaneceram consistentes, sem distorções substantivas, reforçando a adequação do método adotado e indicando que o procedimento de imputação não introduziu alterações artificiais nas principais características da coorte (Stekhoven; Bühlmann, 2012; Waljee *et al.*, 2013).

Com a base final consolidada, contendo 1.094.567 indivíduos com informações completas nas covariáveis de interesse, estamos metodologicamente preparados para a realização das análises de sobrevivência, que serão apresentadas no Capítulo 5.

Para concluir, destaca-se que o rigor metodológico aplicado em todas as etapas de pré-processamento, imputação e análise exploratória foi essencial para assegurar a robustez e a validade das inferências estatísticas que sustentam os resultados deste trabalho.

4.8 Tempo até a Interrupção da TARV segundo Características Sociodemográficas

Para além da caracterização descritiva das variáveis sociodemográficas e clínicas, foi realizada uma análise exploratória bivariada com o objetivo de avaliar a distribuição do tempo até a interrupção da TARV em função das principais covariáveis do estudo.

A Figura 4.5 apresenta, por meio de boxplots, a distribuição do tempo de seguimento em TARV (em anos) para cada categoria de sexo de nascimento, raça/cor declarada, escolaridade e faixa etária ao início do tratamento. De modo geral, as medianas de tempo de seguimento mostraram-se relativamente próximas na maioria dos grupos, porém com

algumas variações que merecem destaque.

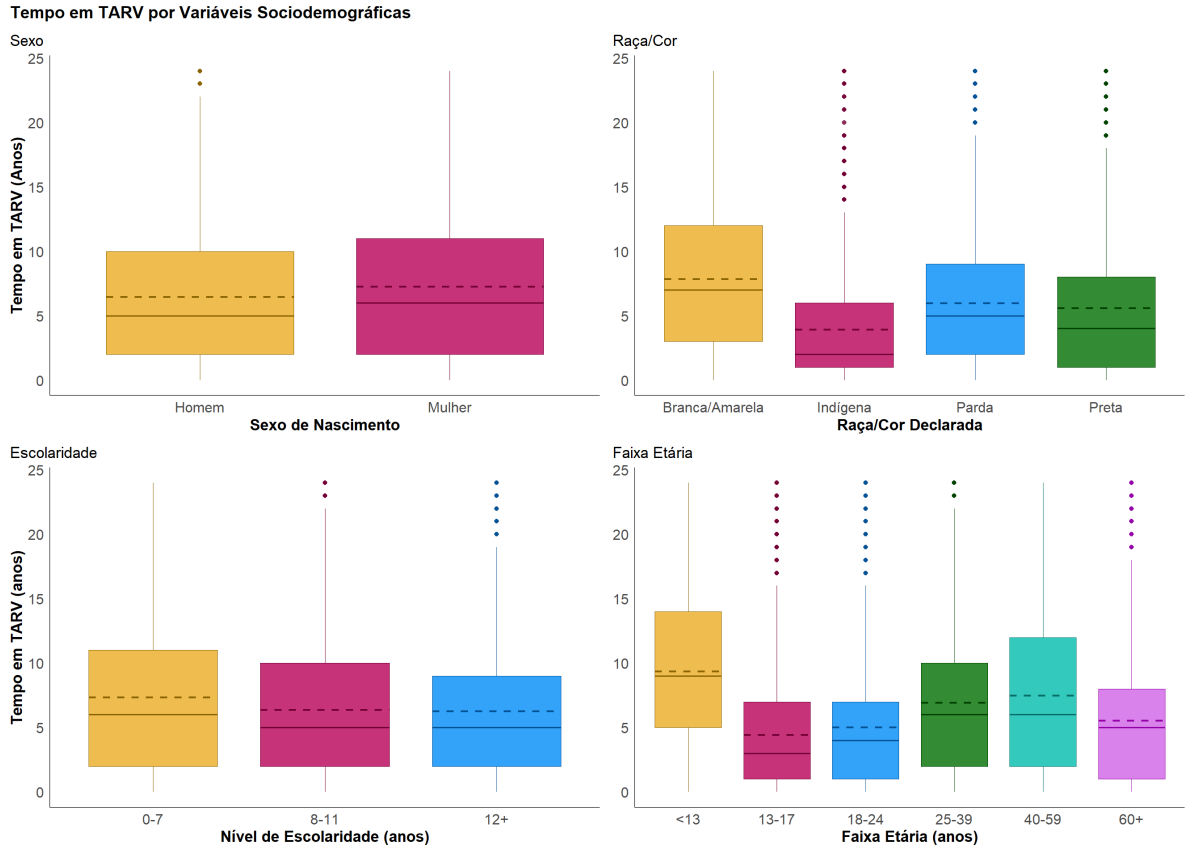

Figura 4.5 – Boxplots do tempo de permanência em TARV segundo sexo de nascimento, raça/cor, escolaridade e faixa etária ao início do tratamento. A linha contínua representa a mediana e a linha tracejada representa a média de cada grupo.

Fonte: Elaboração própria.

Observa-se que todas as distribuições apresentaram assimetria à direita (positiva), caracterizada pela presença da média sistematicamente superior à mediana e pela concentração dos dados nos valores mais baixos de tempo de seguimento, com caudas longas estendendo-se em direção aos tempos mais prolongados. Esse padrão é esperado em estudos de sobrevivência com alta taxa de censura, refletindo a predominância de interrupções precoces do tratamento e a presença de um subgrupo menor de indivíduos que permanecem em TARV por períodos substancialmente mais longos. A assimetria à direita sugere heterogeneidade nos padrões de adesão e reforça a adequação de métodos de análise de sobrevivência que não assumem distribuições simétricas ou normais dos tempos de evento (Kleinbaum; Klein, 2020).

Em relação ao sexo de nascimento, observou-se que os homens apresentaram uma

mediana de 5 anos em TARV (média de 6,45 anos), enquanto as mulheres alcançaram uma mediana de 6 anos (média de 7,24 anos). A diferença de aproximadamente 0,8 ano nas médias reforça a importância de explorar o potencial impacto do sexo biológico em análises de sobrevivência. Possivelmente, esse resultado reflete fatores sociocomportamentais, como o fato de as mulheres, em média, serem mais cuidadosas e apresentarem maior organização em relação ao cuidado com a saúde, favorecendo uma maior adesão e permanência no tratamento. Além disso, é importante ressaltar que a variável utilizada (`sexo_cat`) refere-se ao sexo atribuído ao nascimento, não contemplando identidade de gênero. Tal limitação pode resultar em subnotificação de pessoas transgênero, reconhecidamente mais vulneráveis à interrupção do tratamento (Brasil. Ministério da Saúde, 2024a).

Na análise segundo raça/cor, indivíduos autodeclarados como *Brancos/Amarelos* apresentaram a maior mediana de tempo em TARV (7 anos) e também a maior média (7,84 anos), seguidos por pessoas pardas (mediana de 5 anos, média de 5,97 anos), pretas (mediana de 4 anos, média de 5,58 anos) e indígenas (mediana de 2 anos, média de 3,92 anos). A diferença de quase 4 anos entre as médias do grupo Branca/Amarela e do grupo Indígena evidencia a magnitude das desigualdades estruturais que afetam o acesso e a permanência no tratamento, reconhecidas como determinantes sociais da saúde no Brasil. Estudos recentes apontam que pessoas negras e indígenas enfrentam maiores barreiras ao diagnóstico precoce, ao início oportuno da TARV e à manutenção da adesão, muitas vezes em decorrência de fatores como discriminação institucional, desigualdades socioeconômicas e limitações na oferta de serviços especializados (Brasil. Ministério da Saúde, 2024e).

Quanto à escolaridade, o tempo mediano de permanência em TARV foi de 6 anos entre pessoas com até 7 anos de estudo (média de 7,32 anos), e de 5 anos para as demais faixas de escolaridade (médias de 6,35 e 6,27 anos para 8–11 anos e 12 anos ou mais, respectivamente). Embora numericamente superior, a mediana e a média observadas no grupo de menor escolaridade devem ser interpretadas com cautela, uma vez que a amplitude interquartil (IQR) indicou maior heterogeneidade nesse grupo. Tal padrão, ainda que contraintuitivo, pode estar relacionado a políticas públicas específicas e programas de suporte intensivo do SUS voltados a populações de maior vulnerabilidade, os quais buscam mitigar as barreiras de acesso e promover a retenção no cuidado. Estudos anteriores destacam que indivíduos com menor escolaridade, em geral, enfrentam mais desafios para compreender o regime terapêutico e têm maior risco de interrupção do tratamento (Cunha *et al.*, 2025; Menezes *et al.*, 2018). Nesse sentido, a maior mediana observada pode refletir o efeito de programas compensatórios ou estratégias de busca ativa e acompanhamento

próximo, mas deve ser interpretada à luz das desigualdades estruturais que afetam a adesão ao tratamento.

Em relação à faixa etária ao início da TARV, destacaram-se as crianças menores de 13 anos, que apresentaram a maior mediana de tempo em TARV (9 anos) e também a maior média (9,35 anos), seguidas pelos adultos de 40 a 59 anos (mediana de 6 anos, média de 7,48 anos) e de 25 a 39 anos (mediana de 6 anos, média de 6,92 anos). Por outro lado, adolescentes de 13 a 17 anos (mediana de 3 anos, média de 4,41 anos) e jovens de 18 a 24 anos (mediana de 4 anos, média de 5,01 anos) apresentaram os menores tempos medianos e médios. Nota-se que idosos com 60 anos ou mais apresentaram mediana de 5 anos e média de 5,52 anos, valores intermediários que podem refletir tanto maior vulnerabilidade clínica quanto menor tempo potencial de seguimento devido à idade avançada ao início do tratamento. Esses achados corroboram a literatura que identifica a adolescência e o início da vida adulta como períodos de maior vulnerabilidade para a manutenção da adesão ao tratamento, em razão de fatores psicossociais, comportamentais e contextuais específicos dessas faixas etárias (Guerra; Seidl, 2010).

Esses resultados, ainda que descritivos, reforçam a importância de incluir essas covariáveis como fatores de ajuste nos modelos de sobrevivência, considerando sua potencial associação com o risco de interrupção da TARV. Além disso, evidenciam a necessidade de interpretar os desfechos clínicos à luz das desigualdades sociais, raciais e etárias.

Além das diferenças nas medianas e médias, a análise da amplitude interquartil (IQR) revelou padrões relevantes de dispersão nos tempos de seguimento. Por exemplo, a IQR entre os homens foi de 8 anos, enquanto entre as mulheres alcançou 9 anos, indicando maior heterogeneidade temporal entre estas. No que tange à raça/cor, o grupo *Branca/Amarela* apresentou a maior IQR (9 anos), ao passo que o grupo indígena exibiu a menor dispersão, com IQR de 5 anos.

No caso da escolaridade, destaca-se que o grupo com 0–7 anos de estudo apresentou tanto a maior mediana e média quanto a maior amplitude interquartil (9 anos), refletindo a coexistência de diferentes padrões de adesão dentro desse subgrupo.

A análise por faixa etária revelou um padrão não monótono. Crianças menores de 13 anos apresentaram a maior mediana (9 anos) e média (9,35 anos), além de ampla IQR (9 anos), enquanto adolescentes de 13 a 17 anos apresentaram a menor mediana (3 anos), a menor média (4,41 anos) e IQR igualmente reduzida (6 anos). Tal resultado reforça a evidência de que a adolescência constitui um período crítico em termos de risco para a interrupção do tratamento.

Observa-se ainda que, em todas as categorias analisadas, o tempo máximo de seguimento registrado foi de 24 anos, correspondente ao limite superior do período de observação disponível nas bases. Por outro lado, o tempo mínimo foi de 0 anos em todos os grupos, refletindo a ocorrência de eventos ou censuras antes de completar um ano da TARV.

A presença de valores extremos (outliers), concentrados nos tempos mais longos de seguimento, justifica a adoção de métodos estatísticos robustos e a realização de análises de sensibilidade nas etapas subsequentes da modelagem de sobrevivência (Kleinbaum; Klein, 2020).

Complementando a análise gráfica apresentada na Figura 4.5, as Tabelas 4.7 a 4.10 detalham as estatísticas descritivas do tempo de seguimento em TARV segundo sexo ao nascimento, raça/cor declarada, escolaridade e faixa etária ao início do tratamento. Os resultados incluem, para cada categoria, o número total de indivíduos (N), a média, os quartis (Q1, mediana e Q3) e o intervalo interquartil (IQR).

Tabela 4.7 – Estatísticas descritivas do tempo de seguimento em TARV por sexo de nascimento.

Sexo	***N***	**Média**	**Q1**	**Mediana**	**Q3**	**IQR**
Homem	735.706	6,45	2	5	10	8
Mulher	358.861	7,24	2	6	11	9

Fonte: Elaboração própria.

Tabela 4.8 – Estatísticas descritivas do tempo de seguimento em TARV por raça/cor declarada.

Raça/Cor	***N***	**Média**	**Q1**	**Mediana**	**Q3**	**IQR**
Branca/Amarela	474.483	7,84	3	7	12	9
Parda	467.069	5,97	2	5	9	7
Preta	143.819	5,58	1	4	8	7
Indígena	9.196	3,92	1	2	6	5

Fonte: Elaboração própria.

Tabela 4.9 – Estatísticas descritivas do tempo de seguimento em TARV por escolaridade.

Escolaridade	*N*	Média	Q1	Mediana	Q3	IQR
0–7 anos	428.208	7,32	2	6	11	9
8–11 anos	345.122	6,35	2	5	10	8
12 anos ou mais	321.237	6,27	2	5	9	7

Fonte: Elaboração própria.

Tabela 4.10 – Estatísticas descritivas do tempo de seguimento em TARV por faixa etária ao início do tratamento.

Faixa Etária	*N*	Média	Q1	Mediana	Q3	IQR
13 anos ou menos	18.252	9,35	5	9	14	9
13–17 anos	15.734	4,41	1	3	7	6
18–24 anos	186.486	5,01	1	4	7	6
25–39 anos	538.131	6,92	2	6	10	8
40–59 anos	301.092	7,48	2	6	12	10
60 anos ou mais	34.872	5,52	1	5	8	7

Fonte: Elaboração própria.

4.8.1 Distribuição Geral do Tempo de Seguimento e Classificação dos Eventos

Além das análises estratificadas por covariáveis, procedeu-se a uma avaliação global da distribuição do tempo de seguimento em TARV e da frequência dos diferentes tipos de evento observados na População Válida. Esses resultados encontram-se sintetizados na Figura 4.6 e na Tabela 4.11.

O histograma apresentado no painel superior esquerdo da Figura 4.6 evidencia uma concentração expressiva de indivíduos com tempos de seguimento entre 0 e 5 anos, seguida por um declínio progressivo no número de pacientes nos anos subsequentes. Tal padrão é

característico de coortes longitudinais com extensos períodos de acompanhamento, nas quais a combinação entre o risco acumulado de ocorrência do evento de interesse e o aumento da censura ao longo do tempo resulta em uma redução gradual da população em risco (Kleinbaum; Klein, 2020).

A distribuição dos tipos de evento ao final do período de observação, apresentada no painel superior direito da mesma figura e detalhada na Tabela 4.11, revela que, dos 1.094.567 indivíduos incluídos na População Válida, 714.659 (65,3%) foram classificados como censurados ($\delta_i = 0$), enquanto 379.908 (34,7%) apresentaram o desfecho de interesse ($\delta_i = 1$), definido como interrupção da terapia antirretroviral. A predominância de censura à direita constitui um aspecto metodológico de destaque, amplamente esperado em estudos de coorte com acompanhamento prolongado.

Esse tipo de censura, frequentemente denominado *censura administrativa*, ocorre quando o acompanhamento de um indivíduo é encerrado por razões relacionadas ao término do período de observação definido pelo pesquisador, e não pela ocorrência de um evento clínico específico (Kleinbaum; Klein, 2020). Tal cenário reforça a pertinência da Análise de Sobrevivência como abordagem estatística, ao permitir a consideração adequada desses casos na estimação dos riscos de interrupção da TARV. Adicionalmente, a elevada proporção de censura observada pode refletir tanto a efetividade das estratégias de retenção em cuidado implementadas nos serviços de saúde quanto limitações inerentes aos próprios sistemas de informação, como atrasos no registro de eventos ou a ocorrência de perdas administrativas ao longo do acompanhamento (Colosimo; Giolo, 2024).

Para complementar a avaliação global, foram elaborados histogramas específicos para os pacientes que apresentaram interrupção e para aqueles classificados como censurados, permitindo uma análise diferenciada dos perfis temporais. Observa-se que os eventos de interrupção concentram-se predominantemente nos primeiros anos de seguimento, sugerindo maior vulnerabilidade à descontinuação no início do tratamento. Em contrapartida, os indivíduos censurados apresentaram maior concentração de tempos longos, o que pode estar associado à permanência prolongada em TARV ou ao encerramento administrativo do acompanhamento. Por fim, destaca-se que, em todos os subgrupos analisados, o tempo máximo de seguimento registrado foi de 24 anos, valor que corresponde ao período compreendido entre o início da consolidação dos registros no PIMC (ano 2000) e o ano-base da análise (2024), não configurando uma restrição técnica específica, mas sim uma característica estrutural decorrente da janela histórica de observação definida na base (Brasil. Ministério da Saúde, 2024e).

Tabela 4.11 – Distribuição dos tipos de evento observados na População Válida.

Status do Evento	*N*	Proporção
Censurado ($\delta_i = 0$)	714.659	65,3%
Interrupção ($\delta_i = 1$)	379.908	34,7%

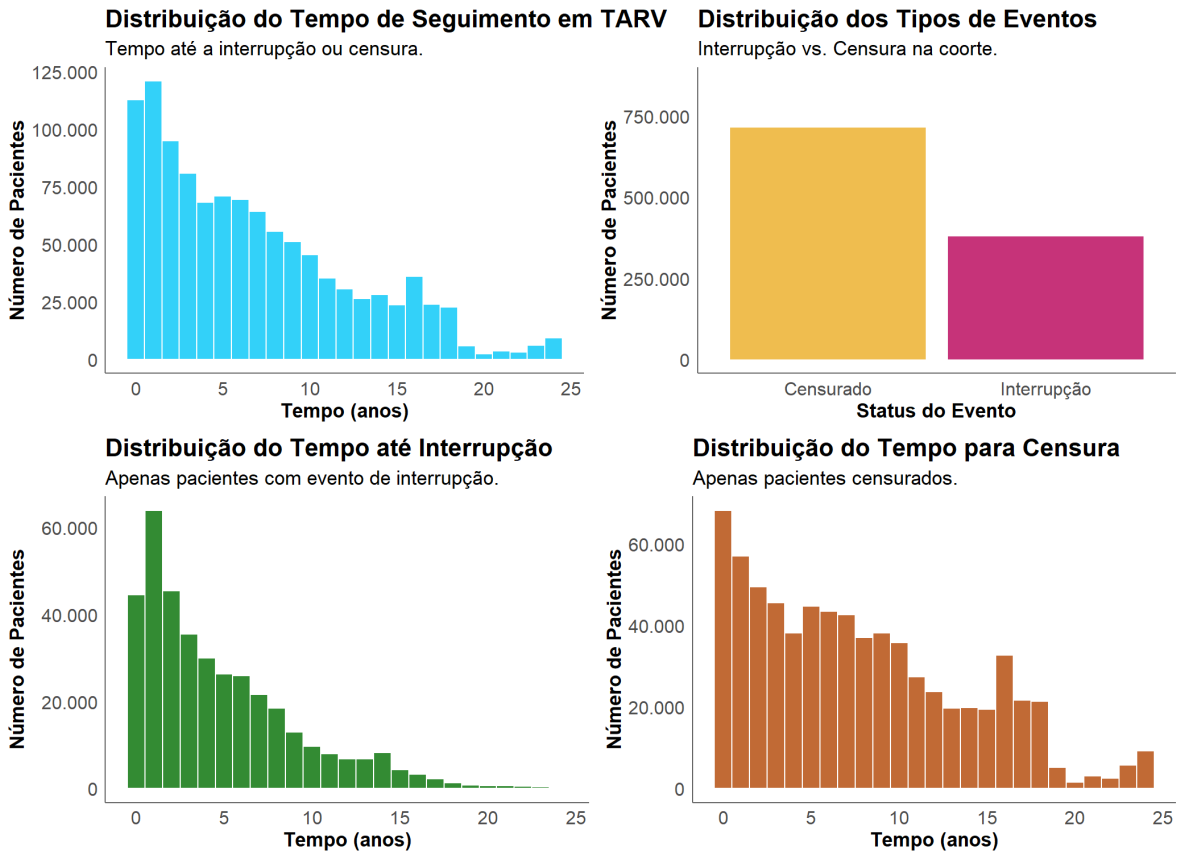

Figura 4.6 – Distribuição do tempo de seguimento em TARV (painel superior esquerdo), frequência dos tipos de evento observados (painel superior direito) e histogramas específicos para interrupção (inferior esquerdo) e censura (inferior direito), considerando a População Válida.

Fonte: Elaboração própria.

4.9 Interpretação Crítica dos Achados

A análise exploratória realizada ao longo deste capítulo permitiu uma compreensão detalhada das principais características sociodemográficas, clínicas e geográficas da coorte

analisada, bem como das limitações inerentes aos dados utilizados. Os resultados obtidos refletem não apenas o histórico da epidemia de HIV/Aids no Brasil, mas também as especificidades do processo de cuidado oferecido pelo Sistema Único de Saúde (SUS), evidenciando fatores estruturais e sociais que influenciam os desfechos clínicos observados.

Sob a perspectiva sociodemográfica, destaca-se o predomínio da doença entre homens adultos jovens, com baixos níveis de escolaridade e majoritariamente autodeclarados como *Pardos* ou *Brancos/Amarelos*. Tal perfil é consistente com o panorama epidemiológico descrito em investigações anteriores sobre a população vivendo com HIV atendida pelo SUS em diferentes regiões do país (Cunha *et al.*, 2025; Menezes *et al.*, 2018). Contudo, a análise estratificada por raça/cor e escolaridade revelou disparidades importantes no que se refere à permanência em TARV. Indivíduos autodeclarados como pretos e indígenas apresentaram os menores tempos medianos de seguimento, além de menores amplitudes interquartis, indicando maior vulnerabilidade a interrupções precoces no tratamento.

Esses achados dialogam com uma literatura nacional robusta, que destaca o papel das desigualdades sociais e raciais na trajetória de cuidado de pessoas vivendo com HIV/Aids no contexto brasileiro. Relatórios recentes do Ministério da Saúde apontam a raça/cor como um determinante social crítico, associado a barreiras institucionais, discriminação estrutural, desigualdades regionais na oferta de serviços especializados e dificuldades persistentes de acesso a recursos sociais fundamentais para a manutenção da adesão ao tratamento (Brasil. Ministério da Saúde, 2024e).

A escolaridade também foi considerada na análise. Embora a mediana de tempo de seguimento tenha sido ligeiramente superior entre indivíduos com menor nível educacional, a análise dos quartis e da dispersão revelou maior heterogeneidade nesse grupo. Tal padrão sugere a existência de subgrupos com perfis distintos de adesão e de permanência no tratamento. Pesquisas anteriores têm demonstrado uma associação consistente entre baixa escolaridade e maior risco de não adesão ou de interrupção da TARV, reforçando o papel da educação como um importante determinante social da saúde (Cunha *et al.*, 2025; Menezes *et al.*, 2018; Costa *et al.*, 2021).

A análise segundo faixa etária ao início da TARV reafirma o padrão epidemiológico brasileiro, caracterizado por predominância de adultos entre 25 e 39 anos. Entretanto, a maior vulnerabilidade observada entre adolescentes (13 a 17 anos), que apresentaram os menores tempos medianos de seguimento, merece atenção. Tal resultado é consistente com a literatura que aponta os desafios específicos enfrentados por adolescentes vivendo com HIV, incluindo conflitos familiares, dificuldades de aceitação do diagnóstico, menor

aderência ao tratamento e lacunas nos processos de transição entre serviços pediátricos e de adultos (Guerra; Seidl, 2010).

Do ponto de vista clínico, os dados indicam que uma parcela expressiva da população iniciou a TARV em estágios avançados de imunossupressão, com cerca de um quarto da coorte apresentando contagem de CD4 inferior a 200 células/mm^3 no início do tratamento. Esses achados reforçam preocupações já descritas na literatura sobre o diagnóstico e o início tardio da TARV no contexto brasileiro (Miranda *et al.*, 2022; Cunha *et al.*, 2025).

A elevada taxa de censura identificada (65,3%) é uma característica esperada em coortes longitudinais com longos períodos de acompanhamento, o que reforça a adequação da Análise de Sobrevivência como abordagem estatística. Contudo, a ausência de informações detalhadas sobre os motivos da censura (como óbito, perda de acompanhamento do paciente ou transferência de serviço) representa uma limitação importante, a ser considerada na interpretação dos modelos inferenciais subsequentes (Colosimo; Giolo, 2024).

A análise das desigualdades regionais também evidenciou diferenças significativas nas taxas de exclusão por inconsistências cronológicas e no tempo médio de seguimento entre as Unidades Federativas. Essas disparidades podem refletir desigualdades históricas na infraestrutura de saúde, na qualidade dos registros administrativos e nas políticas estaduais de vigilância e cuidado ao HIV/Aids (Cunha *et al.*, 2025; Brasil. Ministério da Saúde, 2022).

Por fim, destaca-se que o rigor metodológico empregado em todas as etapas iniciais de processamento dos dados, incluindo a imputação e a análise exploratória, contribuiu para a consistência e a robustez dos achados apresentados. As evidências aqui discutidas reforçam a necessidade de incluir variáveis sociodemográficas, clínicas e contextuais como fatores de ajuste nos modelos de regressão de sobrevivência, visando aprofundar a compreensão dos determinantes da interrupção da TARV no Brasil.

Capítulo 5

Resultados

5.1 Informações Preliminares

A etapa de modelagem estatística constitui o núcleo analítico dos estudos de tempo até evento. No presente trabalho, a análise de sobrevivência foi empregada para investigar os fatores associados ao tempo até a interrupção da terapia antirretroviral (TARV) entre pessoas vivendo com HIV/Aids no Brasil. Essa modelagem visa quantificar o efeito das covariáveis sociodemográficas e clínicas descritas no Capítulo 4, identificando como essas características influenciam a probabilidade de manutenção do tratamento ao longo dos anos.

Modelos clássicos, como o de riscos proporcionais de Cox (Cox, 1972), ou modelos paramétricos baseados em distribuições específicas (Weibull, exponencial, log-normal) (Kalbfleisch; Prentice, 2002), são amplamente utilizados na literatura. Entretanto, conforme discutido na Seção 4.8, a heterogeneidade temporal observada em variáveis como raça/cor, faixa etária e escolaridade sugere violação do pressuposto de proporcionalidade dos riscos, elemento central da formulação de Cox. Além disso, relações não lineares e interações complexas entre covariáveis limitam o poder explicativo de modelos lineares tradicionais.

Diante desse cenário, adotou-se o modelo *Nnet-Survival* (Gensheimer; Narasimhan, 2019), uma rede neural projetada especificamente para análise de sobrevivência em tempo discreto. Essa abordagem, detalhada no Capítulo 3, permite modelar diretamente o risco condicional por intervalo sem assumir relações lineares entre covariáveis e risco. No presente estudo, o horizonte de acompanhamento foi discretizado em $J = 25$ *hazards* discretos correspondentes aos tempos de evento observados $t \in \{0, 1, 2, \ldots, 24\}$ anos. A

rede neural estima o *hazard* discreto h_{it} para cada tempo $t = 0, 1, \ldots, 24$, representando a probabilidade condicional de interrupção da TARV no tempo t dado que o indivíduo permaneceu em tratamento até esse momento. O horizonte confiável para análise, definido por $\hat{G}(t) > 0{,}10$ segundo Graf *et al.* (1999), corresponde aos primeiros $K = 17$ anos.

A escolha do *Nnet-Survival* justifica-se pela dimensão e complexidade da coorte: mais de um milhão de indivíduos, múltiplas covariáveis categóricas e censura à direita predominante ($\delta_i = 0$ em 65,3% dos casos). Essa configuração inviabilizaria a aplicação eficiente de modelos de riscos proporcionais e reforça a adequação de uma arquitetura neural capaz de capturar dependências não lineares e interações de alta ordem. Além disso, a formulação discreta do modelo é compatível com a granularidade temporal dos registros administrativos do SUS, dispensando aproximações contínuas e garantindo estabilidade numérica.

A seguir, são apresentadas as especificidades de implementação, os parâmetros de treinamento, as métricas de desempenho e a análise interpretativa dos resultados, sempre referenciando as definições matemáticas previamente estabelecidas na Seção 3.3.

5.2 Formulação e Implementação do Modelo

Conforme formalizado no Capítulo 3, a modelagem de sobrevivência em tempo discreto baseia-se na estimação do risco condicional $h_{it} = P(T_i = t \mid T_i \geq t, \mathbf{x}_i)$ e na reconstrução da função de sobrevivência individual $\hat{S}_i(t) = \prod_{s=0}^{t-1}(1 - \hat{h}_{is})$, conforme a Subseção 3.2.3. Esses conceitos fundamentam a arquitetura do *Nnet-Survival* implementada neste estudo.

5.2.1 Função de Perda e Otimização

A verossimilhança discreta dos tempos de falha, conforme a formulação geral apresentada em (3.1), é maximizada numericamente por meio da minimização de sua forma negativa. Para evitar instabilidades numéricas causadas por logaritmos de valores nulos ou muito próximos de zero durante a retropropagação do erro (Seção 3.1), adiciona-se uma constante $\varepsilon = 10^{-7}$ aos argumentos dos logaritmos. A função de perda implementada é então:

$$\mathcal{L}(\Theta) = -\sum_{i=1}^{n} \left[\delta_i \log(h_{it_i}(\mathbf{x}_i; \Theta) + \varepsilon) + \sum_{s=0}^{t_i - \delta_i} \log(1 - h_{is}(\mathbf{x}_i; \Theta) + \varepsilon) \right], \tag{5.1}$$

onde Θ representa o conjunto de pesos e vieses da rede, e t_i é o tempo em que o indivíduo i experimentou o evento ou foi censurado, conforme definido na Subseção 3.2.3.

O algoritmo de otimização adotado foi o *Adam* (Kingma; Ba, 2017), que combina adaptação de passo por gradiente e controle de momento, proporcionando convergência estável mesmo em bases de grande escala. A taxa de aprendizado[1] inicial foi definida em $\eta = 10^{-3}$, controlando a velocidade de ajuste dos pesos: valores muito altos (acima de 10^{-1}) podem causar instabilidade e divergência, enquanto valores muito baixos (abaixo de 10^{-5}) tornam o treinamento excessivamente lento. O termo "inicial" refere-se ao fato de que o *Adam* ajusta adaptativamente essa taxa ao longo das iterações, baseando-se nas estimativas do primeiro e segundo momentos dos gradientes. Adicionalmente, empregou-se regularização L2 (*weight decay*) com parâmetro $\lambda = 10^{-5}$, que adiciona um termo de penalização $\lambda\|\Theta\|^2$ à função de perda para prevenir sobreajuste ao desencorajar pesos excessivamente grandes. Esse hiperparâmetro, assim como a taxa de *dropout* de 30% aplicada nas camadas ocultas (detalhada na Subseção 5.2.2), foi selecionado por meio de experimentos preliminares no conjunto de validação, testando valores em $\lambda \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ e selecionando aquele que minimizou a perda de validação.

5.2.2 Arquitetura da Rede Neural

A rede implementada segue uma topologia *feedforward* (conforme definida na Seção 3.1), com camada de entrada de dimensão $p = 19$, duas camadas ocultas densamente conectadas e uma camada de saída com $J = 25$ neurônios, correspondentes aos $J = 25$ *hazards* discretos para os tempos $t = 0, 1, \ldots, 24$.

Covariáveis Preditoras

As 19 covariáveis preditoras correspondem às variáveis sociodemográficas e clínicas descritas no Capítulo 4, após codificação *one-hot* (*one-hot encoding*). Essa codificação consiste em representar cada categoria de uma variável qualitativa como uma variável binária (0 ou 1), de modo que variáveis com c categorias são convertidas em $c - 1$ indicadores. A categoria omitida, denominada referência, corresponde ao perfil em que todos os indicadores daquela variável assumem valor zero. Por exemplo, no caso de raça/cor (4 categorias), a categoria Indígena é a referência: um indivíduo indígena é representado por (0, 0, 0) nos três indicadores de raça/cor, enquanto um indivíduo branco/amarelo é representado por (1, 0, 0). Essa escolha não altera os resultados

[1] A taxa de aprendizado (*learning rate*) é um hiperparâmetro que controla a magnitude das atualizações dos pesos a cada iteração do algoritmo de otimização. Valores típicos situam-se entre 10^{-4} e 10^{-2} (Goodfellow; Bengio; Courville, 2016).

preditivos do modelo, pois a rede neural pode reconstruir qualquer combinação de efeitos a partir das categorias codificadas.

1. **Sexo** (1 variável): indicador binário para sexo feminino (referência: masculino);
2. **Raça/Cor** (3 variáveis): indicadores para Branca/Amarela, Parda e Preta (referência: Indígena);
3. **Escolaridade** (2 variáveis): indicadores para 8 a 11 anos e 12 anos ou mais de estudo (referência: 0 a 7 anos);
4. **Faixa Etária ao Início da TARV** (5 variáveis): indicadores para menores de 13 anos, 13 a 17 anos, 18 a 24 anos, 25 a 39 anos e 40 a 59 anos (referência: 60 anos ou mais);
5. **CD4 Basal** (2 variáveis): indicadores para CD4 $<$ 200 células/mm^3 e CD4 ausente/não informado (referência: CD4 $\geq$ 200 células/mm^3);
6. **Carga Viral Basal** (2 variáveis): indicadores para CV detectável ($>$ 50 cópias/mL) e CV ausente/não informada (referência: CV indetectável $\leq$ 50 cópias/mL);
7. **Região Geográfica** (4 variáveis): indicadores para Centro-Oeste, Nordeste, Norte e Sudeste (referência: Sul).

Especificação da Arquitetura

A rede implementada possui duas camadas ocultas densamente conectadas, com 128 e 64 neurônios, respectivamente (a justificativa para essa escolha arquitetural é apresentada na Subseção 5.2.3). Formalmente, seja $a^{(0)} = \mathbf{x}_i \in \{0,1\}^{19}$ o vetor de entrada (covariáveis do indivíduo i). As ativações intermediárias são obtidas por

$$a^{(\ell)} = \varphi(W^{(\ell)} a^{(\ell-1)} + b^{(\ell)}), \quad \ell = 1, 2,$$

onde $\varphi(\cdot)$ denota a função ReLU (Subseção 3.1.1), $W^{(1)} \in \mathbb{R}^{128\times 19}$ e $W^{(2)} \in \mathbb{R}^{64\times 128}$ são as matrizes de pesos das camadas ocultas, e $b^{(1)} \in \mathbb{R}^{128}$ e $b^{(2)} \in \mathbb{R}^{64}$ são os vetores de vieses correspondentes.

As camadas ocultas utilizam ativação ReLU, introduzindo não linearidade necessária para modelar interações complexas. Aplicou-se regularização do tipo *dropout* (Srivastava *et al.*, 2014) com taxa de 30% nas camadas ocultas, desativando aleatoriamente 30% dos

neurônios em cada iteração do treinamento para promover representações mais robustas e generalizáveis. Durante a fase de teste, todos os neurônios são ativados, mas suas saídas são escaladas pelo fator de retenção (0,7) para compensar o maior número de unidades ativas. Detalhes adicionais sobre a regularização e a reprodutibilidade são apresentados na Subseção 5.2.4.

A camada de saída emprega ativação sigmoide (Subseção 3.1.1), gerando um vetor de $J = 25$ entradas correspondentes aos *hazards* discretos estimados para cada tempo:

$$\hat{\mathbf{h}}(\mathbf{x}_i) = \sigma(W^{(3)}a^{(2)} + b^{(3)}) \in (0,1)^{25},$$

onde $\sigma(\cdot)$ é a função sigmoide aplicada elemento a elemento, $W^{(3)} \in \mathbb{R}^{25\times 64}$ e $b^{(3)} \in \mathbb{R}^{25}$. Isso garante que os *hazards* discretos estimados $\hat{h}_{it}$ satisfaçam $0 < \hat{h}_{it} < 1$ para todo i e t, conforme a Subseção 3.2.3. As desigualdades são estritas pois a função sigmoide assume valores em $(0, 1)$, nunca atingindo exatamente 0 ou 1 para entradas finitas. Note que cada $\hat{h}_{it}$ representa uma probabilidade condicional de falha no tempo t dado que o indivíduo sobreviveu até o tempo anterior, de modo que não há restrição de que a soma das 25 estimativas seja menor ou igual a 1; a restrição probabilística relevante é que a função de sobrevivência $\hat{S}_i(t) = \prod_{s=0}^{t-1}(1 - \hat{h}_{is})$ permaneça em $[0, 1]$, o que é automaticamente satisfeito.

A Figura 5.1 apresenta uma representação esquemática da arquitetura completa do modelo, ilustrando o fluxo de informação desde as 19 covariáveis de entrada até os 25 *hazards* discretos de saída, evidenciando as conexões densas entre as camadas e a transformação progressiva dos dados através da rede (omitindo o índice i, para simplificar).

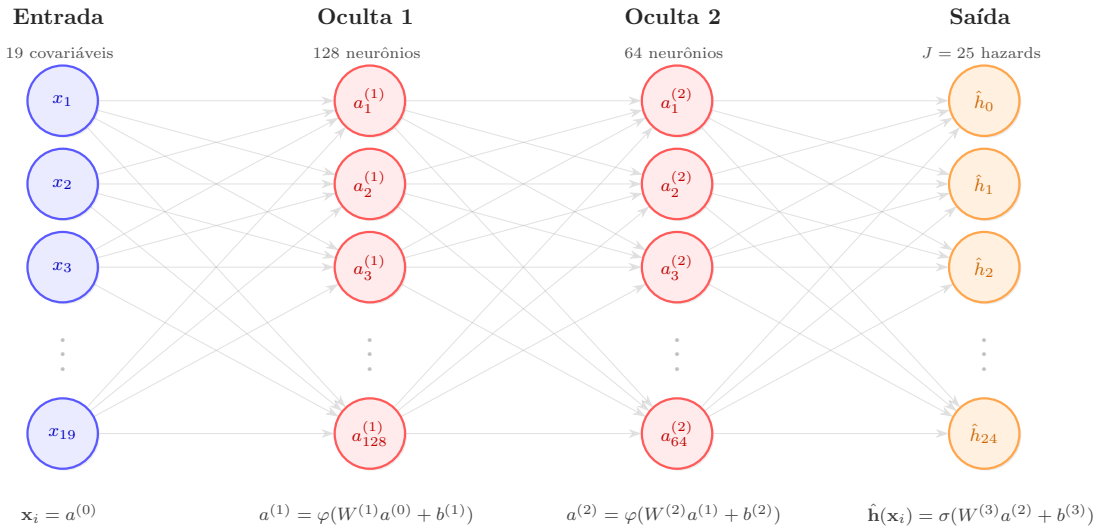

Figura 5.1 – Arquitetura da rede neural *Nnet-Survival* implementada. Embora a topologia seja a de uma rede *feedforward* densamente conectada convencional, sua adaptação à análise de sobrevivência reside na camada de saída e na função de perda: os $J = 25$ neurônios de saída, com ativação sigmoide, estimam diretamente os *hazards* discretos $\hat{h}_{it}$ para cada intervalo de tempo $t = 0, 1, \ldots, 24$, e o treinamento minimiza a log-verossimilhança negativa em tempo discreto (Equação 5.1), que incorpora naturalmente a censura. As duas camadas ocultas (128 e 64 neurônios) utilizam ativação ReLU e regularização por *dropout*[2].

Fonte: Elaboração própria.

O vetor de saída $\hat{\mathbf{h}}(\mathbf{x}_i) = (\hat{h}_{i0}, \hat{h}_{i1}, \ldots, \hat{h}_{i24})^\top$ contém os *hazards* discretos estimados, onde $\hat{h}_{it} = P(T_i = t \mid T_i \geq t, \mathbf{x}_i)$ representa a probabilidade condicional de ocorrência da interrupção do TARV no tempo t, conforme definido na Subseção 3.2.3. Note que $\hat{h}_{i0}$ modela o risco de interrupção antes de completar 1 ano de tratamento ($t = 0$), enquanto os demais *hazards* $\hat{h}_{it}$ para $t = 1, 2, \ldots, 24$ modelam o risco de interrupção em cada ano subsequente.

A função de sobrevivência individual é então reconstruída como:

$$\hat{S}_i(t) = \prod_{s=0}^{t-1} (1 - \hat{h}_{is}), \quad t = 1, 2, \ldots, 24,$$

representando a probabilidade estimada de o indivíduo i sobreviver (permanecer em TARV) até o tempo t anos, conforme definido na Subseção 3.2.3.

[2] O *dropout* não está representado visualmente no diagrama, mas é aplicado durante o treinamento entre as camadas ocultas.

5.2.3 Escolha do Número de Camadas e Neurônios

A arquitetura final com $L = 2$ camadas ocultas contendo 128 e 64 neurônios, respectivamente, foi determinada com base em três critérios:

1. **Fundamentação teórica:** O teorema da aproximação universal garante que redes neurais com uma única camada oculta de largura suficiente podem aproximar funções contínuas arbitrariamente bem (Hastie; Tibshirani; Friedman, 2009). Porém, arquiteturas mais profundas (múltiplas camadas) podem representar funções complexas de forma mais eficiente, com menos parâmetros, ao extrair hierarquias de representações abstratas dos dados (Goodfellow; Bengio; Courville, 2016). Para problemas de sobrevivência com $p = 19$ covariáveis, arquiteturas com 2–3 camadas ocultas são tipicamente suficientes (Katzman *et al.*, 2018; Kvamme; Borgan; Scheel, 2019).

2. **Validação empírica:** Experimentos preliminares no conjunto de validação compararam arquiteturas com diferentes profundidades ($L \in \{1, 2, 3\}$) e larguras (número de neurônios por camada variando entre 32 e 256). A arquitetura $L = 2$ com (128, 64) neurônios apresentou o melhor equilíbrio entre desempenho preditivo (menor perda de validação) e custo computacional, aparentando convergência sem evidências de sobreajuste. Arquiteturas mais profundas ($L = 3$) não resultaram em ganhos significativos de desempenho, enquanto arquiteturas mais rasas ($L = 1$) apresentaram capacidade representacional insuficiente.

3. **Parcimônia:** A regra prática de reduzir progressivamente o número de neurônios entre camadas consecutivas ($128 \to 64 \to 25$) cria um "gargalo" que força a rede a extrair representações compactas e informativas, facilitando a generalização e reduzindo o risco de memorização dos dados de treino (Goodfellow; Bengio; Courville, 2016). A arquitetura resultante possui 12.441 parâmetros treináveis[3]. Considerando que o conjunto de treinamento contém 700.522 observações, a razão entre o número de observações e o número de parâmetros é de aproximadamente 56:1, o que configura um cenário confortável do ponto de vista de sobreajuste e dispensa técnicas de regularização particularmente agressivas. Ainda assim, as estratégias de regularização adotadas (*dropout* e penalização L2) fornecem uma margem adicional de segurança.

[3] O número de parâmetros é calculado como: camada 1: $19 \times 128 + 128 = 2.560$; camada 2: $128 \times 64 + 64 = 8.256$; camada de saída: $64 \times 25 + 25 = 1.625$; total: $2.560 + 8.256 + 1.625 = 12.441$.

Essa escolha alinha-se com as práticas reportadas na literatura de modelos neurais para sobrevivência, onde arquiteturas compactas (2–3 camadas, 64–256 neurônios por camada) demonstram boa performance em bases de dados clínicos e epidemiológicos (Katzman *et al.*, 2018; Kvamme; Borgan; Scheel, 2019; Gensheimer; Narasimhan, 2019).

5.2.4 Regularização e Reprodutibilidade

Conforme mencionado na Subseção 5.2.2, aplicou-se regularização do tipo *dropout* (Srivastava *et al.*, 2014) com taxa de 30% nas camadas ocultas, forçando a rede a não depender excessivamente de neurônios específicos e promovendo a aprendizagem de representações mais robustas e generalizáveis.

Adicionalmente, empregou-se inicialização aleatória controlada por semente fixa (*seed*), assegurando a reprodutibilidade completa dos resultados. Os pesos iniciais foram amostrados de uma distribuição com variância calibrada apropriada para camadas com ativação ReLU (Goodfellow; Bengio; Courville, 2016), prevenindo o desvanecimento ou explosão de gradientes durante as primeiras iterações do treinamento. Os códigos utilizados para implementação do modelo estão disponíveis no Apêndice C.

5.3 Treinamento e Convergência do Modelo

O treinamento buscou minimizar a função de perda $\mathcal{L}(\Theta)$ definida na Equação (5.1), maximizando a verossimilhança dos tempos observados e censurados. Uma *epoch* corresponde a uma passagem completa do algoritmo de otimização por todo o conjunto de treinamento, atualizando os parâmetros da rede com base nos gradientes calculados em cada lote (*batch*) de observações. O procedimento foi monitorado por até 50 *epochs*, com *early stopping* (parada antecipada, conforme definido na Subseção 3.3.4) após 10 *epochs* consecutivas sem melhora na função perda considerando os dados de validação, e redução adaptativa da taxa de aprendizado (*ReduceLROnPlateau*), que diminui a taxa de aprendizado por um fator de 0,5 quando a métrica de validação fica estagnada por 5 *epochs*.

Conforme estabelecido na Subseção 3.3.4, os dados foram divididos aleatoriamente em 64% para treinamento (700.522 observações), 16% para validação (175.131 observações) e 20% para teste (218.914 observações), mantendo a proporção de eventos em cada conjunto por meio de amostragem estratificada. Cada *epoch* envolveu aproximadamente 343 lotes (*batches*) de 2.048 observações.

A Figura 5.2 apresenta a evolução da log-verossimilhança negativa ao longo das *epochs* para os conjuntos de treino e validação. As curvas permaneceram próximas (diferença de apenas 0,62%), sugerindo convergência estável e ausência de sobreajuste, reforçando a adequação das estratégias de regularização (L2, *dropout*, *early stopping*). Observa-se que a perda de treino é ligeiramente superior à de validação ao longo de todo o processo. Esse comportamento, embora contraintuitivo à primeira vista, é uma consequência direta da regularização por *dropout*: durante o treinamento, 30% dos neurônios são desativados aleatoriamente em cada iteração, de modo que a perda é calculada sobre uma rede operando com capacidade reduzida; na fase de validação, o *dropout* é desligado e todos os neurônios ficam ativos (com saídas escaladas pelo fator de retenção de 0,7), permitindo predições mais precisas e, consequentemente, perda menor (Srivastava *et al.*, 2014). Caso o *dropout* fosse removido, o padrão clássico de perda de treino inferior à de validação seria observado, porém com maior risco de sobreajuste.

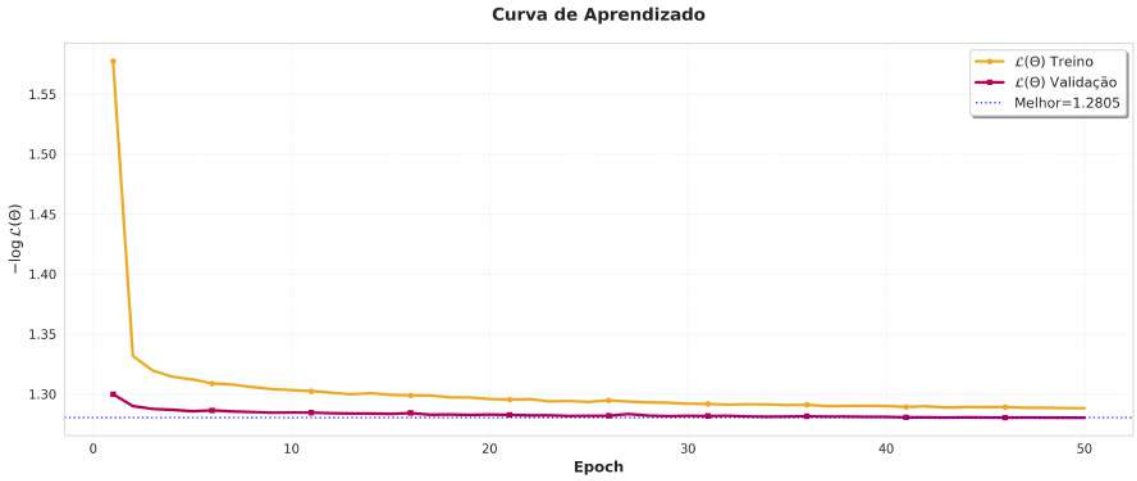

Figura 5.2 – Evolução da log-verossimilhança negativa durante o treinamento ao longo das *epochs*. A perda de treino situa-se acima da perda de validação devido ao efeito do *dropout*, que opera apenas durante o treinamento, reduzindo a capacidade efetiva da rede. A proximidade entre as curvas indica convergência estável e ausência de sobreajuste.

Fonte: Elaboração própria.

5.4 Avaliação de Desempenho

A avaliação do modelo considerou duas características complementares: discriminação, medida pelo *C-index* (Subseção 3.3.1), e calibração, avaliada pelo escore de Brier integrado (Subseção 3.3.2). As métricas foram calculadas no conjunto de teste. Antes de apresentar

os resultados, justificam-se as escolhas metodológicas específicas adotadas neste estudo.

5.4.1 Escolhas Metodológicas para Avaliação

Truncamento de Pesos IPCW

Conforme discutido na Subseção 3.3.2, em cenários de alta censura os pesos IPCW podem tornar-se instáveis. No presente estudo (65,3% de censura), adotou-se $w_{\max} = 5{,}0$ com base em três critérios:

1. **Análise da função $\hat{G}(t)$:** A função de sobrevivência da censura permanece acima de 0,20 até $t = 15$ anos, implicando pesos originais $w_i(t) \leq 5{,}0$ nesse horizonte. O truncamento só afeta tempos posteriores ($t > 15$ anos), preservando a informação no período em que a censura não é excessivamente pesada e a maioria dos eventos é observada.

2. **Fundamentação teórica:** Kvamme, Borgan e Scheel (2019) demonstram que o truncamento de pesos IPCW em um limite superior máximo previne instabilidade numérica quando $\hat{G}(t)$ se aproxima de zero, situação observada em horizontes temporais longos sob alta censura. O valor $w_{\max} = 5{,}0$ corresponde ao limiar $\hat{G}(t) \geq 0{,}20$, garantindo que apenas períodos com pelo menos 20% dos indivíduos livres de censura contribuam com pesos não truncados para o cálculo do escore de Brier.

3. **Análise de sensibilidade:** Testes exploratórios com $w_{\max} \in \{3, 5, 10\}$ indicaram variação mínima do IBS no horizonte confiável ($< 1\%$), confirmando robustez da escolha. No entanto, valores $w_{\max} \geq 10$ resultaram em variância substancialmente maior em horizontes longos ($t > 15$ anos), indicando que o truncamento em 5,0 representa um equilíbrio adequado entre estabilidade e preservação da informação para este cenário de 65,3% de censura.

Horizonte Confiável de Avaliação

No presente estudo, a mediana do seguimento é de 5 anos, porém adotou-se o critério operacional mais liberal $\hat{G}(t) > 0{,}10$ para definir o horizonte confiável, permitindo avaliação em períodos mais longos clinicamente relevantes para adesão à TARV. Este limiar implica que pelo menos 10% dos indivíduos permanecem livres de censura até o tempo t, garantindo que os pesos IPCW não excedam $w_i(t) = 1/0{,}10 = 10$ antes do

truncamento em $w_{\max} = 5{,}0$. Valores $\hat{G}(t) < 0{,}10$ resultariam em pesos originais superiores a 10, elevando substancialmente a variância das estimativas mesmo após truncamento e comprometendo a confiabilidade da avaliação (Kvamme; Borgan; Scheel, 2019).

O horizonte confiável foi delimitado em $K = 17$ anos, já que $\hat{G}(17) = 0{,}114 > 0{,}10$ e $\hat{G}(18) = 0{,}087 < 0{,}10$. Este período abrange a ampla maioria dos eventos observados e permite avaliar o desempenho do modelo em horizontes de longo prazo relevantes para o acompanhamento de pacientes em TARV. Resultados para $t > 17$ anos são reportados para completude, mas devem ser interpretados com cautela devido à instabilidade crescente dos pesos IPCW nesses horizontes.

O IBS foi calculado primariamente no horizonte confiável ($t \in \{1, 2, \ldots, 17\}$ anos), fornecendo uma avaliação robusta onde os pesos IPCW são estáveis. Adicionalmente, o IBS foi calculado para o período completo ($t \in \{1, 2, \ldots, 24\}$ anos) para fins de comparação, embora com menor confiabilidade devido à instabilidade dos pesos em horizontes longos.

Decisão sobre Recalibração

Optou-se por não aplicar recalibração pós-treinamento (conforme o método de *temperature scaling* apresentado na Subseção 3.3.3) pelas seguintes razões: (i) a função de perda baseada em verossimilhança de tempo discreto incentiva o modelo a predizer probabilidades bem calibradas (Gensheimer; Narasimhan, 2019), embora calibração perfeita não seja garantida; (ii) a regularização (*dropout* de 30% e L2 com $\lambda = 10^{-5}$) previne *overconfidence*[4], reduzindo o risco de descalibração sistemática; (iii) o IBS obtido (0,1559) indica calibração moderada, substancialmente abaixo do limiar aleatório (0,25) e dentro de faixas reportadas para modelos neurais de sobrevivência em aplicações clínicas similares; e (iv) a recalibração pós-treinamento é tipicamente recomendada quando há evidência clara de má calibração (IBS $> 0{,}20$), cenário não observado neste estudo.

5.4.2 Métricas Globais

A Tabela 5.1 resume os resultados das métricas de desempenho no conjunto de teste.

[4] *Overconfidence* (excesso de confiança) ocorre quando o modelo atribui probabilidades preditas excessivamente próximas de 0 ou 1, superestimando sua certeza sobre os desfechos individuais.

Tabela 5.1 – Métricas de desempenho do modelo *Nnet-Survival* no conjunto de teste.

Métrica	**Valor**	**Interpretação**	**Comentário**
C-index	0,7270	Boa discriminação	Ordenação correta em 72,70% dos pares
IBS (horizonte confiável)	0,1559	Calibração moderada	Redução de 37,6% vs. acaso
IBS (período completo)	0,1556	Calibração moderada	Inclui horizontes instáveis
Taxa de censura	65,3%	Alta	Típica de coortes longas
Epochs treinadas	50	–	Convergência estável

Fonte: Elaboração própria.

O C-index de 0,7270 demonstra boa capacidade discriminativa, ordenando corretamente 72,70% dos pares comparáveis de indivíduos. Esse valor está alinhado com resultados reportados por Katzman *et al.* (2018) e Kvamme, Borgan e Scheel (2019) em aplicações similares de redes neurais para análise de sobrevivência.

O IBS no horizonte confiável de 0,1559 indica calibração moderada, representando redução de 37,6% do erro em relação ao desempenho aleatório (IBS = 0,25). A diferença mínima entre o IBS no horizonte confiável (0,1559) e o IBS no período completo (0,1556) sugere que a inclusão de horizontes com $\hat{G}(t) < 0{,}10$ não alterou substancialmente a avaliação global, embora as análises principais sejam restritas ao horizonte confiável por razões de rigor metodológico.

A calibração moderada reflete o equilíbrio entre a função de perda baseada em verossimilhança, que naturalmente incentiva calibração, e a complexidade inerente ao problema de predição em horizontes longos com alta censura (65,3%). O valor de IBS obtido situa-se substancialmente abaixo do limiar de desempenho aleatório, indicando que o modelo captura adequadamente tanto a ordenação de riscos (discriminação) quanto a acurácia das probabilidades absolutas (calibração). A regularização por *dropout* (30%) e penalização L2 ($\lambda = 10^{-5}$) mostraram-se suficientes para prevenir *overconfidence* durante o treinamento, produzindo predições probabilísticas adequadas para uso em contextos de saúde pública sem necessidade de recalibração pós-treinamento.

5.4.3 Distribuição do Risco Predito

A Figura 5.3 apresenta a distribuição das probabilidades de risco acumulado preditas pelo modelo, estratificada pelo status de evento observado. O risco acumulado foi calculado como $1 - \hat{S}_i(24)$, ou seja, a probabilidade complementar de sobrevivência até o final do período de observação (24 anos). O painel esquerdo mostra histogramas sobrepostos, enquanto o painel direito apresenta boxplots comparativos.

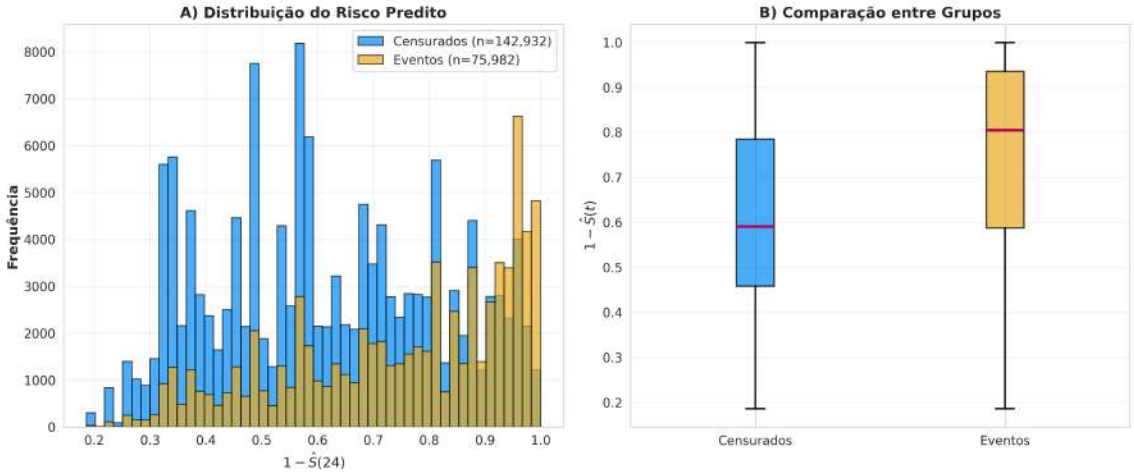

Figura 5.3 – Distribuição do risco acumulado predito segundo o status de evento. Indivíduos com interrupção da TARV apresentaram maiores riscos preditos, confirmando a capacidade discriminativa do modelo.

Fonte: Elaboração própria.

A clara separação entre as distribuições dos grupos censurados (mediana de risco = 0,5907) e os que experimentaram o evento (mediana de risco = 0,8051) demonstra que o modelo atribui sistematicamente maior risco acumulado aos indivíduos que interromperam a TARV, evidenciando coerência entre as predições e os desfechos observados. A diferença de 0,2144 entre as medianas confirma a capacidade discriminativa do modelo.

O histograma à esquerda evidencia que indivíduos censurados (que permaneceram em tratamento) concentram-se em valores mais baixos de risco acumulado, com moda aproximadamente em 0,55, enquanto indivíduos que interromperam a TARV apresentam distribuição deslocada para valores mais altos, com moda próxima a 1,0. Os boxplots à direita quantificam essa separação: a sobreposição limitada entre as distribuições, evidenciada pelos intervalos interquartis pouco sobrepostos, confirma que o modelo captura sistematicamente o padrão diferencial de risco entre os grupos. Nota-se, contudo, heterogeneidade intragrupo: alguns indivíduos censurados apresentam riscos elevados

(possivelmente aqueles com características adversas mas que, por razões não capturadas pelo modelo, permaneceram em tratamento), enquanto alguns eventos ocorreram em indivíduos com riscos preditos moderados, refletindo a natureza estocástica do fenômeno e a presença de fatores não observados que influenciam a interrupção.

5.4.4 Calibração ao Longo do Tempo

A Figura 5.4 apresenta a evolução do escore de Brier BS(t) e da função de sobrevivência da censura $\hat{G}(t)$ ao longo do tempo, para $t = 1, 2, \ldots, 24$. O painel superior mostra que BS(t) permanece relativamente estável entre os horizontes de 1 a 17 anos (horizonte confiável, onde $t \leq K = 17$), variando de 0,0872 (excelente calibração no curto prazo) a 0,1680 (calibração moderada no longo prazo), com mediana de 0,1574. O painel inferior ilustra o decaimento de $\hat{G}(t)$, que se mantém acima de 0,10 até $t = 17$ (linha laranja vertical), delimitando o horizonte confiável conforme operacionalização do critério de Graf *et al.* (1999) para evitar períodos com censura excessivamente pesada. Para $t > 17$, $\hat{G}(t)$ decai acentuadamente, resultando em pesos IPCW potencialmente instáveis ($w_i > 10$) mesmo após truncamento em $w_{\max} = 5{,}0$. O "N avaliável" indicado no painel inferior representa o número de indivíduos ainda sob risco (não censurados e sem evento) em cada tempo, ilustrando a redução progressiva da amostra disponível para avaliação.

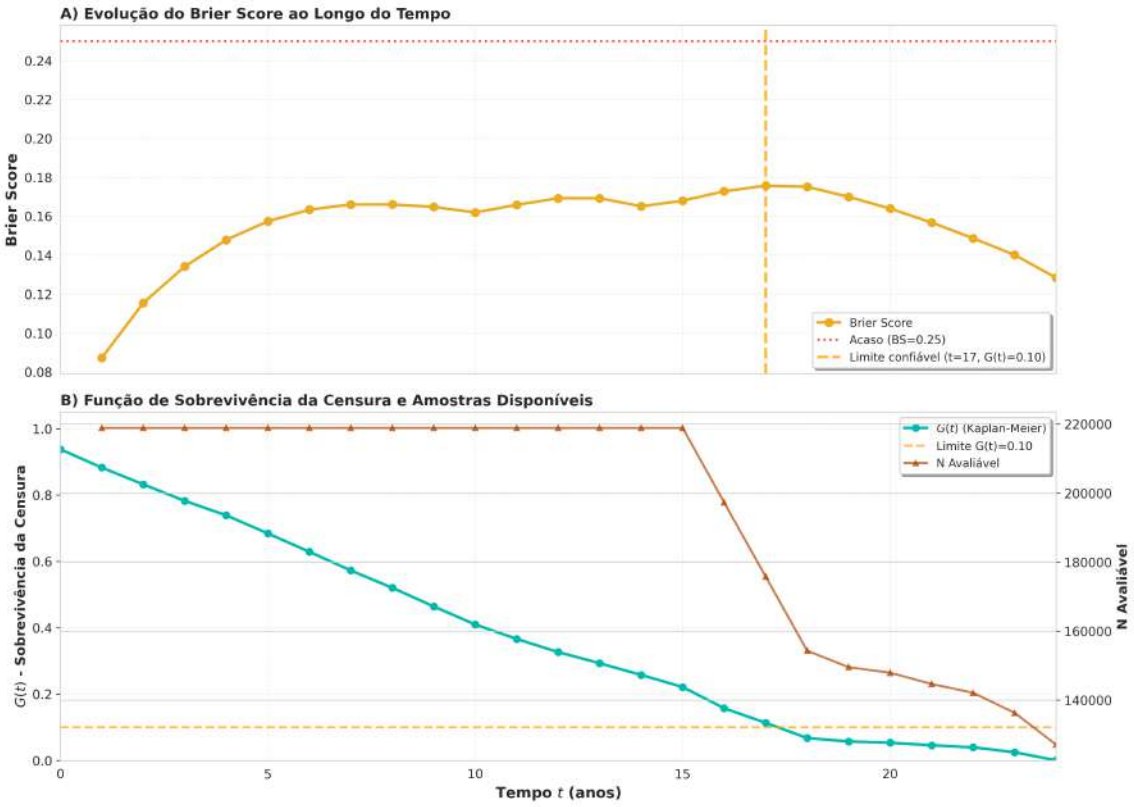

Figura 5.4 – Evolução do escore de Brier (painel superior) e da função $\hat{G}(t)$ de sobrevivência da censura (painel inferior) ao longo do tempo. A linha vertical laranja demarca o horizonte confiável ($t \leq K = 17$ anos) onde $\hat{G}(t) > 0{,}10$. O "N avaliável" representa o número de indivíduos ainda sob risco em cada tempo.

Fonte: Elaboração própria.

A análise conjunta dos dois painéis revela a relação entre qualidade de calibração e disponibilidade de dados. Nos primeiros anos ($t \in \{1, 2, 3, 4, 5\}$), o BS(t) é baixo (0,0872; 0,1342) e $\hat{G}(t)$ é alto (0,6843; 0,8828), indicando ampla disponibilidade de observações não censuradas que permitem estimativas confiáveis. À medida que o tempo avança, $\hat{G}(t)$ decai progressivamente, refletindo o acúmulo de censuras, e o BS(t) aumenta gradualmente, atingindo valores entre 0,1619 e 0,1680 no período de 10 a 17 anos. Esse padrão é esperado em coortes com alta censura: horizontes mais longos contêm menos informação empírica para validação das predições, aumentando a incerteza estatística.

Nota-se, entretanto, que mesmo no limite do horizonte confiável ($t = 17$), o BS(17) permanece abaixo de 0,17, substancialmente inferior ao limiar de desempenho aleatório (0,25). Após $t = 17$, observa-se um decaimento nos valores de BS(t), o que pode parecer contraintuitivo. Esse comportamento decorre da combinação de dois fatores: (i) a escassez

de eventos observados em horizontes longos reduz a variabilidade do escore; e (ii) os pesos IPCW truncados suavizam artificialmente as estimativas. Portanto, os valores de BS(t) para $t > 17$ não devem ser interpretados como evidência de melhor calibração, mas sim como artefato da instabilidade estatística, justificando a restrição das análises principais ao horizonte confiável. A linha vertical laranja demarca claramente essa transição, orientando a interpretação dos resultados: predições para $t \leq 17$ anos são estatisticamente confiáveis, enquanto predições para $t > 17$ anos devem ser interpretadas com cautela devido à escassez de dados de validação.

A decomposição do IBS por períodos temporais revela calibração consistente ao longo da trajetória de seguimento. Calculou-se o IBS$(1, 5)$ para o curto prazo (0–5 anos), obtendo-se 0,1284 (calibração boa); o IBS$(6, 10)$ para o médio prazo (6–10 anos), com valor de 0,1644 (calibração moderada); e o IBS$(11, 17)$ para o longo prazo (11–17 anos), resultando em 0,1694 (calibração moderada).

A deterioração gradual da calibração em horizontes mais longos é esperada e reflete a combinação de três fatores: (i) redução do tamanho efetivo da amostra sob risco (apenas 10–15% da amostra original permanece em $t > 15$ anos); (ii) aumento da incerteza estatística devido à escassez de eventos observados; e (iii) possível presença de efeitos tempo-variantes não capturados pelo modelo. Apesar dessa deterioração, os valores de BS(t) permanecem consistentemente abaixo de 0,20, indicando que o modelo mantém calibração adequada mesmo em horizontes longos.

5.4.5 Importância das Variáveis

A interpretabilidade do modelo foi aprimorada por análise de importância via *Permutation Importance* (Algoritmo 1), que quantifica a contribuição de cada covariável através da redução no C-index após permutação aleatória de seus valores (Breiman, 2001; Fisher; Rudin; Dominici, 2019). A Figura 5.5 apresenta todas as 19 covariáveis em ordem decrescente de impacto discriminatório.

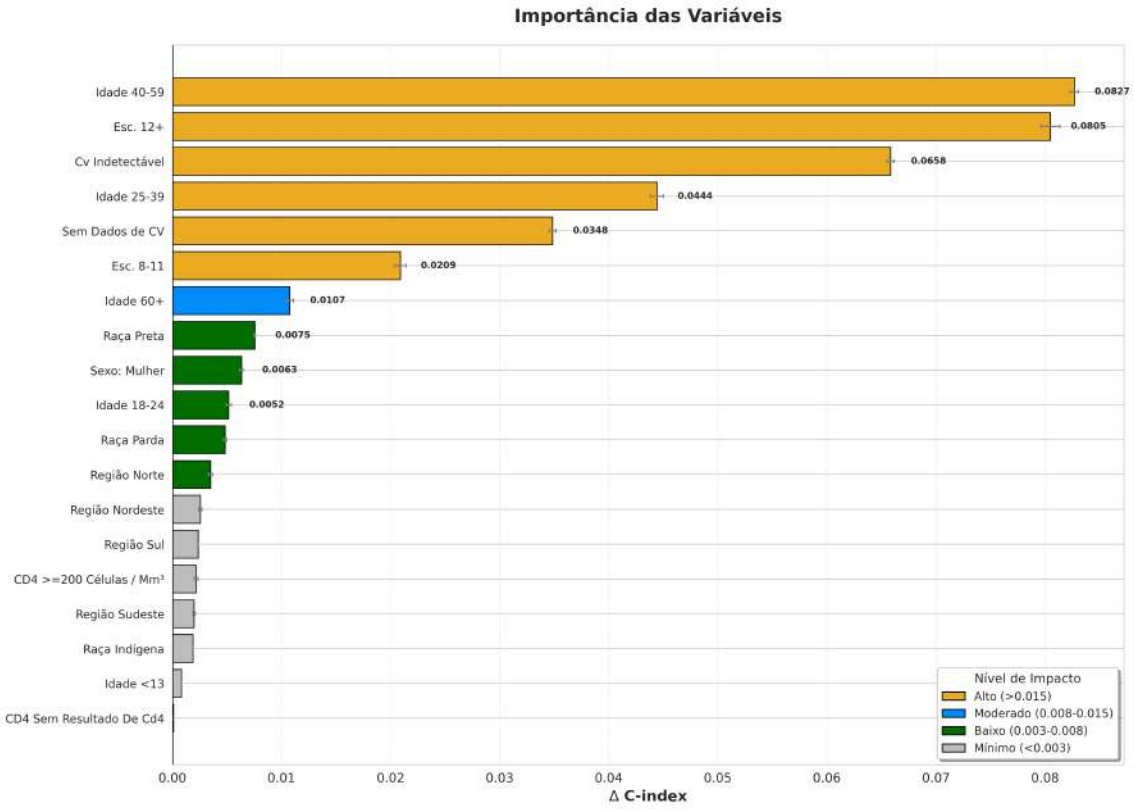

Figura 5.5 – Importância relativa das variáveis estimada pelo método de *Permutation Importance*. Todas as 19 covariáveis são exibidas em ordem decrescente de impacto no C-index.

Fonte: Elaboração própria.

A hierarquia de importância revela um padrão notável: a parte superior do *ranking* é dominada por variáveis associadas a faixa etária, nível de escolaridade e carga viral. A faixa etária de 40 a 59 anos (redução de 0,0827 no C-index) emerge como o preditor mais importante, seguida por escolaridade de 12 anos ou mais (0,0805) e carga viral indetectável (0,0658). Esse padrão evidencia que fatores sociais (especialmente idade e educação) exercem influência comparável ou até superior aos marcadores clínicos na predição da retenção em cuidado. A escolaridade elevada provavelmente reflete não apenas maior compreensão da importância da adesão, mas também melhores condições socioeconômicas, acesso facilitado aos serviços de saúde e menor exposição a barreiras estruturais. As faixas etárias de maior importância representam períodos da vida adulta com maior estabilidade psicossocial e percepção de risco, em contraste com adolescentes de 13 a 17 anos e jovens de 18 a 24 anos, que apresentam menor retenção conforme observado nas curvas estratificadas da Subseção 5.4.6.

Dentre os marcadores clínicos, a carga viral indetectável (0,0658) e a ausência de dados

de carga viral (0,0348) figuram entre as cinco variáveis de maior impacto. A importância da carga viral indetectável confirma que o controle virológico inicial é preditor consistente de melhor evolução no tratamento, sendo coerente com evidências da literatura sobre adesão à TARV (Costa *et al.*, 2021; Cunha *et al.*, 2025). Curiosamente, a ausência de dados de carga viral também emerge como preditora relevante, possivelmente sinalizando desigualdades no acesso a exames laboratoriais ou falhas na sistematização de registros, fatores que podem estar associados a piores desfechos e que merecem investigação aprofundada em estudos futuros.

Variáveis de raça/cor (preta: 0,0075; parda: 0,0053) apresentam importância moderada, sugerindo que seu efeito é parcialmente absorvido por escolaridade e outras covariáveis socioeconômicas que provavelmente apresentam correlação expressiva com a raça/cor, embora desigualdades raciais permaneçam evidentes nas curvas estratificadas. Região geográfica e sexo (mulher: 0,0063) apresentam importância mais baixa (0,0042; 0,0063), indicando que, embora relevantes, seu impacto é secundário quando ajustado por outros determinantes. Esse padrão reforça a natureza multidimensional da interrupção do TARV, na qual fatores sociais, demográficos e clínicos interagem de forma complexa para determinar a trajetória de cuidado.

5.4.6 Curvas de Sobrevivência Estratificadas

A Figura 5.6 apresenta as curvas de sobrevivência preditas $\hat{S}_i(t)$ estratificadas por características sociodemográficas. Como o modelo estima *hazards* discretos h_{ik} para cada intervalo anual $k = 0, 1, \ldots, 24$, a função de sobrevivência $\hat{S}_i(t_k) = \prod_{s=0}^{k-1}(1 - \hat{h}_{is})$ é definida apenas nos tempos discretos t_k; os segmentos de reta que conectam os pontos consecutivos nas curvas são um recurso gráfico de visualização, não uma interpolação do modelo. As diferenças entre grupos refletem desigualdades estruturais já observadas na análise exploratória (Capítulo 4) e confirmam os achados da análise de importância (Subseção 5.4.5).

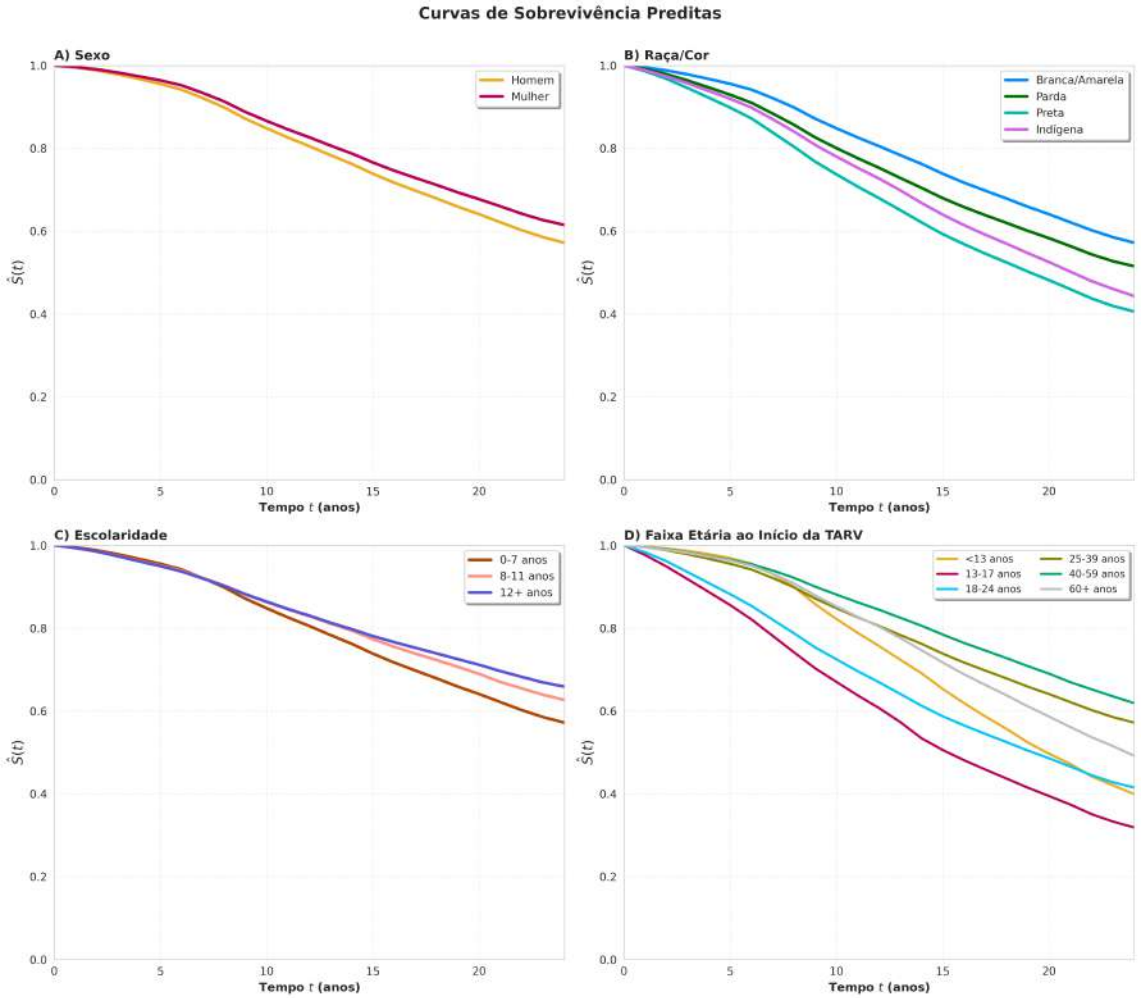

Figura 5.6 – Curvas de sobrevivência preditas estratificadas por características sociodemográficas: (A) sexo; (B) raça/cor; (C) escolaridade; (D) faixa etária ao início da TARV.

Fonte: Elaboração própria.

Disparidades de gênero (Painel A da Figura 5.6). Mulheres apresentam probabilidade de permanência no TARV consistentemente superior à dos homens ao longo de todo o horizonte temporal, com diferença crescente a partir do 5º ano. Aos 24 anos de seguimento, aproximadamente 60% das mulheres permanecem em tratamento, contra cerca de 55% dos homens. Essa disparidade reflete vulnerabilidades estruturais e barreiras específicas enfrentadas por homens no acesso e na continuidade do cuidado.

Desigualdades raciais (Painel B da Figura 5.6). O painel evidencia profundas disparidades: pessoas pretas apresentam as menores taxas de retenção (aproximadamente 45% aos 24 anos), seguidas por pessoas indígenas (cerca de 50%) e pardas (aproximadamente 55%). Pessoas brancas/amarelas mantêm-se mais tempo em TARV, com cerca de 60%

de permanência no mesmo horizonte. Essas diferenças refletem o racismo estrutural no sistema de saúde e desigualdades socioeconômicas historicamente construídas (Cunha; Cruz; Pedroso, 2022; Brasil. Ministério da Saúde, 2024f).

Gradiente educacional (Painel C da Figura 5.6). A estratificação por escolaridade revela padrão dose-resposta: indivíduos com 12 anos ou mais de estudo apresentam probabilidade de permanência superior (aproximadamente 65% aos 24 anos), seguidos por aqueles com 8 a 11 anos (cerca de 60%) e com 0 a 7 anos de estudo (aproximadamente 55%). Maior escolaridade associa-se a melhor compreensão da importância da adesão, maior autonomia financeira e acesso facilitado aos serviços de saúde (Miranda *et al.*, 2022; Melo *et al.*, 2019).

Efeito da idade (Painel D da Figura 5.6). A faixa etária ao início da TARV emerge como determinante crítico da retenção. Adolescentes de 13 a 17 anos e jovens de 18 a 24 anos apresentam as menores taxas de permanência por 24 anos (aproximadamente 30 a 40%). Crianças menores de 13 anos mantêm retenção intermediária (cerca de 40%), com uma queda acentuada começando após 8 anos de tratamento, provavelmente associada à entrada desses indivíduos na fase adulta e consequente menor interferência dos pais na continuidade do tratamento. Adultos de 25 a 39 anos alcançam aproximadamente 55% de permanência. Os grupos com melhor desempenho são adultos de 40 a 59 anos, com aproximadamente 65% de permanência aos 24 anos, seguidos por idosos de 60 anos ou mais, com cerca de 60%. Esse gradiente etário sugere que maturidade psicossocial e percepção de risco influenciam positivamente a adesão (Guerra; Seidl, 2010; Brasil. Ministério da Saúde, 2006).

5.5 Discussão dos Resultados

Os resultados confirmam que o *Nnet-Survival* alcançou desempenho satisfatório na modelagem do tempo até interrupção da TARV, conciliando boa discriminação (C-index = 0,7270) e calibração moderada (IBS = 0,1559 no horizonte confiável). As curvas preditas reproduzem as desigualdades sociodemográficas e clínicas identificadas empiricamente no Capítulo 4, validando a consistência interna do modelo.

A predominância de interrupções entre homens, pessoas pretas e indígenas, indivíduos de menor escolaridade e faixas etárias mais jovens evidencia a persistência de desigualdades estruturais no cuidado, em consonância com a literatura sobre adesão e vulnerabilidade social (Costa *et al.*, 2021; Cunha *et al.*, 2025). A análise de importância das variáveis

(Subseção 5.4.5) revelou que determinantes sociais (especialmente idade e escolaridade) exercem influência comparável ou superior aos marcadores clínicos (CD4, carga viral), reforçando a natureza multidimensional dos fatores que influenciam a retenção em cuidado.

5.5.1 Viés de Sobrevivência Saudável

Um ponto importante a se observar é a possível presença do viés de sobrevivência saudável (*healthy survivor bias*) (Buckley *et al.*, 2015). Esse viés ocorre quando indivíduos que permanecem mais tempo em acompanhamento diferem sistematicamente daqueles que interrompem precocemente, não apenas em características clínicas, mas também em aspectos comportamentais e socioeconômicos não mensurados. No presente estudo, a alta taxa de censura (65,3%) indica que a maioria dos indivíduos não experimentou o evento de interesse (interrupção da TARV) durante o período de observação, seja porque permaneceram em tratamento até o final do acompanhamento, seja porque foram perdidos de seguimento por outras razões (óbito, transferência, perda de vínculo com o serviço). Os indivíduos que permanecem mais tempo sob acompanhamento potencialmente representam um subgrupo com melhor aderência e condições socioeconômicas mais favoráveis. Ainda assim, as comparações relativas entre subgrupos permanecem válidas, pois o padrão diferencial é consistente ao longo do tempo e reflete desigualdades estruturais bem documentadas na literatura.

Capítulo 6

Considerações Finais

A compreensão dos fatores que influenciam a continuidade da terapia antirretroviral (TARV) constitui um elemento central para o enfrentamento da epidemia de HIV/Aids no Brasil. O presente estudo teve como propósito principal investigar o tempo até a interrupção do tratamento antirretroviral, aplicando uma abordagem metodológica inovadora que combina os fundamentos clássicos da análise de sobrevivência com a flexibilidade das redes neurais artificiais. O objetivo foi, portanto, não apenas quantificar a influência de variáveis sociodemográficas e clínicas sobre a manutenção da TARV, mas também demonstrar o potencial analítico das técnicas de aprendizado profundo na modelagem de fenômenos complexos em saúde pública, e que pode inclusive ser aplicado em várias outras áreas de pesquisa.

A utilização do modelo *Nnet-Survival*, fundamentado em tempo discreto e compatível com a presença de censura à direita, mostrou-se particularmente adequada à natureza dos dados analisados. Conforme discutido no Capítulo 5, o modelo apresentou bom desempenho preditivo e calibração satisfatória, refletindo robustez e capacidade de generalização mesmo diante de uma base de dados extensa e heterogênea. O índice de concordância (*C-index*) de 0,7270 e o escore de Brier integrado (IBS) de 0,1559 no horizonte confiável demonstraram que a rede neural foi capaz de ordenar adequadamente os riscos individuais e reconstruir funções de sobrevivência coerentes com os padrões epidemiológicos observados. A calibração moderada obtida (IBS substancialmente abaixo do limiar de desempenho aleatório de 0,25, representando redução de 37,6% do erro) atesta a adequação das estratégias de regularização empregadas. Esses resultados validam empiricamente o uso de abordagens não lineares na modelagem de sobrevivência e evidenciam o amadurecimento dessas técnicas no campo das ciências biomédicas, com potencial de ser expandido para

vários outros campos, como o das ciências atuariais.

A análise dos resultados confirmou a persistência de desigualdades estruturais na continuidade terapêutica, evidenciando que vulnerabilidades sociais e contextuais ainda exercem influência decisiva sobre a adesão e permanência no tratamento. Homens, pessoas pretas, indígenas, indivíduos com menor escolaridade e faixas etárias mais jovens apresentaram maior probabilidade de interrupção, mesmo após o controle de fatores clínicos. Esses achados corroboram os resultados empíricos apresentados no Capítulo 4 e reiteram que a adesão terapêutica é condicionada por determinantes sociais que transcendem o nível individual, refletindo desigualdades históricas de acesso e permanência no cuidado (Costa *et al.*, 2021; Cunha *et al.*, 2025). Tais evidências reforçam a necessidade de políticas públicas orientadas pelos princípios de equidade e integralidade do SUS, capazes de articular intervenções clínicas, psicossociais e estruturais voltadas ao fortalecimento da adesão e da retenção em cuidado.

Do ponto de vista clínico, o modelo confirmou a relevância da condição imunológica inicial na trajetória terapêutica. Pacientes com contagem reduzida de linfócitos CD4 (< 200 células/mm^3) e carga viral detectável no início do tratamento (> 50 cópias/mL) apresentaram maior risco estimado de interrupção, sugerindo que o diagnóstico precoce e o início imediato da TARV permanecem determinantes essenciais para a efetividade do cuidado. Esses resultados reiteram que a detecção tardia do HIV compromete o sucesso terapêutico, pois indivíduos que ingressam tardiamente nos serviços de saúde tendem a exibir quadros clínicos mais graves e maior vulnerabilidade a descontinuidade. Assim, políticas que integrem testagem ampliada, início precoce do tratamento e monitoramento sistemático são fundamentais para reduzir a mortalidade e o abandono.

Um aspecto metodológico e interpretativo relevante refere-se à possível presença do *viés de sobrevivência saudável* (*healthy survivor bias*), analisado na Seção 5.5.1. Esse viés ocorre quando os indivíduos que permanecem mais tempo sob acompanhamento clínico diferem sistematicamente dos que interrompem precocemente, tanto em condições clínicas quanto em características comportamentais e socioeconômicas. A alta taxa de censura observada (65,3%) indica que a maioria dos indivíduos não experimentou o evento de interesse (interrupção da TARV) durante o período de observação, seja porque permaneceram em tratamento até o final do acompanhamento, seja porque foram perdidos de seguimento por outras razões (óbito, transferência, perda de vínculo com o serviço). Os indivíduos que permanecem mais tempo sob acompanhamento potencialmente representam um subgrupo com melhor aderência e condições socioeconômicas mais favoráveis. Embora essa distorção

não invalide as comparações relativas entre subgrupos, ela ressalta a importância de interpretações cautelosas e do uso de modelos que considerem a dinâmica temporal da adesão. Estudos futuros podem incorporar variáveis longitudinais de monitoramento clínico (como evolução de CD4 e carga viral) e medidas objetivas de adesão farmacológica, mitigando esse viés e refinando as estimativas de sobrevivência.

Sob a perspectiva metodológica, a aplicação do *Nnet-Survival* reforça a viabilidade do uso de redes neurais artificiais na análise de sobrevivência, campo ainda incipiente na literatura nacional. A abordagem empregada demonstrou ser possível aliar o rigor estatístico da modelagem de tempo até a ocorrência do evento de interesse, fundamentada em verossimilhanças de tempo discreto (Equação 3.1), à flexibilidade e à expressividade das redes neurais profundas. A implementação de funções de perda derivadas dessa formulação, associada a técnicas de regularização (L2 com $\lambda = 10^{-5}$, *dropout* de 30%) e otimização adaptativa (*Adam* com taxa de aprendizado inicial de 10^{-3}), garantiram estabilidade numérica e evitaram sobreajuste mesmo em bases de grande porte. A convergência estável em 50 *epochs*, com diferença de apenas 0,62% entre os pontos de máximo da log-verossimilhança negativa quando avaliada com os dados de treino e validação, confirma a ausência de sobreajuste. Essa combinação entre consistência estatística e eficiência computacional confirma o potencial das arquiteturas neurais como ferramentas complementares às abordagens paramétricas clássicas em estudos epidemiológicos e atuariais.

A principal contribuição deste trabalho reside, portanto, na demonstração empírica de que modelos de aprendizado profundo podem ser empregados de forma eficaz na modelagem de fenômenos de sobrevivência com dados administrativos de larga escala. Além de fornecer estimativas preditivas robustas, a abordagem adotada permitiu identificar interações não lineares entre determinantes clínicos e sociais, capturando padrões de risco que escapam à linearidade dos modelos tradicionais. Em um contexto de forte heterogeneidade populacional, tal capacidade de generalização constitui avanço relevante para a pesquisa quantitativa em saúde pública.

Sob o ponto de vista das políticas públicas, os resultados obtidos oferecem subsídios para o aprimoramento das estratégias de adesão e retenção em cuidado. A estimação individual de probabilidades de permanência em TARV por k anos possibilita a identificação de grupos com maior risco de descontinuidade, favorecendo o direcionamento de recursos e o desenvolvimento de intervenções focalizadas. Ao incorporar modelos preditivos às rotinas de vigilância e monitoramento clínico, o sistema de saúde pode adotar práticas

mais dinâmicas, baseadas em evidências, e promover respostas antecipadas a sinais de abandono do tratamento. Tais aplicações alinham-se às metas de eliminação do HIV/Aids como problema de saúde pública, em consonância com as diretrizes do Ministério da Saúde e da UNAIDS (Brasil. Ministério da Saúde, 2024a; Joint United Nations Programme on HIV/AIDS (UNAIDS), 2025).

Por outro lado, é imprescindível reconhecer as limitações do estudo. O uso de dados secundários, ainda que provenientes de bases oficiais de ampla cobertura, impõe restrições quanto à completude e qualidade de certas variáveis. A ausência de dimensões comportamentais e psicossociais, como apoio familiar, estigma e uso de substâncias ilícitas, reduz a capacidade do modelo de capturar aspectos subjetivos da adesão e permanência no tratamento. Além disso, a natureza observacional da base impede inferências causais diretas, restringindo as conclusões ao campo associativo. Apesar dessas limitações, a amplitude amostral (1.094.567 indivíduos), o rigor analítico e a coerência dos resultados conferem solidez às inferências realizadas.

Os resultados obtidos neste estudo abrem caminhos para diversas extensões metodológicas e aplicadas que podem aprimorar a compreensão dos determinantes da interrupção terapêutica e ampliar a utilidade dos modelos preditivos desenvolvidos. Do ponto de vista metodológico, a incorporação de arquiteturas de redes neurais mais sofisticadas representa uma direção promissora. Redes neurais recorrentes (RNNs), incluindo variantes como LSTM (*Long Short-Term Memory*) e GRU (*Gated Recurrent Unit*), ou *transformers* poderiam explorar a natureza sequencial dos dados longitudinais, capturando padrões temporais de adesão, trocas de esquemas terapêuticos e episódios de perda de seguimento que antecedem a interrupção definitiva. Mecanismos de atenção temporal permitiriam identificar janelas críticas no acompanhamento que exercem maior influência sobre o risco futuro, oferecendo insights para intervenções preventivas oportunas.

A adoção de abordagens bayesianas constitui outra extensão metodológica relevante. A formulação de modelos de sobrevivência em arcabouço bayesiano, utilizando métodos de inferência como MCMC (*Markov Chain Monte Carlo*) ou técnicas de aproximação variacional, permitiria quantificar adequadamente a incerteza nas estimativas de risco individualizadas, produzindo intervalos de credibilidade para as probabilidades de sobrevivência preditas (Ibrahim; Chen; Sinha, 2001). Algoritmos MCMC modernos como o NUTS (*No-U-Turn Sampler*) (Hoffman; Gelman, 2011), implementado em frameworks como Stan (Carpenter *et al.*, 2017) e PyMC (Salvatier; Wiecki; Fonnesbeck, 2016), oferecem convergência eficiente mesmo em modelos complexos de alta dimensionalidade,

eliminando a necessidade de ajuste manual de hiperparâmetros de amostragem através de adaptação automática do tamanho de passo e da matriz de massa. Para bases de dados de larga escala como a analisada neste estudo, abordagens de inferência aproximada como a inferência variacional automática (*Automatic Differentiation Variational Inference*, ADVI) (Kucukelbir *et al.*, 2017) ou métodos de subamostragem estocástica poderiam ser empregados para tornar a estimação bayesiana computacionalmente tratável, ainda que com alguma perda de precisão em relação à inferência MCMC exata. Além disso, prioris baseadas em conhecimento epidemiológico prévio poderiam ser incorporadas para estabilizar estimativas em subgrupos com poucos eventos observados, aumentando a robustez das predições em populações vulneráveis. A inferência bayesiana também facilitaria a atualização dinâmica do modelo à medida que novos dados se tornam disponíveis, permitindo um sistema preditivo adaptativo e continuamente calibrado.

Uma extensão particularmente promissora seria a implementação de modelos *Bayesian Additive Regression Trees* (BART) para análise de sobrevivência (Sparapani *et al.*, 2016). BART é um método não paramétrico que modela a função de regressão como uma soma de múltiplas árvores de decisão, cada uma contribuindo incrementalmente para a predição final. Essa abordagem aditiva permite capturar não linearidades e interações complexas entre covariáveis sem especificação prévia, ao mesmo tempo que a formulação bayesiana produz automaticamente distribuições *a posteriori* completas para as probabilidades de sobrevivência. Diferentemente das redes neurais, que requerem técnicas adicionais como *dropout* Monte Carlo para quantificação de incerteza, BART oferece intervalos de credibilidade naturalmente através da inferência MCMC (Chipman; George; McCulloch, 2010; Hill; Linero; Murray, 2020). Além disso, BART possui robustez à escolha de hiperparâmetros, tipicamente funcionando bem com especificações padrão, o que o torna mais acessível para aplicações práticas em saúde pública.

No contexto específico de análise de sobrevivência com censura, Sparapani *et al.* (2016) adaptaram BART para modelar *hazards* discretos através de um arcabouço de tempo discreto similar ao utilizado por redes neurais de sobrevivência. A estrutura hierárquica de árvores facilita a identificação de interações complexas entre covariáveis e permite detectar automaticamente violações de *hazards* proporcionais, incluindo funções de sobrevivência cruzadas, sem necessidade de especificação manual de termos de interação (Tan; Roy, 2019). Estudos de simulação mostraram desempenho preditivo comparável ou superior a *Random Forests* e redes neurais, com a vantagem de produzir intervalos de credibilidade calibrados para as predições (Hill; Linero; Murray, 2020).

Desenvolvimentos recentes ampliaram substancialmente o escopo de BART. He, Yalov e Hahn (2019) introduziram XBART (*Accelerated BART*), uma implementação computacionalmente eficiente que utiliza algoritmos de escalada estocástica para estimação posterior rápida, reduzindo o tempo computacional em várias ordens de magnitude, aspecto crucial para bases de dados de grande escala como a deste estudo. Uma inovação particularmente relevante é o GS-BART (*Graph-Split BART*) proposto por He, Sang e Zhou (2025), que estende BART para incorporar estruturas de grafos nas regras de decisão. Enquanto BART clássico utiliza regras de divisão eixo-paralelas (por exemplo, `idade > 40`), GS-BART permite partições mais flexíveis que respeitam relações complexas entre covariáveis, como proximidade geográfica ou redes sociais. No contexto da adesão à TARV, isso poderia capturar padrões espaciais de vulnerabilidade (por exemplo, clínicas próximas compartilhando barreiras de acesso) ou redes de apoio social que influenciam a permanência no tratamento. Outras extensões incluem BART para riscos competitivos (Sparapani *et al.*, 2020a), eventos recorrentes (Sparapani *et al.*, 2020b) e modelos de tempo de falha acelerado bayesianos (Henderson *et al.*, 2020), demonstrando a versatilidade da abordagem.

No contexto da adesão à TARV, BART seria especialmente útil para descoberta automática de subgrupos de risco com combinações complexas de características sociodemográficas e clínicas. Por exemplo, poderia revelar que indivíduos com baixa escolaridade, na faixa etária de 18 a 24 anos e residentes na região Norte apresentam risco substancialmente elevado, sem necessidade de especificação prévia dessas interações triplas. GS-BART poderia adicionalmente identificar *clusters* geográficos de alta interrupção, por exemplo, pacientes em municípios adjacentes com infraestrutura de saúde precária, que não seriam detectados por métodos que tratam localização como variável contínua independente. Esta capacidade de identificação automática de perfis de vulnerabilidade multidimensionais seria valiosa para informar estratégias de intervenção personalizadas e direcionar recursos de forma mais eficiente. Os intervalos de credibilidade produzidos por BART forneceriam, ainda, quantificação de incerteza robusta para essas predições de risco, aspecto crucial para decisões clínicas onde a confiança nas estimativas é tão importante quanto as próprias predições pontuais.

No campo da interpretabilidade, métodos complementares como SHAP values (*SHapley Additive exPlanations*) e LIME (*Local Interpretable Model-agnostic Explanations*) poderiam ser empregados para explicar predições individuais de forma mais granular, identificando quais características específicas contribuem mais fortemente para o risco

estimado de cada paciente (Lundberg; Lee, 2017). Essa camada adicional de interpretabilidade é fundamental para a aceitação clínica dos modelos de aprendizado de máquina e para a tradução de predições em recomendações acionáveis no contexto do cuidado. Quanto à integração de novas fontes de dados, a incorporação de variáveis comportamentais e psicossociais, tais como indicadores de saúde mental, histórico de uso de substâncias, percepção de estigma, rede de apoio social e identidade de gênero, poderia capturar dimensões subjetivas da adesão que não estão representadas nos registros administrativos. Parcerias com serviços de atenção primária e programas de acompanhamento psicossocial permitiriam enriquecer a base de dados com essas variáveis, potencialmente aumentando o poder preditivo e a relevância clínica dos modelos. Adicionalmente, a integração de dados geoespaciais detalhados (distância até serviços de saúde, índices de vulnerabilidade social dos bairros) poderia elucidar barreiras estruturais de acesso ao tratamento.

Do ponto de vista de análise causal, a aplicação de métodos de inferência causal a partir de dados observacionais, como modelos de equações estruturais, análise de mediação ou técnicas de aprendizado de máquina causal (*causal machine learning*), poderia avançar além das associações identificadas neste estudo, investigando relações causais entre intervenções específicas e a permanência no tratamento (Pearl, 2009). Modelos de *g-computation* ou *targeted maximum likelihood estimation* (TMLE) permitiriam estimar efeitos causais de políticas de saúde, como a simplificação de esquemas terapêuticos ou a descentralização do cuidado, sobre a retenção em TARV. Finalmente, a validação externa do modelo em outras coortes brasileiras ou latino-americanas, bem como sua implementação prospectiva em ambientes clínicos reais, são etapas essenciais para avaliar a generalização e a utilidade prática das predições. Estudos de implementação poderiam avaliar se alertas preditivos baseados no modelo, integrados a sistemas de informação clínica, de fato auxiliam profissionais de saúde a identificar pacientes em risco e a intervir de forma personalizada e oportuna.

Em síntese, o presente trabalho evidencia que a modelagem de sobrevivência mediada por redes neurais artificiais constitui uma alternativa promissora e metodologicamente sólida para o estudo de desfechos em saúde pública. O modelo *Nnet-Survival* mostrou-se capaz de representar de forma precisa as probabilidades de interrupção da TARV e de refletir a complexidade das interações entre variáveis clínicas e sociais. Ao integrar a tradição estatística da análise de sobrevivência, formalizada no Capítulo 2, às inovações do aprendizado profundo, o estudo amplia as fronteiras metodológicas da pesquisa epidemiológica e contribui para o fortalecimento das abordagens quantitativas voltadas à

promoção da equidade e à melhoria das políticas de saúde no Brasil. As direções futuras aqui delineadas apontam para um programa de pesquisa robusto que pode, progressivamente, aproximar modelos preditivos sofisticados da prática clínica e das decisões de política pública, maximizando seu impacto sobre a saúde populacional.

Apêndice A

Conversão de Base: CSV para Parquet (R)

Este apêndice apresenta o script em linguagem R utilizado para converter os arquivos CSV originais em formato *Parquet*, garantindo maior eficiência de leitura, compressão e manipulação durante as etapas de análise, conforme descrito na Seção 4.3.

```r
#------------------------------------------------------------------------------
# Carregamento dos pacotes
#------------------------------------------------------------------------------
library(data.table)
library(arrow)
library(janitor)
```

```r
# --- Arquivo primário ---
message("Iniciando a leitura do arquivo 'Banco_PIMC_2024_prim.csv'...")
prim_dt <- fread("dados/Banco_PIMC_2024_prim.csv", encoding = "Latin-1")
message("Leitura do arquivo primário concluída.")
prim_dt <- clean_names(prim_dt)
message("Escrevendo arquivo Parquet para 'prim'...")
write_parquet(prim_dt, "dados/pimc_prim.parquet")
message("Arquivo 'pimc_prim.parquet' criado com sucesso!")
rm(prim_dt)
```

```r
# --- Arquivo secundário ---
message("Iniciando a leitura do arquivo 'Banco_PIMC_2024_ult.csv'...")
ult_dt <- fread("dados/Banco_PIMC_2024_ult.csv", encoding = "Latin-1")
```

```r
message("Leitura concluída.")
ult_dt <- clean_names(ult_dt)
message("Escrevendo arquivo Parquet para 'ult'...")
write_parquet(ult_dt, "dados/pimc_ult.parquet")
message("Arquivo 'pimc_ult.parquet' criado com sucesso!")
rm(ult_dt)
```

```r
gc() # Força a limpeza da memória RAM
```

Apêndice B

Código do Capítulo 4 (R)

Este apêndice apresenta o código completo em linguagem R utilizado na preparação dos dados e na análise exploratória do Capítulo 4. Os blocos a seguir correspondem aos *chunks* do documento original, implementados com as bibliotecas `tidyverse`, `lubridate`, `arrow` e `ggplot2`.

```r
#------------------------------------------------------------------------------
# Bloco de Configuração Inicial (Setup)
#------------------------------------------------------------------------------
Sys.setlocale("LC_ALL", "pt_BR.UTF-8")
knitr::opts_chunk$set(message = FALSE, warning = FALSE, echo = FALSE)

# Carregamento dos pacotes
library(tidyverse)
library(lubridate)
library(scico)
library(geobr)
library(knitr)
library(sf)
library(patchwork)
library(missForest)
library(arrow)
library(scales)
library(colorspace)
library(glue)
library(stringr)
library(forcats)

# Definições de paletas e temas
```

```r
cores <- c("#ebac23", "#b80058", "#008cf9", "#006e00", "#00bbad", "#d163e6", "#b24502", "#ff9287", "#5954d6", "#00c6f8", "#878500", "#00a76c", "#bdbdbd")

# Criar uma versão escurecida das cores para beirada de boxplots
cores_escurecidas <- darken(cores, amount = 0.4)

paleta_continua <- scale_fill_scico(palette = "imola", direction = -1, name = NULL, na.value = "gray90")

theme_clean <- theme_minimal() +
  theme(
    panel.grid = element_blank(),
    axis.line = element_line(color = "gray40"),
    plot.title = element_text(hjust = 0, face = "bold", size = 16),
    plot.subtitle = element_text(size = 13, hjust = 0),
    plot.caption = element_text(hjust = 0.5, size = 9),
    axis.title = element_text(face = "bold", size = 13),
    axis.text = element_text(size = 12),
    legend.title = element_text(face = "bold", size = 10),
    legend.text = element_text(size = 10)
  )
```

```r
# Dicionário de UFs com correção de inconsistências do geobr
ufs_dict <- geobr::read_state(code_state = "all", year = 2020) %>%
  sf::st_drop_geometry() %>%
  mutate(name_state = case_when(
    name_state == "Amazônas" ~ "Amazonas",
    name_state == "Rio Grande Do Sul" ~ "Rio Grande do Sul",
    name_state == "Mato Grosso Do Sul" ~ "Mato Grosso do Sul",
    name_state == "Rio Grande Do Norte" ~ "Rio Grande do Norte",
    name_state == "Rio De Janeiro" ~ "Rio de Janeiro",
    name_state == "Paraiba" ~ "Paraíba",
    TRUE ~ name_state
  )) %>%
  select(name_state, abbrev_state)

# Função global para padronização de UF
padronizar_uf <- function(df, coluna) {
  df %>%
    mutate(
      !!sym(coluna) := case_when(
        str_detect(!!sym(coluna), regex("^Acre$", TRUE)) ~ "Acre",
        str_detect(!!sym(coluna), regex("^Alagoas$", TRUE)) ~ "Alagoas",
        str_detect(!!sym(coluna), regex("^Amapá$|^Amapa$", TRUE)) ~ "Amapá",
```

```
        str_detect(!!sym(coluna), regex("^Amazonas$", TRUE)) ~ "Amazonas",
        str_detect(!!sym(coluna), regex("^Bahia$", TRUE)) ~ "Bahia",
        str_detect(!!sym(coluna), regex("^Ceará$|^Ceara$", TRUE)) ~ "Ceará",
        str_detect(!!sym(coluna), regex("^Distrito Federal$", TRUE)) ~ "Distrito
        ↪  Federal",
        str_detect(!!sym(coluna), regex("^Espírito Santo$|^Espirito Santo$", TRUE)) ~
        ↪  "Espírito Santo",
        str_detect(!!sym(coluna), regex("^Goiás$|^Goias$", TRUE)) ~ "Goiás",
        str_detect(!!sym(coluna), regex("^Maranhão$|^Maranhao$", TRUE)) ~ "Maranhão",
        str_detect(!!sym(coluna), regex("^Mato Grosso do Sul$", TRUE)) ~ "Mato Grosso
        ↪  do Sul",
        str_detect(!!sym(coluna), regex("^Mato Grosso$", TRUE)) ~ "Mato Grosso",
        str_detect(!!sym(coluna), regex("^Minas Gerais$", TRUE)) ~ "Minas Gerais",
        str_detect(!!sym(coluna), regex("^Pará$", TRUE)) ~ "Pará",
        str_detect(!!sym(coluna), regex("^Paraíba$|^Paraiba$", TRUE)) ~ "Paraíba",
        str_detect(!!sym(coluna), regex("^Paraná$|^Parana$", TRUE)) ~ "Paraná",
        str_detect(!!sym(coluna), regex("^Pernambuco$", TRUE)) ~ "Pernambuco",
        str_detect(!!sym(coluna), regex("^Piauí$|^Piaui$", TRUE)) ~ "Piauí",
        str_detect(!!sym(coluna), regex("^Rio de Janeiro$", TRUE)) ~ "Rio de Janeiro",
        str_detect(!!sym(coluna), regex("^Rio Grande do Norte$", TRUE)) ~ "Rio Grande
        ↪  do Norte",
        str_detect(!!sym(coluna), regex("^Rio Grande do Sul$", TRUE)) ~ "Rio Grande do
        ↪  Sul",
        str_detect(!!sym(coluna), regex("^Rondônia$|^Rondonia$", TRUE)) ~ "Rondônia",
        str_detect(!!sym(coluna), regex("^Roraima$", TRUE)) ~ "Roraima",
        str_detect(!!sym(coluna), regex("^Santa Catarina$", TRUE)) ~ "Santa Catarina",
        str_detect(!!sym(coluna), regex("^São Paulo$|^Sao Paulo$", TRUE)) ~ "São
        ↪  Paulo",
        str_detect(!!sym(coluna), regex("^Sergipe$", TRUE)) ~ "Sergipe",
        str_detect(!!sym(coluna), regex("^Tocantins$", TRUE)) ~ "Tocantins",
        TRUE ~ NA_character_
      )
    ) %>%
    filter(!is.na(!!sym(coluna)) & !!sym(coluna) != "")
}
```

```
#------------------------------------------------------------------------------
# Bloco de Processamento de Dados
#------------------------------------------------------------------------------

# --- ETAPA 1: Leitura dos Dados ---
prim_ds <- open_dataset("dados/pimc_prim.parquet")
ult_ds  <- open_dataset("dados/pimc_ult.parquet")

```

```r
# --- ETAPA 2: Construção da População Válida ---

# Resumo da ULT (primeiro evento de interrupção + último ano observado)
ult_summary_ds <- ult_ds %>%
  group_by(cod_unificado) %>%
  summarise(
    first_event_year   = min(if_else(status_ano == "Interrupção de Tarv", ano,
    ↪  NA_integer_), na.rm = TRUE),
    last_observed_year = max(ano, na.rm = TRUE)
  )

# Join PRIM + ULT
prim_ult_join <- prim_ds %>%
  left_join(ult_summary_ds, by = "cod_unificado") %>%
  mutate(ano_inicio_tarv = year(as_date(data_dispensa))) %>%
  filter(!is.na(ano_inicio_tarv), !is.na(last_observed_year)) %>%
  collect()

# --- Contagem do Total Inicial ---
total_inicial <- nrow(prim_ult_join)
print(glue::glue("Total inicial antes da exclusão de t<0: {total_inicial}"))

# --- Aplicação do cálculo de time_to_event e exclusão de t < 0 ---
populacao_valida_ds <- prim_ult_join %>%
  mutate(
    time_to_event = if_else(
      is.finite(first_event_year),
      first_event_year - ano_inicio_tarv,
      last_observed_year - ano_inicio_tarv
    )
  ) %>%
  filter(time_to_event >= 0)

# --- Criação do dataframe final com as UFs padronizadas ---
df_populacao_valida <- populacao_valida_ds %>%
  collect() %>%
  padronizar_uf("uf_inst_completo") %>%
  mutate(
    across(c(cd4_cat200, cd4_cat350, raca_cat2, escol_cat, sexo_cat), ~ na_if(., "")),
    event_status = if_else(is.finite(first_event_year), "Interrupção", "Censurado")
  )

# --- ETAPA 3: Variável de Carga Viral (CV) ---
valid_codes <- df_populacao_valida$cod_unificado

```

```r
cv_data <- ult_ds %>%
  filter(cod_unificado %in% valid_codes) %>%
  select(cod_unificado, ano, deteccao_cv50) %>%
  collect() %>%
  arrange(cod_unificado, ano) %>%
  group_by(cod_unificado) %>%
  summarise(deteccao_cv50 = first(na.omit(deteccao_cv50), default = NA_character_))

df_longitudinal_para_imputar <- df_populacao_valida %>%
  left_join(cv_data, by = "cod_unificado") %>%
  mutate(across(c(sexo_cat, raca_cat2, escol_cat, idade_vinc_cat, cd4_cat200,
    ↪  cd4_cat350, deteccao_cv50), ~ na_if(., "Não Informado")))

df_longitudinal_pre_imputacao <<- df_longitudinal_para_imputar

# --- ETAPA 4: Imputação com missForest ---
df_for_missforest <- df_longitudinal_para_imputar %>%
  select(sexo_cat, raca_cat2, escol_cat, idade_vinc_cat, cd4_cat200, cd4_cat350,
         time_to_event, event_status, deteccao_cv50, uf_inst_completo) %>%
  mutate(across(where(is.character), as.factor))

set.seed(123)
imputacao_rf <- missForest(as.data.frame(df_for_missforest), maxiter = 2, ntree = 20)
df_imputed <- imputacao_rf$ximp

# --- ETAPA 5: Reconstrução Final ---
df_ids <- df_longitudinal_para_imputar %>% select(cod_unificado)

df_longitudinal_final <- bind_cols(df_ids, df_imputed) %>%
  rename(
    sexo = sexo_cat, raca = raca_cat2, escolaridade = escol_cat,
    faixa_etaria_inicio = idade_vinc_cat
  ) %>%
  mutate(
    escolaridade = fct_recode(escolaridade,
                              "0-7" = "0 a 7 anos",
                              "8-11" = "8 a 11 anos",
                              "12+" = "12 e mais anos"),
    escolaridade = fct_relevel(escolaridade, "0-7", "8-11", "12+"),
    faixa_etaria_inicio = fct_recode(faixa_etaria_inicio,
                                     "<13" = "Menos de 13 anos",
                                     "13-17" = "13 a 17 anos",
                                     "18-24" = "18 a 24 anos",
                                     "25-39" = "25 a 39 anos",
                                     "40-59" = "40 a 59 anos",
```

```
                                      "60+"  = "Mais de 60 anos"),
    faixa_etaria_inicio = fct_relevel(faixa_etaria_inicio, "<13", "13-17", "18-24",
    ↪  "25-39", "40-59", "60+")
  )

df_longitudinal <<- df_longitudinal_final

save(df_longitudinal, file = "df_longitudinal.RData")
```

```
load("df_longitudinal.RData")
```

```
#-------------------------------------------------------------------------------
# Bloco: Resumo das Proporções de Dados Ausentes - Antes e Depois da Imputação
#-------------------------------------------------------------------------------

# Calcula % de dados ausentes antes da imputação
pre_missing <- df_longitudinal_pre_imputacao %>%
  summarise(
    `Sexo de nascimento` = mean(is.na(sexo_cat)) * 100,
    `Raça/cor declarada` = mean(is.na(raca_cat2)) * 100,
    `Escolaridade` = mean(is.na(escol_cat)) * 100,
    `Faixa etária ao início do tratamento` = mean(is.na(idade_vinc_cat)) * 100,
    `CD4 inicial` = mean(is.na(cd4_cat200)) * 100,
    `Carga viral (< 50 cópias/mL)` = mean(is.na(deteccao_cv50)) * 100
  ) %>%
  pivot_longer(everything(), names_to = "Variável", values_to = "Percentual de Dados
  ↪  Ausentes (Pré-Imputação)")

# Calcula % de dados ausentes depois da imputação
post_missing <- df_longitudinal %>%
  summarise(
    `Sexo de nascimento` = mean(is.na(sexo)) * 100,
    `Raça/cor declarada` = mean(is.na(raca)) * 100,
    `Escolaridade` = mean(is.na(escolaridade)) * 100,
    `Faixa etária ao início do tratamento` = mean(is.na(faixa_etaria_inicio)) * 100,
    `CD4 inicial` = mean(is.na(cd4_cat200)) * 100,
    `Carga viral (< 50 cópias/mL)` = mean(is.na(deteccao_cv50)) * 100
  ) %>%
  pivot_longer(everything(), names_to = "Variável", values_to = "Percentual de Dados
  ↪  Ausentes (Pós-Imputação)")

# Junta as duas
```

```r
missing_summary_table <- pre_missing %>%
  left_join(post_missing, by = "Variável") %>%
  mutate(across(starts_with("Percentual"), ~ scales::percent(.x / 100, accuracy = 0.1)))

# Exibe a tabela
knitr::kable(
  missing_summary_table,
  caption = "Resumo das proporções de dados ausentes antes e após a imputação via missForest."
)
```

```r
# --- Lista de UFs de Referência (27 estados + DF) ---
ufs_ref <- ufs_dict %>%
  distinct(name_state) %>%
  rename(uf_inst_completo = name_state)

# --- Contagem Antes da Limpeza ---
contagem_antes <- prim_ds %>%
  collect() %>%
  padronizar_uf("uf_inst_completo") %>%
  count(uf_inst_completo, name = "N_Inicial")

# --- Contagem Depois da Limpeza ---
contagem_depois <- df_populacao_valida %>%
  count(uf_inst_completo, name = "N_Final_Valido")

# --- Geração da Tabela de Atrito ---
tabela_atricao <- ufs_ref %>%
  left_join(contagem_antes, by = "uf_inst_completo") %>%
  left_join(contagem_depois, by = "uf_inst_completo") %>%
  mutate(
    N_Inicial = replace_na(N_Inicial, 0),
    N_Final_Valido = replace_na(N_Final_Valido, 0),
    N_Excluidos = N_Inicial - N_Final_Valido,
    Proporcao_Excluida = if_else(N_Inicial > 0, N_Excluidos / N_Inicial, NA_real_)
  )

# --- Adiciona Linha Total ---
linha_total <- tabela_atricao %>%
  summarise(
    uf_inst_completo = "Total",
    N_Inicial = sum(N_Inicial, na.rm = TRUE),
    N_Final_Valido = sum(N_Final_Valido, na.rm = TRUE),
```

```
    N_Excluidos = sum(N_Excluidos, na.rm = TRUE),
    Proporcao_Excluida = sum(N_Excluidos, na.rm = TRUE) / sum(N_Inicial, na.rm = TRUE)
  )

tabela_atricao_final <- bind_rows(tabela_atricao, linha_total) %>%
  mutate(Proporcao_Excluida = scales::percent(Proporcao_Excluida, accuracy = 0.1))

# --- Impressão da Tabela ---
knitr::kable(
  tabela_atricao_final,
  col.names = c("UF", "N Inicial", "N Válido Final", "N Excluídos", "% Excluídos"),
  caption = "Número de indivíduos por UF antes e depois da limpeza de dados para elegibilidade na coorte."
)
```

```
#------------------------------------------------------------------------------
# Bloco: Visualização da Estrutura dos Dados - Antes e Depois da Imputação
#------------------------------------------------------------------------------

cat("\n\n================ ESTRUTURA DOS DADOS PRÉ-IMPUTAÇÃO ================\n\n")
glimpse(df_longitudinal_pre_imputacao)

cat("\n\n================ ESTRUTURA DOS DADOS PÓS-IMPUTAÇÃO ================\n\n")
glimpse(df_longitudinal)
```

```
#------------------------------------------------------------------------------
# Bloco: Análise Exploratória - Variáveis Sociodemográficas
#------------------------------------------------------------------------------

# --- Distribuição por Sexo ---
p_sexo <- df_longitudinal %>%
  ggplot(aes(x = fct_infreq(sexo), fill = sexo)) +
  geom_bar(alpha = 0.9) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark = ",")) +
  labs(
    title = "Distribuição das Variáveis Sociodemográficas da Coorte",
    subtitle = "Sexo",
    x = "Sexo de Nascimento",
    y = "Número de Pacientes"
  ) +
```

```r
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Distribuição por Raça/Cor ---
p_raca <- df_longitudinal %>%
  ggplot(aes(x = fct_infreq(raca), fill = raca)) +
  geom_bar(alpha = 0.9) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    subtitle = "Raça/Cor",
    x = "Raça/Cor Declarada",
    y = NULL
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Distribuição por Escolaridade ---
p_escolaridade <- df_longitudinal %>%
  ggplot(aes(x = escolaridade, fill = escolaridade)) +
  geom_bar(alpha = 0.9) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    subtitle = "Escolaridade",
    x = "Nível de Escolaridade (anos)",
    y = "Número de Pacientes"
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Distribuição por Faixa Etária no Início da TARV ---
p_faixa_etaria <- df_longitudinal %>%
  ggplot(aes(x = faixa_etaria_inicio, fill = faixa_etaria_inicio)) +
```

```
  geom_bar(alpha = 0.9) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    subtitle = "Faixa Etária (Início da TARV)",
    x = "Faixa Etária (anos)",
    y = NULL
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Painel Final: 2x2 ---
(p_sexo | p_raca) / (p_escolaridade | p_faixa_etaria)
```

```
#------------------------------------------------------------------------------
# Bloco: Tabelas de Distribuição - Variáveis Sociodemográficas
#------------------------------------------------------------------------------

# Função auxiliar para geração de tabelas de frequência com proporção
generate_freq_table <- function(df, var, titulo) {
  df %>%
    count({{ var }}) %>%
    mutate(Proporcao = scales::percent(n / sum(n), accuracy = 0.01)) %>%
    knitr::kable(caption = titulo)
}

# --- Tabela: Sexo ---
generate_freq_table(df_longitudinal, sexo, "Distribuição por Sexo")

# --- Tabela: Raça/Cor ---
generate_freq_table(df_longitudinal, raca, "Distribuição por Raça/Cor")

# --- Tabela: Escolaridade ---
generate_freq_table(df_longitudinal, escolaridade, "Distribuição por Escolaridade")

# --- Tabela: Faixa Etária ---
generate_freq_table(df_longitudinal, faixa_etaria_inicio, "Distribuição por Faixa
↪  Etária no Início da TARV")
```

```r
#------------------------------------------------------------------------------
# Bloco: Pirâmide Etária por Sexo
#------------------------------------------------------------------------------

# Pré-processamento dos dados
age_pyramid_data_plot <- df_longitudinal %>%
  filter(!is.na(sexo)) %>%
  count(faixa_etaria_inicio, sexo, name = "n") %>%
  mutate(
    count_for_pyramid = if_else(sexo == "Homem", -n, n),
    label_pos = if_else(sexo == "Homem", -n - 50, n + 50)
  )

# Gráfico
ggplot(age_pyramid_data_plot, aes(x = faixa_etaria_inicio, y = count_for_pyramid, fill
↪  = sexo)) +
  geom_col(alpha = 0.8) +
  coord_flip() +
  labs(
    title = "Distribuição Etária da Coorte segundo o Sexo de nascimento",
    subtitle = "Pirâmide Etária da Coorte por Sexo",
    y = "Número de Pacientes",
    x = "Faixa Etária",
    fill = "Sexo de Nascimento"
  ) +
  scale_fill_manual(values = cores[1:2]) +
  scale_y_continuous(
    labels = function(x) scales::comma_format(big.mark = ".", decimal.mark =
    ↪  ",")(abs(x)),
    breaks = pretty(age_pyramid_data_plot$count_for_pyramid, n = 5)
  ) +
  theme_clean +
  theme(
    plot.title.position = "plot"
  )
```

```r
#------------------------------------------------------------------------------
# Bloco: Tabela de Distribuição por Faixa Etária e Sexo
#------------------------------------------------------------------------------

age_pyramid_data_table <- df_longitudinal %>%
  filter(!is.na(sexo)) %>%
  count(faixa_etaria_inicio, sexo, name = "n") %>%
```

```r
  group_by(faixa_etaria_inicio) %>%
  mutate(Proporcao_na_faixa = scales::percent(n / sum(n), accuracy = 0.1)) %>%
  ungroup()

knitr::kable(
  age_pyramid_data_table,
  caption = "Distribuição por Faixa Etária e Sexo"
)
```

```r
#-------------------------------------------------------------------------------
# Bloco: Distribuição das Variáveis Clínicas Iniciais
#-------------------------------------------------------------------------------

# --- Gráfico CD4 <= 200 ---
p_cd4_200 <- df_longitudinal %>%
  mutate(cd4_cat200_label = fct_recode(cd4_cat200,
    "<200" = "<200 células / mm³",
    "200" = ">=200 células / mm³",
    "Sem CD4" = "Sem resultado de CD4"
  )) %>%
  ggplot(aes(x = cd4_cat200_label, fill = cd4_cat200_label)) +
  geom_bar(alpha = 0.8) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    title = "Distribuição das Variáveis Clínicas Iniciais",
    subtitle = "Contagem de CD4 ( 200 células/mm³)",
    x = "CD4 (células/mm³)",
    y = "Número de Pacientes"
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot",
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )

# --- Gráfico CD4 <= 350 ---
p_cd4_350 <- df_longitudinal %>%
  mutate(cd4_cat350_label = fct_recode(cd4_cat350,
    "<350" = "<350 células / mm³",
    "350" = ">=350 células / mm³",
    "Sem CD4" = "Sem resultado de CD4"
```

```
  )) %>%
  ggplot(aes(x = cd4_cat350_label, fill = cd4_cat350_label)) +
  geom_bar(alpha = 0.8) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    subtitle = "Contagem de CD4 ( 350 células/mm³)",
    x = "CD4 (células/mm³)",
    y = NULL
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot",
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )

# --- Gráfico Carga Viral < 50 ---
p_cv_50 <- df_longitudinal %>%
  mutate(cv50_label = fct_recode(deteccao_cv50,
    "Indetectável" = "CV indetectável",
    "Detectável" = "CV detectável",
    "Sem CV" = "Sem dados de CV"
  )) %>%
  ggplot(aes(x = cv50_label, fill = cv50_label)) +
  geom_bar(alpha = 0.8) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    subtitle = "Carga Viral (< 50 cópias/mL)",
    x = "Detectabilidade (cópias/mL)",
    y = NULL
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot",
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )

# --- Painel combinado ---
p_cd4_200 | p_cd4_350 | p_cv_50
```

```r
#-------------------------------------------------------------------------------
# Bloco: Tabelas de Distribuição das Variáveis Clínicas Iniciais
#-------------------------------------------------------------------------------

# --- Tabela CD4 (corte 200) ---
knitr::kable(
  df_longitudinal %>%
    count(cd4_cat200) %>%
    mutate(Proporcao = scales::percent(n / sum(n), accuracy = 0.01)),
  caption = "Distribuição de CD4 ( 200 células/mm³)"
)

# --- Tabela CD4 (corte 350) ---
knitr::kable(
  df_longitudinal %>%
    count(cd4_cat350) %>%
    mutate(Proporcao = scales::percent(n / sum(n), accuracy = 0.01)),
  caption = "Distribuição de CD4 ( 350 células/mm³)"
)

# --- Tabela Carga Viral (corte 50 cópias/mL) ---
knitr::kable(
  df_longitudinal %>%
    count(deteccao_cv50) %>%
    mutate(Proporcao = scales::percent(n / sum(n), accuracy = 0.01)),
  caption = "Distribuição da Carga Viral (corte 50 cópias/mL)"
)
```

```r
#-------------------------------------------------------------------------------
# Bloco: Distribuição do Tempo de Seguimento e Status do Evento
#-------------------------------------------------------------------------------

# --- Histograma do Tempo de Seguimento (Geral) ---
p_hist <- df_longitudinal %>%
  ggplot(aes(x = time_to_event)) +
  geom_histogram(binwidth = 1, fill = cores[10], color = "white", alpha = 0.8) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
    ↪  ",")) +
  labs(
    title = "Distribuição do Tempo de Seguimento em TARV",
    subtitle = "Tempo até a interrupção ou censura.",
    x = "Tempo (anos)",
    y = "Número de Pacientes"
```

```r
  ) +
  theme_clean

# --- Gráfico de Barras do Status do Evento ---
event_counts <- df_longitudinal %>%
  count(event_status, name = "Contagem") %>%
  mutate(Proporcao = Contagem / sum(Contagem))

p_event <- ggplot(event_counts, aes(x = event_status, y = Contagem, fill =
↪  event_status)) +
  geom_col(alpha = 0.8, show.legend = FALSE) +
  scale_fill_manual(values = cores) +
  scale_y_continuous(
    limits = c(0, max(event_counts$Contagem) * 1.2),
    labels = scales::comma_format(big.mark = ".", decimal.mark = ",")
  ) +
  labs(
    title = "Distribuição dos Tipos de Eventos",
    subtitle = "Interrupção vs. Censura na coorte.",
    x = "Status do Evento",
    y = "Número de Pacientes"
  ) +
  theme_clean

# --- Histograma apenas para Interrupção ---
p_hist_interrupcao <- df_longitudinal %>%
  filter(event_status == "Interrupção") %>%
  ggplot(aes(x = time_to_event)) +
  geom_histogram(binwidth = 1, fill = cores[4], color = "white", alpha = 0.8) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    title = "Distribuição do Tempo até Interrupção",
    subtitle = "Apenas pacientes com evento de interrupção.",
    x = "Tempo (anos)",
    y = "Número de Pacientes"
  ) +
  theme_clean

# --- Histograma apenas para Censura ---
p_hist_censura <- df_longitudinal %>%
  filter(event_status == "Censurado") %>%
  ggplot(aes(x = time_to_event)) +
  geom_histogram(binwidth = 1, fill = cores[7], color = "white", alpha = 0.8) +
```

```
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark =
  ↪  ",")) +
  labs(
    title = "Distribuição do Tempo para Censura",
    subtitle = "Apenas pacientes censurados.",
    x = "Tempo (anos)",
    y = "Número de Pacientes"
  ) +
  theme_clean

# --- Painel final com os quatro gráficos ---
(p_hist | p_event) / (p_hist_interrupcao | p_hist_censura)

# --- Impressão da Tabela com Contagens e Proporções ---
cat("\n\n================ Distribuição dos Tipos de Evento ================\n\n")
print(event_counts)

# --- Total de Pacientes ---
total_pacientes <- sum(event_counts$Contagem)
cat("\nTotal de pacientes na coorte: ", total_pacientes, "\n")
```

```
#-------------------------------------------------------------------------------
# Bloco: Boxplots e Tabelas Descritivas do Tempo em TARV por Variáveis
↪  Sociodemográficas
#-------------------------------------------------------------------------------

#-------------------------------------------------------------------------------
# Função auxiliar para calcular estatísticas descritivas para boxplots
get_boxplot_stats <- function(df, group_var) {
  df %>%
    group_by({{ group_var }}) %>%
    summarise(
      N = n(),
      Média = mean(time_to_event, na.rm = TRUE),
      Mediana = median(time_to_event, na.rm = TRUE),
      Q1 = quantile(time_to_event, 0.25, na.rm = TRUE),
      Q3 = quantile(time_to_event, 0.75, na.rm = TRUE),
      IQR = Q3 - Q1,
      Mínimo = min(time_to_event, na.rm = TRUE),
      Máximo = max(time_to_event, na.rm = TRUE),
      .groups = 'drop'
    )
}

```

```r

# --- Boxplot por Sexo ---
p_sexo <- ggplot(df_longitudinal, aes(x = sexo, y = time_to_event, fill = sexo, color = sexo)) +
  geom_boxplot(linewidth = 0.3, alpha = 0.8) +
  stat_summary(
    fun = mean,
    geom = "crossbar",
    width = 0.75,
    linetype = "dashed",
    linewidth = 1.5,
    fatten = 0,
    aes(color = sexo),
    show.legend = FALSE
  ) +
  scale_fill_manual(values = cores) +
  scale_color_manual(values = cores_escurecidas) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark = ",")) +
  labs(
    subtitle = "Sexo",
    x = "Sexo de Nascimento",
    y = "Tempo em TARV (Anos)"
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Boxplot por Raça/Cor ---
p_raca <- ggplot(df_longitudinal, aes(x = raca, y = time_to_event, fill = raca, color = raca)) +
  geom_boxplot(linewidth = 0.3, alpha = 0.8) +
  stat_summary(
    fun = mean,
    geom = "crossbar",
    width = 0.75,
    linetype = "dashed",
    linewidth = 1.5,
    fatten = 0,
    aes(color = raca),
    show.legend = FALSE
  ) +
  scale_fill_manual(values = cores) +
```

```r
  scale_color_manual(values = cores_escurecidas) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark = ",")) +
  labs(
    subtitle = "Raça/Cor",
    x = "Raça/Cor Declarada",
    y = NULL
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Boxplot por Escolaridade ---
p_escolaridade <- ggplot(df_longitudinal, aes(x = escolaridade, y = time_to_event, fill = escolaridade, color = escolaridade)) +
  geom_boxplot(linewidth = 0.3, alpha = 0.8) +
  stat_summary(
    fun = mean,
    geom = "crossbar",
    width = 0.75,
    linetype = "dashed",
    linewidth = 1.5,
    fatten = 0,
    aes(color = escolaridade),
    show.legend = FALSE
  ) +
  scale_fill_manual(values = cores) +
  scale_color_manual(values = cores_escurecidas) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark = ",")) +
  labs(
    subtitle = "Escolaridade",
    x = "Nível de Escolaridade (anos)",
    y = "Tempo em TARV (anos)"
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Boxplot por Faixa Etária ---
```

```r
p_faixa_etaria <- ggplot(df_longitudinal, aes(x = faixa_etaria_inicio, y = time_to_event, fill = faixa_etaria_inicio, color = faixa_etaria_inicio)) +
  geom_boxplot(linewidth = 0.3, alpha = 0.8) +
  stat_summary(
    fun = mean,
    geom = "crossbar",
    width = 0.75,
    linetype = "dashed",
    linewidth = 1.5,
    fatten = 0,
    aes(color = faixa_etaria_inicio),
    show.legend = FALSE
  ) +
  scale_fill_manual(values = cores) +
  scale_color_manual(values = cores_escurecidas) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".", decimal.mark = ",")) +
  labs(
    subtitle = "Faixa Etária",
    x = "Faixa Etária (anos)",
    y = NULL
  ) +
  theme_clean +
  theme(
    legend.position = "none",
    plot.title.position = "plot"
  )

# --- Painel final ---
((p_sexo | p_raca) / (p_escolaridade | p_faixa_etaria)) +
  plot_annotation(
    title = "Tempo em TARV por Variáveis Sociodemográficas",
    theme = theme(
      plot.title = element_text(
        face = "bold",
        size = 14,
        hjust = 0
      )
    )
  )

#------------------------------------------------------------------------------
# Bloco: Tabelas de Estatísticas Descritivas
#------------------------------------------------------------------------------

```

```r
# Calcula as estatísticas
sexo_stats <- get_boxplot_stats(df_longitudinal, sexo)
raca_stats <- get_boxplot_stats(df_longitudinal, raca)
escolaridade_stats <- get_boxplot_stats(df_longitudinal, escolaridade)
faixa_etaria_stats <- get_boxplot_stats(df_longitudinal, faixa_etaria_inicio)

# Exibe as tabelas
knitr::kable(sexo_stats, caption = "Estatísticas do Tempo em TARV por Sexo", digits = 2)
knitr::kable(raca_stats, caption = "Estatísticas do Tempo em TARV por Raça/Cor", digits = 2)
knitr::kable(escolaridade_stats, caption = "Estatísticas do Tempo em TARV por Escolaridade", digits = 2)
knitr::kable(faixa_etaria_stats, caption = "Estatísticas do Tempo em TARV por Faixa Etária", digits = 2)
```

```r
#------------------------------------------------------------------------------
# Bloco: Análise Geográfica - Distribuição por UF e Tempo Médio de Permanência
#------------------------------------------------------------------------------
# Junta df_longitudinal com ufs_dict para adicionar as siglas das UFs
df_longitudinal_geo <- df_longitudinal %>%
  left_join(ufs_dict, by = c("uf_inst_completo" = "name_state"))

# Verificação de integridade: alguma UF sem sigla após o join?
ufs_na_geo <- df_longitudinal_geo %>%
  filter(is.na(abbrev_state)) %>%
  distinct(uf_inst_completo)
if(nrow(ufs_na_geo) > 0) {
  print("--- UFs com problemas no join ---")
  print(ufs_na_geo)
}

#------------------------------------------------------------------------------
# Geometria dos estados
#------------------------------------------------------------------------------
estados_geo <- read_state(code_state = "all", year = 2020)

#------------------------------------------------------------------------------
# Mapa coroplético 1: Proporção de pacientes por UF
#------------------------------------------------------------------------------
prop_por_uf <- df_longitudinal_geo %>%
  filter(!is.na(abbrev_state)) %>%
  count(abbrev_state) %>%
  mutate(Prop = n / sum(n))
```

```r

estados_mapa_prop <- estados_geo %>%
  left_join(prop_por_uf, by = "abbrev_state")

p_mapa_prop <- ggplot(data = estados_mapa_prop) +
  geom_sf(aes(fill = Prop), color = "white", linewidth = 0.5) +
  geom_sf_text(
    aes(label = paste0(abbrev_state, "\n", scales::percent(Prop, accuracy = 0.1,
    ↪  decimal.mark = ","))),
    color = "black", size = 2.5, fontface = "bold", check_overlap = TRUE
  ) +
  scale_fill_scico(
    palette = "buda",
    direction = -1,
    name = "Proporção",
    labels = scales::percent_format(accuracy = 1, decimal.mark = ","),
    na.value = "gray90"
  ) +
  labs(title = "Distribuição de Pacientes por UF") +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0, face = "bold", size = 14),
    legend.position = "right"
  )

#------------------------------------------------------------------------------
# Mapa coroplético 2: Tempo Médio de Permanência por UF
#------------------------------------------------------------------------------
tempo_medio_por_uf <- df_longitudinal_geo %>%
  group_by(abbrev_state) %>%
  summarise(
    tempo_medio_anos = mean(time_to_event, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  filter(!is.na(abbrev_state))

estados_mapa_tempo <- estados_geo %>%
  left_join(tempo_medio_por_uf, by = "abbrev_state")

p_mapa_tempo <- ggplot(data = estados_mapa_tempo) +
  geom_sf(aes(fill = tempo_medio_anos), color = "white", linewidth = 0.5) +
  geom_sf_text(
    aes(label = paste0(abbrev_state, "\n", scales::number(tempo_medio_anos, accuracy =
    ↪  0.1, decimal.mark = ","))),
    color = "black", size = 2.5, fontface = "bold", check_overlap = TRUE
```

```r
  ) +
  scale_fill_scico(
    palette = "imola",
    direction = -1,
    name = "Tempo Médio em TARV (anos)",
    na.value = "gray90"
  ) +
  labs(title = "Tempo Médio de Permanência em TARV por UF") +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0, face = "bold", size = 14),
    legend.position = "right"
  )

#------------------------------------------------------------------------------
# Exibe os dois mapas lado a lado
#------------------------------------------------------------------------------
(p_mapa_prop | p_mapa_tempo) +
  plot_layout(widths = c(1, 1)) &
  theme(plot.margin = margin(0, 0, 0, 0))
```

Apêndice C

Código do Capítulo 5 (Python)

Este apêndice contém o código completo em linguagem Python utilizado para as análises do Capítulo 5. O script foi implementado em ambiente `Jupyter Notebook`, com uso das bibliotecas `PyTorch`, `lifelines`, `scikit-survival`, `scikit-learn`, `matplotlib` e `seaborn`. A arquitetura neural implementada segue o modelo *Nnet-Survival* em tempo discreto com função de perda baseada em log-verossimilhança negativa. Os blocos estão organizados conforme a sequência de execução no notebook original.

```
# ============================================================================
# 1. CONFIGURAÇÃO INICIAL E IMPORTAÇÃO DE BIBLIOTECAS
# ============================================================================

# Instalação de dependências necessárias
import sys
!pip install lifelines pyreadr scikit-survival -q

# Importação de bibliotecas principais
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import brier_score_loss
from sksurv.metrics import concordance_index_censored
from lifelines import KaplanMeierFitter
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import Dataset, DataLoader
import pyreadr
import warnings
warnings.filterwarnings('ignore')

# Configuração de visualização
```

```
sns.set_style("whitegrid")
plt.rcParams['figure.figsize'] = (12, 6)
plt.rcParams['font.size'] = 11

# Paleta de cores personalizada para os gráficos
CORES_CUSTOM = ["#ebac23", "#b80058", "#008cf9", "#006e00",
                "#00bbad", "#d163e6", "#b24502", "#ff9287",
                "#5954d6", "#00c6f8", "#878500", "#00a76c", "#bdbdbd"]

# Seed para reprodutibilidade dos resultados
SEED = 42
np.random.seed(SEED)
torch.manual_seed(SEED)
if torch.cuda.is_available():
    torch.cuda.manual_seed(SEED)
    torch.backends.cudnn.deterministic = True
    torch.backends.cudnn.benchmark = False

torch.use_deterministic_algorithms(True)

# Criar um generator para o DataLoader
g = torch.Generator()
g.manual_seed(SEED)

print("="*80)
print("CONFIGURAÇÃO INICIAL CONCLUÍDA")
print("="*80)
print(f"NumPy version: {np.__version__}")
print(f"PyTorch version: {torch.__version__}")
print(f"CUDA disponível: {torch.cuda.is_available()}")
print(f"Seed fixada: {SEED}")
```

```
# ============================================================================
# 2. CARREGAMENTO E EXPLORAÇÃO INICIAL DOS DADOS
# ============================================================================

print("\n" + "="*80)
print("CARREGAMENTO DA BASE DE DADOS")
print("="*80)

# Carregar arquivo RData com os dados da coorte
caminho_arquivo = 'df_longitudinal.RData'
resultado = pyreadr.read_r(caminho_arquivo)
nome_df = list(resultado.keys())[0]
df = resultado[nome_df]

print(f"Objeto carregado: {nome_df}")
print(f"Dimensões: {df.shape[0]:,} indivíduos × {df.shape[1]} variáveis")

# Exibir estrutura dos dados
print("\n" + "-"*80)
print("ESTRUTURA DOS DADOS")
print("-"*80)
print(df.info())

```

```
# Primeiras observações
print("\n" + "-"*80)
print("PRIMEIRAS OBSERVAÇÕES")
print("-"*80)
print(df.head())

# Verificar valores ausentes
print("\n" + "-"*80)
print("VALORES AUSENTES")
print("-"*80)
valores_ausentes = df.isnull().sum()
if valores_ausentes.sum() == 0:
    print("Nenhum valor ausente detectado")
else:
    print(valores_ausentes[valores_ausentes > 0])

# Estatísticas da variável resposta
print("\n" + "-"*80)
print("VARIÁVEL RESPOSTA: event_status")
print("-"*80)
print(df['event_status'].value_counts())
print(f"\nTaxa de eventos: {(df['event_status']=='Interrupção').mean()*100:.2f}%")
print(f"Taxa de censura: {(df['event_status']=='Censurado').mean()*100:.2f}%")

# Estatísticas do tempo de seguimento
print("\n" + "-"*80)
print("TEMPO DE SEGUIMENTO: time_to_event")
print("-"*80)
print(df['time_to_event'].describe())
```

```
# ============================================================================
# 3. PRÉ-PROCESSAMENTO DOS DADOS
# ============================================================================

print("\n" + "="*80)
print("PRÉ-PROCESSAMENTO")
print("="*80)

# Criar cópia para manipulação
df_model = df.copy()

# Codificar variável resposta binária (delta)
# delta = 1 indica evento observado (interrupção do tratamento)
# delta = 0 indica censura
print("\n1. Codificação da variável resposta (delta)")
df_model['delta'] = (df_model['event_status'] == 'Interrupção').astype(int)
print(f"   Eventos (delta=1): {(df_model['delta']==1).sum():,} "
      f"({(df_model['delta']==1).mean()*100:.2f}%)")
print(f"   Censurados (delta=0): {(df_model['delta']==0).sum():,} "
      f"({(df_model['delta']==0).mean()*100:.2f}%)")

# Remover variáveis não utilizadas no modelo
print("\n2. Remoção de variáveis")
colunas_remover = ['cod_unificado', 'event_status', 'cd4_cat350']
df_model = df_model.drop(columns=[c for c in colunas_remover
```

```
                                    if c in df_model.columns])
print(f"   Removidas: {[c for c in colunas_remover if c in df.columns]}")

# Agregar UF em Região para reduzir cardinalidade
print("\n3. Agregação UF → Região")
uf_para_regiao = {
    'Acre': 'Norte', 'Amapá': 'Norte', 'Amazonas': 'Norte',
    'Pará': 'Norte', 'Rondônia': 'Norte', 'Roraima': 'Norte',
    'Tocantins': 'Norte',
    'Alagoas': 'Nordeste', 'Bahia': 'Nordeste', 'Ceará': 'Nordeste',
    'Maranhão': 'Nordeste', 'Paraíba': 'Nordeste',
    'Pernambuco': 'Nordeste', 'Piauí': 'Nordeste',
    'Rio Grande do Norte': 'Nordeste', 'Sergipe': 'Nordeste',
    'Distrito Federal': 'Centro-Oeste', 'Goiás': 'Centro-Oeste',
    'Mato Grosso': 'Centro-Oeste', 'Mato Grosso do Sul': 'Centro-Oeste',
    'Espírito Santo': 'Sudeste', 'Minas Gerais': 'Sudeste',
    'Rio de Janeiro': 'Sudeste', 'São Paulo': 'Sudeste',
    'Paraná': 'Sul', 'Rio Grande do Sul': 'Sul', 'Santa Catarina': 'Sul'
}

if 'uf_inst_completo' in df_model.columns:
    df_model['regiao'] = df_model['uf_inst_completo'].map(uf_para_regiao)
    df_model = df_model.drop(columns=['uf_inst_completo'])
    print("   Variável 'regiao' criada")
    print(f"\n   Distribuição por região:")
    print(df_model['regiao'].value_counts().sort_index())

# Codificação One-Hot das variáveis categóricas
print("\n4. Codificação One-Hot das variáveis categóricas")

variaveis_categoricas = []
for col in df_model.columns:
    if col not in ['time_to_event', 'delta']:
        if df_model[col].dtype in ['object', 'category']:
            variaveis_categoricas.append(col)

print(f"   Variáveis identificadas: {variaveis_categoricas}")

df_encoded = df_model.copy()
for var in variaveis_categoricas:
    dummies = pd.get_dummies(df_encoded[var], prefix=var,
                             drop_first=True, dtype=float)
    df_encoded = pd.concat([df_encoded, dummies], axis=1)
    df_encoded = df_encoded.drop(columns=[var])
    print(f"   {var}: {dummies.shape[1]} variáveis binárias criadas")

print(f"\nDimensões finais: {df_encoded.shape[0]:,} × {df_encoded.shape[1]}")
```

```
# ============================================================================
# 4. DIVISÃO TREINO-VALIDAÇÃO-TESTE E PADRONIZAÇÃO
# ============================================================================

print("\n" + "="*80)
print("DIVISÃO DOS DADOS")
print("="*80)
```

```

# Separar features (X), tempo (T) e indicador de evento (delta)
colunas_features = [col for col in df_encoded.columns
                    if col not in ['time_to_event', 'delta']]
X = df_encoded[colunas_features].values.astype(np.float32)
T = df_encoded['time_to_event'].values.astype(np.int64)
delta = df_encoded['delta'].values.astype(np.float32)

print(f"Features (X): {X.shape}")
print(f"Tempo (T): {T.shape}, range=[{T.min()}, {T.max()}]")
print(f"Evento (delta): {delta.shape}, eventos={delta.sum():.0f} "
      f"({delta.mean()*100:.2f}%)")

# Divisão estratificada: 80% treino+validação, 20% teste
X_train_val, X_test, T_train_val, T_test, delta_train_val, delta_test = \
    train_test_split(X, T, delta, test_size=0.2, random_state=SEED,
                     stratify=delta)

# Subdivisão: 80% treino, 20% validação
X_train, X_val, T_train, T_val, delta_train, delta_val = \
    train_test_split(X_train_val, T_train_val, delta_train_val,
                     test_size=0.2, random_state=SEED,
                     stratify=delta_train_val)

print(f"\nTreino: {len(X_train):,} ({len(X_train)/len(X)*100:.1f}%)")
print(f"  Eventos: {delta_train.sum():.0f} ({delta_train.mean()*100:.2f}%)")

print(f"Validação: {len(X_val):,} ({len(X_val)/len(X)*100:.1f}%)")
print(f"  Eventos: {delta_val.sum():.0f} ({delta_val.mean()*100:.2f}%)")

print(f"Teste: {len(X_test):,} ({len(X_test)/len(X)*100:.1f}%)")
print(f"  Eventos: {delta_test.sum():.0f} ({delta_test.mean()*100:.2f}%)")

# Padronização Z-score
print("\n" + "-"*80)
print("PADRONIZAÇÃO")
print("-"*80)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_val_scaled = scaler.transform(X_val)
X_test_scaled = scaler.transform(X_test)

print(f"Treino - Média (primeiras 5 features): "
      f"{X_train_scaled[:, :5].mean(axis=0)}")
print(f"Treino - Std (primeiras 5 features): "
      f"{X_train_scaled[:, :5].std(axis=0)}")
```

```

# ============================================================================
# 5. PREPARAÇÃO DOS DATALOADERS (PyTorch)
# ============================================================================

print("\n" + "="*80)
print("CRIAÇÃO DOS DATALOADERS")
print("="*80)
```

```

class SurvivalDataset(Dataset):
    """
    Dataset customizado para dados de sobrevivência.
    Retorna triplas (X, T, delta) para cada observação.
    """
    def __init__(self, X, T, delta):
        self.X = torch.FloatTensor(X)
        self.T = torch.LongTensor(T)
        self.delta = torch.FloatTensor(delta)

    def __len__(self):
        return len(self.X)

    def __getitem__(self, idx):
        return self.X[idx], self.T[idx], self.delta[idx]

# Criar datasets PyTorch
train_dataset = SurvivalDataset(X_train_scaled, T_train, delta_train)
val_dataset = SurvivalDataset(X_val_scaled, T_val, delta_val)
test_dataset = SurvivalDataset(X_test_scaled, T_test, delta_test)

# Criar DataLoaders para batch processing
batch_size = 2048
train_loader = DataLoader(train_dataset, batch_size=batch_size,
                          shuffle=True, num_workers=0)
val_loader = DataLoader(val_dataset, batch_size=batch_size,
                        shuffle=False, num_workers=0)
test_loader = DataLoader(test_dataset, batch_size=batch_size,
                         shuffle=False, num_workers=0)

print(f"Batch size: {batch_size:,}")
print(f"Batches - Treino: {len(train_loader):,}, "
      f"Validação: {len(val_loader):,}, Teste: {len(test_loader):,}")
```

```

# ============================================================================
# 6. DEFINIÇÃO DO MODELO NNET-SURVIVAL
# ============================================================================

print("\n" + "="*80)
print("ARQUITETURA DO MODELO")
print("="*80)

class NnetSurvival(nn.Module):
    """
    Rede Neural para Análise de Sobrevivência em Tempo Discreto.

    Arquitetura:
    - Camada de entrada: p features
    - Camada oculta 1: hidden_dim1 neurônios + ReLU + Dropout
    - Camada oculta 2: hidden_dim2 neurônios + ReLU + Dropout
    - Camada de saída: K+1 neurônios (intervalos de tempo) + Sigmoid
    """
    def __init__(self, input_dim, num_time_intervals,
                 hidden_dim1=128, hidden_dim2=64, dropout_rate=0.3):
```

```
        super(NnetSurvival, self).__init__()

        # Primeira camada oculta
        self.fc1 = nn.Linear(input_dim, hidden_dim1)
        self.relu1 = nn.ReLU()
        self.dropout1 = nn.Dropout(dropout_rate)

        # Segunda camada oculta
        self.fc2 = nn.Linear(hidden_dim1, hidden_dim2)
        self.relu2 = nn.ReLU()
        self.dropout2 = nn.Dropout(dropout_rate)

        # Camada de saída
        self.fc_out = nn.Linear(hidden_dim2, num_time_intervals)
        self.sigmoid = nn.Sigmoid()

    def forward(self, x):
        """Forward pass: X → h1 → h2 → hazards"""
        x = self.fc1(x)
        x = self.relu1(x)
        x = self.dropout1(x)

        x = self.fc2(x)
        x = self.relu2(x)
        x = self.dropout2(x)

        hazards = self.fc_out(x)
        hazards = self.sigmoid(hazards)

        return hazards

# Hiperparâmetros da arquitetura
input_dim = X_train_scaled.shape[1]
num_time_intervals = 25  # k = 0, 1, ..., 24 anos
hidden_dim1 = 128
hidden_dim2 = 64
dropout_rate = 0.3

# Selecionar dispositivo (GPU se disponível, caso contrário CPU)
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
print(f"Dispositivo: {device}")

# Instanciar modelo e mover para dispositivo
model = NnetSurvival(input_dim=input_dim,
                     num_time_intervals=num_time_intervals,
                     hidden_dim1=hidden_dim1,
                     hidden_dim2=hidden_dim2,
                     dropout_rate=dropout_rate).to(device)

# Contar parâmetros treináveis
total_params = sum(p.numel() for p in model.parameters())
print(f"\nArquitetura:")
print(f"  Entrada: {input_dim} features")
print(f"  Oculta 1: {hidden_dim1} neurônios (ReLU + Dropout {dropout_rate})")
print(f"  Oculta 2: {hidden_dim2} neurônios (ReLU + Dropout {dropout_rate})")
print(f"  Saída: {num_time_intervals} intervalos (Sigmoid)")
```

```
print(f"\nParâmetros treináveis: {total_params:,}")
```

```
# ============================================================================
# 7. FUNÇÃO DE PERDA (LOG-VEROSSIMILHANÇA NEGATIVA)
# ============================================================================

def discrete_time_loss(hazards, time, event, eps=1e-7):
    """
    Calcula a log-verossimilhança negativa de tempo discreto.

    Formulação matemática:
    L() = -[_i log h_{ik_i} + _{j=1}^{k_i - _i} log(1 - h_{ij})]

    Parâmetros:
    -----------
    hazards : torch.Tensor (batch_size, num_time_intervals)
        Probabilidades de risco discreto h_{ik}
    time : torch.Tensor (batch_size,)
        Tempo observado k_i para cada indivíduo
    event : torch.Tensor (batch_size,)
        Indicador _i (1=evento, 0=censura)
    eps : float
        Epsilon para estabilidade numérica

    Retorna:
    --------
    loss : torch.Tensor (escalar)
        Log-verossimilhança negativa média
    """
    batch_size = hazards.shape[0]
    num_time_intervals = hazards.shape[1]

    # Clipping para evitar log(0)
    hazards = torch.clamp(hazards, eps, 1 - eps)

    # Criar máscara para intervalos j  k_i
    time_expanded = time.unsqueeze(1).expand(-1, num_time_intervals)
    interval_indices = torch.arange(num_time_intervals, device=hazards.device)
    interval_indices = interval_indices.unsqueeze(0).expand(batch_size, -1)
    mask = interval_indices <= time_expanded

    # Calcular  log(1 - h_j) para j  k_i
    log_survival_intervals = torch.log(1 - hazards + eps)
    cumsum_log_survival = (log_survival_intervals * mask.float()).sum(dim=1)

    # Para eventos: adicionar log(h_{k_i}) - log(1 - h_{k_i})
    time_long = time.long()
    log_hazard_at_event = torch.log(
        hazards.gather(1, time_long.unsqueeze(1)).squeeze() + eps
    )
    log_survival_at_event = torch.log(
        1 - hazards.gather(1, time_long.unsqueeze(1)).squeeze() + eps
    )

    # Log-verossimilhança por indivíduo
```

```
    log_likelihood = (event.float() * (log_hazard_at_event -
                      log_survival_at_event) + cumsum_log_survival)

    return -log_likelihood.mean()

print("\n" + "="*80)
print("FUNÇÃO DE PERDA DEFINIDA")
print("="*80)
print("Tipo: Log-verossimilhança negativa de tempo discreto")
print("Componentes:")
print("  - Tratamento diferenciado para eventos e censuras")
print("  - Estabilização numérica (eps=1e-7)")
print("  - Mascaramento temporal (apenas j  k_i)")
```

```
# ================================================================================
# 8. CONFIGURAÇÃO DO TREINAMENTO
# ================================================================================

print("\n" + "="*80)
print("CONFIGURAÇÃO DO TREINAMENTO")
print("="*80)

# Hiperparâmetros de otimização
learning_rate = 0.001
weight_decay = 1e-5  # Regularização L2
num_epochs = 50
patience_early_stop = 10

# Otimizador Adam com weight decay
optimizer = optim.Adam(model.parameters(), lr=learning_rate,
                       weight_decay=weight_decay)

# Scheduler: Reduz learning rate quando perda de validação estagna
scheduler = optim.lr_scheduler.ReduceLROnPlateau(
    optimizer, mode='min', factor=0.5, patience=5
)

print(f"Otimizador: Adam")
print(f"  Learning rate inicial: {learning_rate}")
print(f"  Weight decay (L2): {weight_decay}")
print(f"\nScheduler: ReduceLROnPlateau")
print(f"  Fator de redução: 0.5")
print(f"  Paciência: 5 epochs")
print(f"\nEarly Stopping: {patience_early_stop} epochs sem melhoria")
print(f"Epochs máximas: {num_epochs}")
```

```
# ================================================================================
# 9. LOOP DE TREINAMENTO
# ================================================================================

print("\n" + "="*80)
print("INICIANDO TREINAMENTO")
print("="*80)
```

```
8
9  # Variáveis de controle
10 best_val_loss = float('inf')
11 epochs_without_improvement = 0
12 train_losses = []
13 val_losses = []
14
15 for epoch in range(num_epochs):
16     # Fase de treino
17     model.train()
18     train_loss_epoch = 0.0
19
20     for X_batch, T_batch, delta_batch in train_loader:
21         X_batch = X_batch.to(device)
22         T_batch = T_batch.to(device)
23         delta_batch = delta_batch.to(device)
24
25         optimizer.zero_grad()
26         hazards = model(X_batch)
27         loss = discrete_time_loss(hazards, T_batch, delta_batch)
28         loss.backward()
29         optimizer.step()
30
31         train_loss_epoch += loss.item()
32
33     train_loss_epoch /= len(train_loader)
34     train_losses.append(train_loss_epoch)
35
36     # Fase de validação
37     model.eval()
38     val_loss_epoch = 0.0
39
40     with torch.no_grad():
41         for X_batch, T_batch, delta_batch in val_loader:
42             X_batch = X_batch.to(device)
43             T_batch = T_batch.to(device)
44             delta_batch = delta_batch.to(device)
45
46             hazards = model(X_batch)
47             loss = discrete_time_loss(hazards, T_batch, delta_batch)
48             val_loss_epoch += loss.item()
49
50     val_loss_epoch /= len(val_loader)
51     val_losses.append(val_loss_epoch)
52
53     # Atualização do learning rate
54     old_lr = optimizer.param_groups[0]['lr']
55     scheduler.step(val_loss_epoch)
56     new_lr = optimizer.param_groups[0]['lr']
57
58     if new_lr < old_lr:
59         print(f"   Learning rate reduzido: {old_lr:.6f} → {new_lr:.6f}")
60
61     # Early stopping
62     if val_loss_epoch < best_val_loss:
63         best_val_loss = val_loss_epoch
64         epochs_without_improvement = 0
```

```
        torch.save(model.state_dict(), 'best_nnet_survival_model.pth')
    else:
        epochs_without_improvement += 1

    # Logging
    if (epoch + 1) % 5 == 0 or epoch == 0:
        print(f"Epoch [{epoch+1:>2}/{num_epochs}] | "
              f"L_treino={train_loss_epoch:.4f} | "
              f"L_val={val_loss_epoch:.4f} | "
              f"Best={best_val_loss:.4f} | "
              f"LR={new_lr:.6f}")

    # Parada antecipada
    if epochs_without_improvement >= patience_early_stop:
        print(f"\nEarly stopping em epoch {epoch+1}")
        print(f"Melhor perda de validação: {best_val_loss:.4f}")
        break

# Carregar melhor modelo
model.load_state_dict(torch.load('best_nnet_survival_model.pth'))

print("\n" + "="*80)
print("TREINAMENTO CONCLUÍDO")
print("="*80)
print(f"Epochs treinadas: {len(train_losses)}")
print(f"Melhor L_val: {best_val_loss:.4f}")
print(f"L_treino final: {train_losses[-1]:.4f}")
print(f"L_val final: {val_losses[-1]:.4f}")
```

```
# ============================================================================
# 10. VISUALIZAÇÃO: CURVA DE APRENDIZADO
# ============================================================================

print("\n" + "="*80)
print("FIGURA 5.1: CURVA DE APRENDIZADO")
print("="*80)

plt.figure(figsize=(14, 6))

plt.plot(range(1, len(train_losses)+1), train_losses,
         label=r'$\mathcal{L}(\Theta)$ Treino', linewidth=2.5,
         color=CORES_CUSTOM[0], marker='o', markersize=4,
         markevery=max(1, len(train_losses)//10))

plt.plot(range(1, len(train_losses)+1), val_losses,
         label=r'$\mathcal{L}(\Theta)$ Validação', linewidth=2.5,
         color=CORES_CUSTOM[1], marker='s', markersize=4,
         markevery=max(1, len(train_losses)//10))

plt.axhline(y=best_val_loss, color='blue', linestyle=':', linewidth=1.5,
           alpha=0.7, label=f'Melhor={best_val_loss:.4f}')

plt.xlabel('Epoch', fontsize=13, fontweight='bold')
plt.ylabel(r'$-\log \mathcal{L}(\Theta)$', fontsize=13, fontweight='bold')
plt.title('Curva de Aprendizado', fontsize=15, fontweight='bold', pad=20)
```

```
plt.legend(fontsize=12, loc='best', frameon=True, shadow=True)
plt.grid(True, alpha=0.3, linestyle='--')
plt.tight_layout()
plt.savefig('figura_5_1_curva_aprendizado.png', dpi=300, bbox_inches='tight')
sns.despine()
plt.show()

# Análise de overfitting
diff_prop = abs(train_losses[-1] - val_losses[-1]) / val_losses[-1]
print(f"\nDiferença treino-validação: {diff_prop*100:.2f}%")
if diff_prop < 0.05:
    print("Avaliação: Ausência de sobreajuste")
elif diff_prop < 0.10:
    print("Avaliação: Sobreajuste moderado")
else:
    print("Avaliação: Possível sobreajuste significativo")
```

```
# ================================================================================
# 11. GERAÇÃO DE PREDIÇÕES NO CONJUNTO DE TESTE
# ================================================================================

print("\n" + "="*80)
print("GERAÇÃO DE PREDIÇÕES (CONJUNTO DE TESTE)")
print("="*80)

def calculate_survival_function(hazards):
    """
    Calcula a função de sobrevivência S(k) a partir dos hazards.
    S(k) = _{j=0}^{k} (1 - h_j)
    """
    if isinstance(hazards, torch.Tensor):
        hazards = hazards.detach().cpu().numpy()

    survival = np.cumprod(1 - hazards + 1e-7, axis=1)
    return survival

# Gerar predições para o conjunto de teste
model.eval()
all_hazards_test = []

with torch.no_grad():
    for X_batch, _, _ in test_loader:
        X_batch = X_batch.to(device)
        hazards = model(X_batch)
        all_hazards_test.append(hazards.cpu())

hazards_test = torch.cat(all_hazards_test, dim=0)
survival_test = calculate_survival_function(hazards_test)

print(f"Hazards: {hazards_test.shape}")
print(f"Survival: {survival_test.shape}")
print(f"\nEstatísticas S(t=24):")
print(f"  Mínimo: {survival_test[:, -1].min():.4f}")
print(f"  Mediana: {np.median(survival_test[:, -1]):.4f}")
```

```
print(f"  Máximo: {survival_test[:, -1].max():.4f}")
```

```
# ============================================================================
# 12. AVALIAÇÃO: C-INDEX (CAPACIDADE DISCRIMINATIVA)
# ============================================================================

print("\n" + "="*80)
print("MÉTRICA 1: C-INDEX (CONCORDANCE INDEX)")
print("="*80)

# Score de risco: 1 - S(t)
risk_scores = 1 - survival_test[:, -1]

# Calcular C-index usando sksurv
cindex, _, _, _, _ = concordance_index_censored(
    delta_test.astype(bool), T_test.astype(float), risk_scores
)

print(f"C-index (global): {cindex:.4f}")

# Interpretação
if cindex >= 0.8:
    interpretacao = "Excelente"
elif cindex >= 0.7:
    interpretacao = "Boa"
elif cindex >= 0.6:
    interpretacao = "Moderada"
else:
    interpretacao = "Fraca"

print(f"Interpretação: {interpretacao} capacidade discriminativa")
print(f"O modelo ordena corretamente {cindex*100:.1f}% "
      f"dos pares comparáveis")

# Distribuição de risco por grupo
print("\n" + "-"*80)
print("DISTRIBUIÇÃO DO RISCO POR GRUPO")
print("-"*80)

risk_censurados = risk_scores[delta_test == 0]
risk_eventos = risk_scores[delta_test == 1]

print(f"Censurados (n={len(risk_censurados):,}):")
print(f"  Média: {risk_censurados.mean():.4f}")
print(f"  Mediana: {np.median(risk_censurados):.4f}")

print(f"\nEventos (n={len(risk_eventos):,}):")
print(f"  Média: {risk_eventos.mean():.4f}")
print(f"  Mediana: {np.median(risk_eventos):.4f}")

diff_media = risk_eventos.mean() - risk_censurados.mean()
print(f"\nDiferença (Eventos - Censurados): {diff_media:+.4f}")
```

```
# ============================================================================
```

```python
# 13. VISUALIZAÇÃO: DISTRIBUIÇÃO DO RISCO PREDITO
# ============================================================================

print("\n" + "="*80)
print("FIGURA 5.2: DISTRIBUIÇÃO DO RISCO PREDITO")
print("="*80)

fig, axes = plt.subplots(1, 2, figsize=(14, 6))

# Painel A: Histogramas sobrepostos
axes[0].hist(risk_censurados, bins=50, alpha=0.7, color=CORES_CUSTOM[2],
             label=f'Censurados (n={len(risk_censurados):,})',
             edgecolor='black')
axes[0].hist(risk_eventos, bins=50, alpha=0.7, color=CORES_CUSTOM[0],
             label=f'Eventos (n={len(risk_eventos):,})', edgecolor='black')
axes[0].set_xlabel(r'$1 - \hat{S}(24)$', fontsize=12, fontweight='bold')
axes[0].set_ylabel('Frequência', fontsize=12, fontweight='bold')
axes[0].set_title('A) Distribuição do Risco Predito',
                  fontsize=13, fontweight='bold')
axes[0].legend(fontsize=11)
axes[0].grid(True, alpha=0.3)

# Painel B: Boxplots comparativos
bp = axes[1].boxplot([risk_censurados, risk_eventos],
                     labels=['Censurados', 'Eventos'],
                     patch_artist=True, boxprops=dict(linewidth=1.5),
                     medianprops=dict(color=CORES_CUSTOM[1], linewidth=2.5),
                     whiskerprops=dict(linewidth=1.5),
                     capprops=dict(linewidth=1.5))

bp['boxes'][0].set_facecolor(CORES_CUSTOM[2])
bp['boxes'][0].set_alpha(0.7)
bp['boxes'][1].set_facecolor(CORES_CUSTOM[0])
bp['boxes'][1].set_alpha(0.7)

axes[1].set_ylabel(r'$1 - \hat{S}(t)$', fontsize=12, fontweight='bold')
axes[1].set_title('B) Comparação entre Grupos',
                  fontsize=13, fontweight='bold')
axes[1].grid(True, alpha=0.3, axis='y')

plt.tight_layout()
plt.savefig('figura_5_2_distribuicao_risco.png', dpi=300, bbox_inches='tight')
sns.despine()
plt.show()
```

```python
# ============================================================================
# 14. AVALIAÇÃO: BRIER SCORE COM IPCW TRUNCADO
# ============================================================================

print("\n" + "="*80)
print("MÉTRICA 2: BRIER SCORE COM IPCW TRUNCADO (Kvamme et al., 2019)")
print("="*80)

# Estimar G(t) - função de sobrevivência da censura
print("Estimando G(t) via Kaplan-Meier...")
```

```python

kmf = KaplanMeierFitter()
censoring_indicator = 1 - delta_test
kmf.fit(T_test, event_observed=censoring_indicator)
G_t = kmf.survival_function_at_times(np.arange(0, 25)).values

print(f"G(t) estimado para t=0 até t=24")

# Determinar horizonte confiável (Graf et al., 1999)
max_t_reliable = max([t for t in range(25) if G_t[t] > 0.10])
print(f"\nHorizonte confiável (Graf et al., 1999): "
      f"t  {max_t_reliable} anos (G(t) > 0.10)")

# Análise dos pesos IPCW
print("\n" + "-"*80)
print("ANÁLISE DOS PESOS IPCW")
print("-"*80)
print(f"{'t':<10} {'G(t)':<10} {'Peso Original':<16} "
      f"{'Peso Truncado':<16} {'Status'}")
print("-"*80)

MAX_WEIGHT = 5.0

for t in [1, 3, 5, 10, 15, 20, 23]:
    g_t = G_t[min(t, len(G_t)-1)]
    peso_original = 1.0 / max(g_t, 1e-7)
    peso_truncado = min(peso_original, MAX_WEIGHT)

    if t <= max_t_reliable:
        status = "Confiável"
    else:
        status = "Instável"

    print(f"{t:<10} {g_t:<10.4f} {peso_original:<16.2f} "
          f"{peso_truncado:<16.2f} {status}")

def calculate_brier_score_ipcw_truncated(survival_preds, time_obs,
                                         event_obs, G_t, time_eval,
                                         max_weight=5.0, eps=1e-7):
    """
    Brier Score com IPCW truncado (Kvamme et al., 2019).

    BS(t) = (1/n)  w_i (I(T_i > t) - S(t|X_i))²

    Pesos truncados em max_weight para estabilidade.
    """
    n_samples = len(time_obs)
    t_idx = min(int(time_eval), survival_preds.shape[1] - 1)
    y_pred = survival_preds[:, t_idx]

    G_t_eval = G_t[min(time_eval, len(G_t) - 1)]
    G_t_eval = max(G_t_eval, eps)

    bs_sum = 0.0
    weight_sum = 0.0

    for i in range(n_samples):
```

```
        T_i = time_obs[i]
        delta_i = event_obs[i]

        if T_i <= time_eval and delta_i == 1:
            G_T_i = G_t[min(int(T_i), len(G_t) - 1)]
            G_T_i = max(G_T_i, eps)
            weight = min(1.0 / G_T_i, max_weight)
            y_true = 0

        elif T_i > time_eval:
            weight = min(1.0 / G_t_eval, max_weight)
            y_true = 1

        else:
            continue

        bs_sum += weight * (y_pred[i] - y_true) ** 2
        weight_sum += weight

    if weight_sum > 0:
        bs = bs_sum / weight_sum
        n_eval = int(weight_sum)
    else:
        bs = np.nan
        n_eval = 0

    return bs, n_eval

# Calcular Brier Score para cada horizonte temporal
print("\n" + "-"*80)
print(f"CALCULANDO BRIER SCORE (max_weight={MAX_WEIGHT})")
print("-"*80)

times_eval = list(range(1, 25))
brier_scores_list = []

for t in times_eval:
    bs, n_eval = calculate_brier_score_ipcw_truncated(
        survival_test, T_test, delta_test, G_t, t, max_weight=MAX_WEIGHT
    )
    brier_scores_list.append((t, bs, n_eval))

# Tabela de resultados
print(f"\n{'t (anos)':<10} {'BS(t)':<12} {'N avaliável':<15} "
      f"{'G(t)':<10} {'Interpretação'}")
print("-"*80)

tempos_mostrar = [1, 3, 5, 10, 15, 20, 23]
for t, bs, n_eval in brier_scores_list:
    if t in tempos_mostrar:
        g_t = G_t[min(t, len(G_t)-1)]
        if not np.isnan(bs):
            if bs < 0.10:
                interp = "Excelente"
            elif bs < 0.15:
                interp = "Boa"
            elif bs < 0.20:
```

```
                interp = "Moderada"
            else:
                interp = "Fraca"

            print(f"{t:<10} {bs:<12.4f} {n_eval:<15,} "
                  f"{g_t:<10.4f} {interp}")
        else:
            print(f"{t:<10} {'N/A':<12} {n_eval:<15,} "
                  f"{g_t:<10.4f} {'Insuficiente'}")

# Calcular IBS
bs_values_reliable = [bs for t, bs, _ in brier_scores_list
                      if t <= max_t_reliable and not np.isnan(bs)]
ibs_reliable = np.mean(bs_values_reliable)

bs_values_all = [bs for _, bs, _ in brier_scores_list if not np.isnan(bs)]
ibs_all = np.mean(bs_values_all)

print("\n" + "="*80)
print("INTEGRATED BRIER SCORE (IBS)")
print("="*80)
print(f"IBS (t  {max_t_reliable} anos, horizonte confiável): "
      f"{ibs_reliable:.4f}")
print(f"IBS (todo período, 1-24 anos):                          "
      f"{ibs_all:.4f}")
print("="*80)

# Interpretação do IBS
if ibs_reliable < 0.10:
    interp_ibs = "Excelente"
elif ibs_reliable < 0.15:
    interp_ibs = "Boa"
elif ibs_reliable < 0.20:
    interp_ibs = "Moderada"
else:
    interp_ibs = "Fraca"

print(f"\nInterpretação (horizonte confiável): "
      f"Calibração {interp_ibs.lower()}")
print(f"\nContexto:")
print(f"  - IBS = 0.00: Calibração perfeita (ideal teórico)")
print(f"  - IBS = 0.25: Desempenho aleatório")
print(f"  - IBS = {ibs_reliable:.4f}: Redução de "
      f"{(0.25-ibs_reliable)/0.25*100:.1f}% do erro vs. acaso")

print(f"\nNota metodológica (Graf et al., 1999):")
print(f"  Análise restrita a t  {max_t_reliable} anos devido a "
      f"G({max_t_reliable}) = {G_t[max_t_reliable]:.3f} > 0.10")
print(f"  Horizontes com G(t) < 0.10 apresentam pesos IPCW instáveis")
```

```
# ============================================================================
# 15. VISUALIZAÇÃO: EVOLUÇÃO DO BRIER SCORE E G(t)
# ============================================================================

print("\n" + "="*80)
```

```
print("FIGURA 5.3: EVOLUÇÃO DO BRIER SCORE E FUNÇÃO G(t)")
print("="*80)

fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(14, 10), sharex=True)

times_bs = [t for t, _, _ in brier_scores_list]
bs_values = [bs for _, bs, _ in brier_scores_list]

# Painel A: Evolução do Brier Score
ax1.plot(times_bs, bs_values, 'o-', linewidth=2.5, markersize=7,
         color=CORES_CUSTOM[0], label='Brier Score')

ax1.axhline(y=0.25, color='red', linestyle=':', linewidth=2,
            label='Acaso (BS=0.25)', alpha=0.7)

ax1.axvline(x=max_t_reliable, color='orange', linestyle='--',
           linewidth=2.5,
           label=f'Limite confiável (t={max_t_reliable}, G(t)=0.10)',
           alpha=0.8)

ax1.set_ylabel('Brier Score', fontsize=13, fontweight='bold')
ax1.set_title('A) Evolução do Brier Score ao Longo do Tempo',
              fontsize=13, fontweight='bold', loc='left')
ax1.legend(fontsize=10, loc='best', frameon=True, shadow=True)
ax1.grid(True, alpha=0.3, linestyle='--')

# Painel B: Função G(t) e amostras disponíveis
ax2_twin = ax2.twinx()

times_g = list(range(len(G_t)))
ax2.plot(times_g, G_t, 'o-', linewidth=2.5, markersize=6,
         color=CORES_CUSTOM[4], label=r'$G(t)$ (Kaplan-Meier)')
ax2.axhline(y=0.10, color='orange', linestyle='--', linewidth=2,
           label='Limite G(t)=0.10', alpha=0.7)

n_eval_values = [n for _, _, n in brier_scores_list]
ax2_twin.plot(times_bs, n_eval_values, '^-', linewidth=2, markersize=5,
             color=CORES_CUSTOM[6], alpha=0.7, label='N Avaliável')

ax2.set_xlabel(r'Tempo $t$ (anos)', fontsize=13, fontweight='bold')
ax2.set_ylabel(r'$G(t)$ - Sobrevivência da Censura',
               fontsize=12, fontweight='bold')
ax2_twin.set_ylabel('N Avaliável', fontsize=12, fontweight='bold')
ax2.set_title('B) Função de Sobrevivência da Censura e Amostras Disponíveis',
              fontsize=13, fontweight='bold', loc='left')

lines1, labels1 = ax2.get_legend_handles_labels()
lines2, labels2 = ax2_twin.get_legend_handles_labels()
ax2.legend(lines1 + lines2, labels1 + labels2, fontsize=10,
          loc='upper right', frameon=True, shadow=True)

ax2.grid(True, alpha=0.3, linestyle='--')
ax2.set_xlim(0, 24)
ax2.set_ylim(0, 1.05)

plt.tight_layout()
plt.savefig('figura_5_3_evolucao_bs_e_gt.png', dpi=300, bbox_inches='tight')
```

```
sns.despine()
plt.show()

print("\nInterpretação:")
print("  - Linha laranja vertical: limite de confiabilidade "
      "(Graf et al., 1999)")
print("  - Painel B: G(t) decai com o tempo, aumentando pesos IPCW")
print(f"  - Após t={max_t_reliable} anos: pesos instáveis (G(t) < 0.10)")
```

```
# ============================================================================
# 16. ANÁLISE DE IMPORTÂNCIA DAS VARIÁVEIS (PERMUTATION IMPORTANCE)
# ============================================================================

print("\n" + "="*80)
print("PERMUTATION IMPORTANCE")
print("="*80)

print("Calculando importância das variáveis...")
print("(Isto pode levar alguns minutos)")

baseline_cindex = cindex
importances = []
importances_std = []

for i, feature_name in enumerate(colunas_features):
    if (i + 1) % 5 == 0:
        print(f"  Progresso: {i+1}/{len(colunas_features)} "
              f"({(i+1)/len(colunas_features)*100:.0f}%)")

    cindex_permuted_list = []

    for repeat in range(5):
        X_test_perm = X_test_scaled.copy()
        np.random.seed(SEED + repeat + i*10)
        perm_idx = np.random.permutation(len(X_test_perm))
        X_test_perm[:, i] = X_test_perm[perm_idx, i]

        model.eval()
        with torch.no_grad():
            X_tensor = torch.FloatTensor(X_test_perm).to(device)
            hazards_perm = model(X_tensor)
            survival_perm = calculate_survival_function(hazards_perm)

        risk_perm = 1 - survival_perm[:, -1]

        try:
            cindex_perm, _, _, _, _ = concordance_index_censored(
                delta_test.astype(bool), T_test.astype(float), risk_perm
            )
            cindex_permuted_list.append(cindex_perm)
        except:
            cindex_permuted_list.append(baseline_cindex)

    mean_cindex_perm = np.mean(cindex_permuted_list)
    importance = baseline_cindex - mean_cindex_perm
```

```
    std_cindex_perm = np.std(cindex_permuted_list)

    importances.append(importance)
    importances_std.append(std_cindex_perm)

importances = np.array(importances)
importances_std = np.array(importances_std)
importances_clean = np.maximum(importances, 0)

print(f"\nConcluído!")
print(f"Baseline C-index: {baseline_cindex:.4f}")
print(f"Redução máxima: {importances_clean.max():.4f}")
print(f"Variáveis com impacto positivo: "
      f"{(importances_clean > 0).sum()}/{len(importances_clean)}")
```

```
# ============================================================================
# 17. VISUALIZAÇÃO: PERMUTATION IMPORTANCE
# ============================================================================

print("\n" + "="*80)
print("FIGURA 5.4: IMPORTÂNCIA DAS VARIÁVEIS")
print("="*80)

# Criar nomes legíveis
feature_names_readable = []
for fname in colunas_features:
    fname_clean = fname.replace('_', ' ').title()
    fname_clean = fname_clean.replace('Faixa Etaria Inicio', 'Idade')
    fname_clean = fname_clean.replace('Cd4 Cat200', 'CD4')
    fname_clean = fname_clean.replace('Deteccao Cv50 ', '')
    fname_clean = fname_clean.replace('Cv Detectavel', 'CV Detectável')
    fname_clean = fname_clean.replace('Sem Dados De Cv', 'Sem Dados de CV')
    fname_clean = fname_clean.replace('Cv Indetectavel', 'CV Indetectável')
    fname_clean = fname_clean.replace('Regiao', 'Região')
    fname_clean = fname_clean.replace('Escolaridade', 'Esc.')
    fname_clean = fname_clean.replace('Sexo Mulher', 'Sexo: Mulher')
    fname_clean = fname_clean.replace('Raca', 'Raça')
    feature_names_readable.append(fname_clean)

# Ordenar por importância
importance_order = np.argsort(importances_clean)[::-1]

fig, ax = plt.subplots(figsize=(14, 10))

all_importances = importances_clean[importance_order]
all_names = [feature_names_readable[i] for i in importance_order]
all_std = importances_std[importance_order]

# Cores por magnitude
colors = []
for imp in all_importances:
    if imp > 0.015:
        colors.append(CORES_CUSTOM[0])
    elif imp > 0.008:
        colors.append(CORES_CUSTOM[2])
```

```
    elif imp > 0.003:
        colors.append(CORES_CUSTOM[3])
    else:
        colors.append(CORES_CUSTOM[12])

y_pos = np.arange(len(all_importances))

bars = ax.barh(y_pos, all_importances, xerr=all_std,
               color=colors, edgecolor='black', linewidth=0.8,
               error_kw={'linewidth': 1.2, 'ecolor': 'gray', 'capsize': 2},
               zorder=2)

ax.set_yticks(y_pos)
ax.set_yticklabels(all_names, fontsize=10)
ax.invert_yaxis()
ax.set_xlabel(r'$\Delta$ C-index', fontsize=13, fontweight='bold')
ax.set_title('Importância das Variáveis', fontsize=16,
             fontweight='bold', pad=20)
ax.grid(True, alpha=0.3, axis='x', linestyle='--')
ax.axvline(x=0, color='black', linewidth=1.5, zorder=3)

ax.axvline(x=0.015, color='black', linestyle=':', alpha=0.7,
           linewidth=2, zorder=3)
ax.axvline(x=0.008, color='black', linestyle=':', alpha=0.7,
           linewidth=2, zorder=3)
ax.axvline(x=0.003, color='black', linestyle=':', alpha=0.7,
           linewidth=2, zorder=3)

# Anotar valores no top 10
for i in range(min(10, len(all_importances))):
    if all_importances[i] > 0.001:
        ax.text(all_importances[i] + max(all_std[:10]) + 0.001, i,
                f'{all_importances[i]:.4f}',
                va='center', fontsize=8, fontweight='bold')

# Legenda
from matplotlib.patches import Patch
legend_elements = [
    Patch(facecolor=CORES_CUSTOM[0], edgecolor='black',
          label='Alto (>0.015)'),
    Patch(facecolor=CORES_CUSTOM[2], edgecolor='black',
          label='Moderado (0.008-0.015)'),
    Patch(facecolor=CORES_CUSTOM[3], edgecolor='black',
          label='Baixo (0.003-0.008)'),
    Patch(facecolor=CORES_CUSTOM[12], edgecolor='black',
          label='Mínimo (<0.003)')
]
ax.legend(handles=legend_elements, loc='lower right', fontsize=10,
          title='Nível de Impacto', title_fontsize=11,
          frameon=True, shadow=True)

plt.tight_layout()
plt.savefig('figura_5_4_permutation_importance.png',
            dpi=300, bbox_inches='tight')
sns.despine()
plt.show()

```

```python
# Top 10
print("\n" + "-"*80)
print("TOP 10 VARIÁVEIS MAIS IMPORTANTES")
print("-"*80)
print(f"{'Rank':<6} {'Variável':<45} {'Importância':<12} {'Nível'}")
print("-"*80)

for rank, idx in enumerate(importance_order[:10], 1):
    var_name = feature_names_readable[idx]
    imp = importances_clean[idx]

    if imp > 0.015:
        nivel = "Alto"
    elif imp > 0.008:
        nivel = "Moderado"
    elif imp > 0.003:
        nivel = "Baixo"
    else:
        nivel = "Mínimo"

    print(f"{rank:<6} {var_name:<45} {imp:<12.4f} {nivel}")
```

```python
# ============================================================================
# 18. PREDIÇÕES PARA PERFIS ESPECÍFICOS
# ============================================================================

print("\n" + "="*80)
print("FIGURA 5.5: CURVAS DE SOBREVIVÊNCIA POR CARACTERÍSTICAS")
print("="*80)

# Criar perfil base
perfil_base = np.zeros(input_dim)

def get_feature_idx(substring):
    """Retorna índice da feature que contém o substring."""
    for idx, name in enumerate(colunas_features):
        if substring in name:
            return idx
    return None

# Configurar perfil base
features_base_config = {
    'faixa_etaria_inicio_25-39': 'faixa_etaria_inicio_25-39',
    'cd4_cat200_>=200': 'cd4_cat200_>=200',
    'deteccao_cv50_CV indetectável': 'deteccao_cv50_CV indetectável',
    'regiao_Sudeste': 'regiao_Sudeste'
}

for feature_name, search_string in features_base_config.items():
    idx = get_feature_idx(search_string)
    if idx is not None:
        perfil_base[idx] = 1.0

# Criar figura 2x2
fig, axes = plt.subplots(2, 2, figsize=(16, 14))
```

```
fig.suptitle('Curvas de Sobrevivência Preditas',
             fontsize=16, fontweight='bold', y=0.995)

# PAINEL A: SEXO
ax1 = axes[0, 0]

perfis_sexo = {
    'Homem': perfil_base.copy(),
    'Mulher': perfil_base.copy()
}

idx_mulher = get_feature_idx('sexo_Mulher')
if idx_mulher is not None:
    perfis_sexo['Mulher'][idx_mulher] = 1.0

cores_sexo = {'Homem': CORES_CUSTOM[0], 'Mulher': CORES_CUSTOM[1]}

model.eval()
with torch.no_grad():
    for sexo, perfil in perfis_sexo.items():
        perfil_scaled = scaler.transform(perfil.reshape(1, -1))
        X_tensor = torch.FloatTensor(perfil_scaled).to(device)
        hazards = model(X_tensor)
        survival = calculate_survival_function(hazards)[0]

        ax1.plot(range(num_time_intervals), survival,
                label=sexo, linewidth=3, color=cores_sexo[sexo])

ax1.set_xlabel(r'Tempo $t$ (anos)', fontsize=12, fontweight='bold')
ax1.set_ylabel(r'$\hat{S}(t)$', fontsize=14, fontweight='bold')
ax1.set_title('A) Sexo', fontsize=13, fontweight='bold', loc='left')
ax1.legend(fontsize=11, loc='best', frameon=True, shadow=True)
ax1.grid(True, alpha=0.3, linestyle='--')
ax1.set_xlim(0, 24)
ax1.set_ylim(0, 1)

# PAINEL B: RAÇA/COR
ax2 = axes[0, 1]

perfis_raca = {
    'Branca/Amarela': perfil_base.copy(),
    'Parda': perfil_base.copy(),
    'Preta': perfil_base.copy(),
    'Indígena': perfil_base.copy()
}

ajustes_raca = {
    'Parda': 'raca_Parda',
    'Preta': 'raca_Preta',
    'Indígena': 'raca_Indígena'
}

for raca, feature_name in ajustes_raca.items():
    idx = get_feature_idx(feature_name)
    if idx is not None:
        perfis_raca[raca][idx] = 1.0

```

```
cores_raca = {
    'Branca/Amarela': CORES_CUSTOM[2],
    'Parda': CORES_CUSTOM[3],
    'Preta': CORES_CUSTOM[4],
    'Indígena': CORES_CUSTOM[5]
}

with torch.no_grad():
    for raca, perfil in perfis_raca.items():
        perfil_scaled = scaler.transform(perfil.reshape(1, -1))
        X_tensor = torch.FloatTensor(perfil_scaled).to(device)
        hazards = model(X_tensor)
        survival = calculate_survival_function(hazards)[0]

        ax2.plot(range(num_time_intervals), survival,
                label=raca, linewidth=3, color=cores_raca[raca])

ax2.set_xlabel(r'Tempo $t$ (anos)', fontsize=12, fontweight='bold')
ax2.set_ylabel(r'$\hat{S}(t)$', fontsize=14, fontweight='bold')
ax2.set_title('B) Raça/Cor', fontsize=13, fontweight='bold', loc='left')
ax2.legend(fontsize=11, loc='best', frameon=True, shadow=True)
ax2.grid(True, alpha=0.3, linestyle='--')
ax2.set_xlim(0, 24)
ax2.set_ylim(0, 1)

# PAINEL C: ESCOLARIDADE
ax3 = axes[1, 0]

perfis_escolaridade = {
    '0-7 anos': perfil_base.copy(),
    '8-11 anos': perfil_base.copy(),
    '12+ anos': perfil_base.copy()
}

ajustes_esc = {
    '8-11 anos': 'escolaridade_8-11',
    '12+ anos': 'escolaridade_12+'
}

for esc, feature_name in ajustes_esc.items():
    idx = get_feature_idx(feature_name)
    if idx is not None:
        perfis_escolaridade[esc][idx] = 1.0

cores_esc = {
    '0-7 anos': CORES_CUSTOM[6],
    '8-11 anos': CORES_CUSTOM[7],
    '12+ anos': CORES_CUSTOM[8]
}

with torch.no_grad():
    for esc, perfil in perfis_escolaridade.items():
        perfil_scaled = scaler.transform(perfil.reshape(1, -1))
        X_tensor = torch.FloatTensor(perfil_scaled).to(device)
        hazards = model(X_tensor)
        survival = calculate_survival_function(hazards)[0]

```

```
        ax3.plot(range(num_time_intervals), survival,
                label=esc, linewidth=3, color=cores_esc[esc])

ax3.set_xlabel(r'Tempo $t$ (anos)', fontsize=12, fontweight='bold')
ax3.set_ylabel(r'$\hat{S}(t)$', fontsize=14, fontweight='bold')
ax3.set_title('C) Escolaridade', fontsize=13, fontweight='bold', loc='left')
ax3.legend(fontsize=11, loc='best', frameon=True, shadow=True)
ax3.grid(True, alpha=0.3, linestyle='--')
ax3.set_xlim(0, 24)
ax3.set_ylim(0, 1)

# PAINEL D: FAIXA ETÁRIA
ax4 = axes[1, 1]

perfis_idade = {
    '<13 anos': perfil_base.copy(),
    '13-17 anos': perfil_base.copy(),
    '18-24 anos': perfil_base.copy(),
    '25-39 anos': perfil_base.copy(),
    '40-59 anos': perfil_base.copy(),
    '60+ anos': perfil_base.copy()
}

# Zerar faixa etária base
idx_25_39 = get_feature_idx('faixa_etaria_inicio_25-39')
if idx_25_39 is not None:
    for perfil in perfis_idade.values():
        perfil[idx_25_39] = 0.0

ajustes_idade = {
    '<13 anos': 'faixa_etaria_inicio_<13',
    '13-17 anos': 'faixa_etaria_inicio_13-17',
    '18-24 anos': 'faixa_etaria_inicio_18-24',
    '25-39 anos': 'faixa_etaria_inicio_25-39',
    '40-59 anos': 'faixa_etaria_inicio_40-59',
    '60+ anos': 'faixa_etaria_inicio_60+'
}

for idade, feature_name in ajustes_idade.items():
    idx = get_feature_idx(feature_name)
    if idx is not None:
        perfis_idade[idade][idx] = 1.0

cores_idade = {
    '<13 anos': CORES_CUSTOM[0],
    '13-17 anos': CORES_CUSTOM[1],
    '18-24 anos': CORES_CUSTOM[9],
    '25-39 anos': CORES_CUSTOM[10],
    '40-59 anos': CORES_CUSTOM[11],
    '60+ anos': CORES_CUSTOM[12]
}

with torch.no_grad():
    for idade, perfil in perfis_idade.items():
        perfil_scaled = scaler.transform(perfil.reshape(1, -1))
        X_tensor = torch.FloatTensor(perfil_scaled).to(device)
        hazards = model(X_tensor)
```

```
        survival = calculate_survival_function(hazards)[0]

        ax4.plot(range(num_time_intervals), survival,
                label=idade, linewidth=2.5, color=cores_idade[idade])

ax4.set_xlabel(r'Tempo $t$ (anos)', fontsize=12, fontweight='bold')
ax4.set_ylabel(r'$\hat{S}(t)$', fontsize=14, fontweight='bold')
ax4.set_title('D) Faixa Etária ao Início da TARV',
             fontsize=13, fontweight='bold', loc='left')
ax4.legend(fontsize=10, loc='best', frameon=True, shadow=True, ncol=2)
ax4.grid(True, alpha=0.3, linestyle='--')
ax4.set_xlim(0, 24)
ax4.set_ylim(0, 1)

plt.tight_layout(rect=[0, 0, 1, 0.99])
plt.savefig('figura_5_5_sobrevivencia_estratificada.png',
           dpi=300, bbox_inches='tight')
sns.despine()
plt.show()
```

```
# ============================================================================
# 19. RESUMO FINAL DOS RESULTADOS
# ============================================================================

print("\n" + "="*80)
print("RESUMO FINAL DOS RESULTADOS")
print("="*80)

print(f"\n1. DESEMPENHO DO MODELO")
print("-"*80)
print(f"C-index (global):                          "
      f"{cindex:.4f} ({interpretacao})")
print(f"IBS (t  {max_t_reliable} anos):                          "
      f"{ibs_reliable:.4f} ({interp_ibs})")

print(f"\n2. FUNDAMENTAÇÃO METODOLÓGICA")
print("-"*80)
print(f"Taxa de censura: 65.3%")
print(f"IPCW truncado: max_weight={MAX_WEIGHT} ")
print(f"Horizonte confiável: t  {max_t_reliable} anos "
      f"(Graf et al., 1999)")

print(f"\n3. CALIBRAÇÃO POR PERÍODO")
print("-"*80)

bs_curto = [bs for t, bs, _ in brier_scores_list
           if t <= 5 and not np.isnan(bs)]
ibs_curto = np.mean(bs_curto) if len(bs_curto) > 0 else np.nan

bs_medio = [bs for t, bs, _ in brier_scores_list
           if 6 <= t <= 10 and not np.isnan(bs)]
ibs_medio = np.mean(bs_medio) if len(bs_medio) > 0 else np.nan

bs_longo = [bs for t, bs, _ in brier_scores_list
           if 11 <= t <= max_t_reliable and not np.isnan(bs)]
```

```python
ibs_longo = np.mean(bs_longo) if len(bs_longo) > 0 else np.nan

print(f"Curto prazo (0-5 anos):      IBS={ibs_curto:.4f}")
print(f"Médio prazo (6-10 anos):     IBS={ibs_medio:.4f}")
print(f"Longo prazo (11-{max_t_reliable} anos):    IBS={ibs_longo:.4f}")

print(f"\n4. VARIÁVEIS MAIS IMPORTANTES (TOP 5)")
print("-"*80)
for rank, idx in enumerate(importance_order[:5], 1):
    var_name = feature_names_readable[idx]
    imp = importances_clean[idx]
    print(f"{rank}. {var_name:<45} {imp:.4f}")

print(f"\n5. ARQUITETURA DO MODELO")
print("-"*80)
print(f"Camadas ocultas: L=2 ({hidden_dim1}, {hidden_dim2} neurônios)")
print(f"Dropout: {dropout_rate}")
print(f"Parâmetros: {total_params:,}")
print(f"Epochs treinadas: {len(train_losses)}")

print("\n" + "="*80)
print("ANÁLISE CONCLUÍDA")
print("="*80)
print("\nFiguras geradas:")
print("  - figura_5_1_curva_aprendizado.png")
print("  - figura_5_2_distribuicao_risco.png")
print("  - figura_5_3_evolucao_bs_e_gt.png")
print("  - figura_5_4_permutation_importance.png")
print("  - figura_5_5_sobrevivencia_estratificada.png")

print("\nModelo salvo:")
print("  - best_nnet_survival_model.pth")
print("\n" + "="*80)

```

Referências

JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS (UNAIDS). **AIDS, Crisis and the Power to Transform: UNAIDS Global AIDS Update 2025**. Geneva, 2025. OCLC: 1535395878. Disponível em: <https://unaids.org.br/relatorios-e-publicacoes>. Acesso em: 14 out. 2025.

GALVÃO, J. Access to antiretroviral drugs in Brazil. **Lancet (London, England)**, v. 360, n. 9348, p. 1862–1865, 7 dez. 2002. ISSN 0140-6736. DOI: <10.1016/S0140-6736(02)11775-2>. Acesso em: 15 out. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **Protocolo Clínico e Diretrizes Terapêuticas para Manejo da Infecção pelo HIV em Adultos: Módulo 1: Tratamento**. Brasília, DF: Ministério da Saúde, 8 out. 2024. Acesso em: 6 jun. 2025.

CENTERS FOR DISEASE CONTROL (CDC). Pneumocystis pneumonia–Los Angeles. **MMWR. Morbidity and mortality weekly report**, v. 30, n. 21, p. 250–252, 5 jun. 1981. ISSN 0149-2195.

BARRÉ-SINOUSSI, F.; ROSS, A. L.; DELFRAISSY, J.-F. Past, present and future: 30 years of HIV research. **Nature Reviews. Microbiology**, v. 11, n. 12, p. 877–883, dez. 2013. ISSN 1740-1534. DOI: <10.1038/nrmicro3132>.

CENTERS FOR DISEASE CONTROL (CDC). Update on acquired immune deficiency syndrome (AIDS)–United States. **MMWR. Morbidity and mortality weekly report**, v. 31, n. 37, p. 507–508, 513–514, 24 set. 1982. ISSN 0149-2195.

BARRÉ-SINOUSSI, F. *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). **Science (New York, N.Y.)**, v. 220, n. 4599, p. 868–871, 20 maio 1983. ISSN 0036-8075. DOI: <10.1126/science.6189183>.

GALLO, R. C. The Early Years of HIV/AIDS. **Science**, American Association for the Advancement of Science, v. 298, n. 5599, p. 1728–1730, 29 nov. 2002. DOI: <10.1126/science.1078050>. Acesso em: 23 dez. 2025.

ALEXANDER, T. S. Human Immunodeficiency Virus Diagnostic Testing: 30 Years of Evolution. **Clinical and vaccine immunology: CVI**, v. 23, n. 4, p. 249–253, abr. 2016. ISSN 1556-679X. DOI: <10.1128/CVI.00053-16>.

PALELLA, F. J. *et al.* Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. **The New England Journal of Medicine**, v. 338, n. 13, p. 853–860, 26 mar. 1998. ISSN 0028-4793. DOI: <10.1056/NEJM199803263381301>.

HO, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. **Nature**, v. 373, n. 6510, p. 123–126, 12 jan. 1995. ISSN 0028-0836. DOI: <10.1038/373123a0>.

SAMJI, H. *et al.* Closing the Gap: Increases in Life Expectancy among Treated HIV-Positive Individuals in the United States and Canada. **PLOS ONE**, Public Library of Science, v. 8, n. 12, e81355, 18 dez. 2013. ISSN 1932-6203. DOI: <10.1371/journal.pone.0081355>. Acesso em: 23 dez. 2025.

SCHALKWYK, C. van *et al.* Updated Data and Methods for the 2023 UNAIDS HIV Estimates. **JAIDS Journal of Acquired Immune Deficiency Syndromes**, v. 95, n. 1, e1, 2024. ISSN 1525-4135. DOI: <10.1097/QAI.0000000000003344>. Acesso em: 31 jan. 2026.

BRASIL. MINISTÉRIO DA SAÚDE. **Boletim Epidemiológico - HIV e Aids (2024)**. Brasília, DF, 2024. Disponível em: <https://www.gov.br/aids/pt-br/central-de-conteudo/boletins-epidemiologicos/2024/boletim_hiv_aids_2024e.pdf/view>. Acesso em: 16 jun. 2025.

HAMMAN, E. A AIDS no Brasil (1982-1992). **Cadernos de Saúde Pública**, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 10, p. 400–401, 1994. ISSN 0102-311X, 1678-4464. DOI: <https://doi.org/10.1590/S0102-311X1994000300021>. Acesso em: 23 dez. 2025.

BRASIL. Lei nº 9.313/1996. Versão portuguesa. **Diário Oficial da União**, Lei nº 9.313/1996, 13 nov. 1996. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l9313.htm>. Acesso em: 15 out. 2025.

CASSIER, M.; CORREA, M. Patents, Innovation and Public Health: Brazilian Public-Sector Laboratories' Experience in Copying AIDS Drugs. *In:* ECONOMICS of AIDS and Access to HIV/AIDS Care in Developing Countries. Issues and Challenge, [*s. l.: s. n.*], 2003. Disponível em: <https://shs.hal.science/halshs-02162784>. Acesso em: 23 dez. 2025.

RODRIGUES, W. C. V.; SOLER, O. Compulsory licensing of efavirenz in Brazil in 2007: contextualization. **Revista Panamericana De Salud Publica**, v. 26, n. 6, p. 553–559, dez. 2009. ISSN 1020-4989. DOI: <10.1590/s1020-49892009001200012>.

SILVA, G. J. d. *et al.* Suppression of HIV in the first 12 months of antiretroviral therapy: a comparative analysis of dolutegravir- and efavirenz-based regimens. **Einstein (Sao Paulo, Brazil)**, v. 21, eao0156, 2023. ISSN 2317-6385. DOI: <10.31744/einstein˙journal/2023AO0156>.

PARKER, R.; AGGLETON, P. HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action. **Social Science & Medicine (1982)**, v. 57, n. 1, p. 13–24, jul. 2003. ISSN 0277-9536. DOI: <10.1016/s0277-9536(02)00304-0>.

SHILTS, R. **And the Band Played on: Politics, People, and the AIDS Epidemic**. New York: Griffin, 2007. 630 p. ISBN 978-0-312-37463-1.

BOGHARDT, T. Operation INFEKTION: Soviet Bloc Intelligence and Its AIDS Disinformation Campaign. *In:* Disponível em: <https://www.semanticscholar.org/paper/Operation-INFEKTION%3A-Soviet-Bloc-Intelligence-and-Boghardt/0955bd119fd918b50773d2ecf8d0ad6648bf2e77>. Acesso em: 23 dez. 2025.

BOGART, L. M.; THORBURN, S. Are HIV/AIDS conspiracy beliefs a barrier to HIV prevention among African Americans? **Journal of Acquired Immune Deficiency Syndromes (1999)**, v. 38, n. 2, p. 213–218, 1 fev. 2005. ISSN 1525-4135. DOI: <10.1097/00126334-200502010-00014>.

SHARP, P. M.; HAHN, B. H. Origins of HIV and the AIDS pandemic. **Cold Spring Harbor Perspectives in Medicine**, v. 1, n. 1, a006841, set. 2011. ISSN 2157-1422. DOI: <10.1101/cshperspect.a006841>.

NATTRASS, N. **Mortal Combat: AIDS Denialism and the Struggle for Antiretrovirals in South Africa**. Scottsville, South Africa: University Of KwaZulu-Natal Press, 2007. 269 p. ISBN 978-1-86914-132-5.

CHIGWEDERE, P. *et al.* Estimating the lost benefits of antiretroviral drug use in South Africa. **Journal of Acquired Immune Deficiency Syndromes (1999)**, v. 49, n. 4, p. 410–415, 1 dez. 2008. ISSN 1525-4135. DOI: <10.1097/qai.0b013e31818a6cd5>.

SMITH, T. C. HIV Denial in the COVID Era. **AIDS and Behavior**, v. 29, n. 1, p. 309–316, 1 jan. 2025. ISSN 1573-3254. DOI: <10.1007/s10461-024-04528-3>. Acesso em: 1 fev. 2026.

KALICHMAN, S. C.; EATON, L.; CHERRY, C. "There is no proof that HIV causes AIDS": AIDS denialism beliefs among people living with HIV/AIDS. **Journal of Behavioral Medicine**, v. 33, n. 6, p. 432–440, dez. 2010. ISSN 1573-3521. DOI: <10.1007/s10865-010-9275-7>.

COSTA, L. F. d. *et al.* Fatores psicossociais envolvidos na adesão ao tratamento do HIV/AIDS em adultos: revisão integrativa da literatura. **Saúde Coletiva (Barueri)**, v. 11, n. 61, p. 4990–5005, 1 fev. 2021. Number: 61. ISSN 2675-0244. DOI: <10.36489/saudecoletiva.2021v11i61p4990-5005>. Acesso em: 7 jun. 2025.

CUNHA, A. P. D. *et al.* Fatores associados à interrupção do tratamento antirretroviral de pessoas que vivem com HIV/aids em municípios brasileiros entre 2019 e 2022. **Revista Brasileira de Epidemiologia**, v. 28, e250015, 2025. ISSN 1980-5497, 1415-790X. DOI: <10.1590/1980-549720250015.2>. Acesso em: 7 jun. 2025.

BRITO, A.; CASTILHO, E. A.; SZWARCWALD, C. L. AIDS and HIV infection in Brazil: a multifaceted epidemic. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 34, p. 61–74, 2001. ISSN 0037-8682. DOI: <10.1590/S0037-86822001000700008>. Acesso em: 16 jun. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **Protocolo Clínico e Diretrizes Terapêuticas para Profilaxia Pós-Exposição (PEP) de Risco à Infecção pelo HIV, IST e Hepatites Virais**. Brasília, DF, 2024. Disponível em: <https://www.gov.br/aids/pt-br/central-de-conteudo/pcdts/2021/hiv-aids/prot_clinico_diretrizes_terap_pep_-risco_infeccao_hiv_ist_hv_2021.pdf/view>. Acesso em: 5 dez. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **Protocolo Clínico e Diretrizes Terapêuticas para Profilaxia Pré-Exposição (PrEP) de Risco à Infecção pelo HIV**. Brasília, DF, 2025. Disponível em: <https://www.gov.br/aids/pt-br/central-de-conteudo/pcdts/protocolo-clinico-e-diretrizes-terapeuticas-para-profilaxia-pre-exposicao-prep-oral-a-infeccao-pelo-hiv.pdf/view>. Acesso em: 5 dez. 2025.

GRANT, R. M. *et al.* Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men. **New England Journal of Medicine**, Massachusetts Medical Society, v. 363, n. 27, p. 2587–2599, 30 dez. 2010. _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1011205. ISSN 0028-4793. DOI: <10.1056/NEJMoa1011205>. Acesso em: 23 dez. 2025.

MCCORMACK, S. *et al.* Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label

randomised trial. **Lancet (London, England)**, v. 387, n. 10013, p. 53–60, 2 jan. 2016. ISSN 1474-547X. DOI: <10.1016/S0140-6736(15)00056-2>.

BRASIL. MINISTÉRIO DA SAÚDE. **Relatório de Monitoramento de Profilaxias Pré e Pós-Exposição ao HIV 2023**. Brasília, DF, 2024. Disponível em: <https://www.gov.br/aids/pt-br/central-de-conteudo/publicacoes/2024/relatorio-de-profilaxias-prep-e-pep-2022.pdf/view>. Acesso em: 5 dez. 2025.

RODGER, A. J. *et al.* Sexual Activity Without Condoms and Risk of HIV Transmission in Serodifferent Couples When the HIV-Positive Partner Is Using Suppressive Antiretroviral Therapy. **JAMA**, v. 316, n. 2, p. 171–181, 12 jul. 2016. ISSN 1538-3598. DOI: <10.1001/jama.2016.5148>.

COHEN, M. S. *et al.* Prevention of HIV-1 Infection with Early Antiretroviral Therapy. **New England Journal of Medicine**, Massachusetts Medical Society, v. 365, n. 6, p. 493–505, 11 ago. 2011. _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1105243. ISSN 0028-4793. DOI: <10.1056/NEJMoa1105243>. Acesso em: 23 dez. 2025.

LANDOVITZ, R. J. *et al.* Cabotegravir for HIV Prevention in Cisgender Men and Transgender Women. **The New England Journal of Medicine**, v. 385, n. 7, p. 595–608, 12 ago. 2021. ISSN 1533-4406. DOI: <10.1056/NEJMoa2101016>.

DELANY-MORETLWE, S. *et al.* Cabotegravir for the prevention of HIV-1 in women: results from HPTN 084, a phase 3, randomised clinical trial. **The Lancet**, Elsevier, v. 399, n. 10337, p. 1779–1789, 7 maio 2022. ISSN 0140-6736, 1474-547X. DOI: <10.1016/S0140-6736(22)00538-4>. Acesso em: 5 dez. 2025.

BEKKER, L.-G. *et al.* Twice-Yearly Lenacapavir or Daily F/TAF for HIV Prevention in Cisgender Women. **The New England Journal of Medicine**, v. 391, n. 13, p. 1179–1192, 3 out. 2024. ISSN 1533-4406. DOI: <10.1056/NEJMoa2407001>.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis: A Self-Learning Text**. 3. ed. New York Dordrecht Heidelberg London: Springer, 2020. 700 p. ISBN 978-1-4939-5018-8.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. 2. ed. São Paulo, SP: Blucher, 20 fev. 2024. 362 p. ISBN 978-85-212-2199-9.

COX, D. R. Regression Models and Life-Tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 34, n. 2, p. 187–202, 1972. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x. ISSN 2517-6161. DOI: <10.1111/j.2517-6161.1972.tb00899.x>. Acesso em: 14 out. 2025.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data: 360**. 2. ed. Hoboken, N.J: Wiley-Interscience, 2002. 462 p. ISBN 978-0-471-36357-6.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. New York, NY: Springer, 2009. (Springer Series in Statistics). ISBN 978-0-387-84857-0 978-0-387-84858-7. DOI: <10.1007/978-0-387-84858-7>. Acesso em: 14 out. 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, Massachusetts: The Mit Press, 2016. 775 p. ISBN 978-0-262-03561-3. Disponível em: <http://www.deeplearningbook.org>.

KATZMAN, J. *et al.* DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. **BMC Medical Research Methodology**, v. 18, n. 1, dez. 2018. ISSN 1471-2288. DOI: <10.1186/s12874-018-0482-1>. Acesso em: 10 jul. 2025.

KVAMME, H.; BORGAN, Ø.; SCHEEL, I. **Time-to-Event Prediction with Neural Networks and Cox Regression**. [*S. l.*]: arXiv, 13 set. 2019. DOI: <10.48550/arXiv.1907.00825>. Acesso em: 14 out. 2025.

GENSHEIMER, M. F.; NARASIMHAN, B. A scalable discrete-time survival model for neural networks. **PeerJ**, PeerJ Inc., v. 7, e6257, 25 jan. 2019. ISSN 2167-8359. DOI: <10.7717/peerj.6257>. Acesso em: 14 out. 2025.

BUSTAMANTE-TEIXEIRA, M. T.; FAERSTEIN, E.; LATORRE, M. d. R. Técnicas de análise de sobrevida. **Cadernos de Saúde Pública**, v. 18, n. 3, 1 jun. 2002. ISSN 1678-4464. DOI: <10.1590/S0102-311X2002000300008>. Acesso em: 10 nov. 2025.

NOGUEIRA, M. C. *et al.* Disparidade racial na sobrevivência em 10 anos para o câncer de mama: uma análise de mediação usando abordagem de respostas potenciais. **Cadernos de Saúde Pública**, v. 34, e00211717, 6 set. 2018. Editora: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz. ISSN 0102-311X, 0102-311X, 1678-4464. DOI: <https://doi.org/10.1590/0102-311X00211717>. Acesso em: 10 nov. 2025.

HARTZ, Z. M. d. A. **Avaliação em saúde: dos modelos conceituais à prática na análise da implantação de programas**. [*S. l.*]: Editora FIOCRUZ, 1997. 132 p. ISBN 85-85676-36-1. Disponível em: <https://books.scielo.org/id/3zcft>. Acesso em: 10 nov. 2025.

XAVIER, M. E. F. **Análise comparativa do desempenho de aerogeradores através do estudo das curvas de potência**. 27 dez. 2022. TCC – Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco, Pernambuco. Disponível em: <https://repositorio.ifpe.edu.br/xmlui/handle/123456789/818>. Acesso em: 10 nov. 2025.

MELLO, L. F. d.; SATHLER, D. A demografia ambiental e a emergência dos estudos sobre população e consumo. **Revista Brasileira de Estudos de População**, v. 32, p. 357–380, 2015. Editora: Associação Brasileira de Estudos Populacionais. ISSN 0102-3098, 1980-5519. DOI: <https://doi.org/10.1590/S0102-30982015000000020>. Acesso em: 10 nov. 2025.

BACHINI, N.; CHICARINO, T. S. Os métodos quantitativos, por cientistas sociais brasileiros: entrevistas com Nelson do Valle Silva e Jerônimo Muniz. **Sociedade e Estado**, Departamento de Sociologia da Universidade de Brasília, v. 33, p. 251–279, 2018. ISSN 0102-6992, 1980-5462. DOI: <https://doi.org/10.1590/s0102-699220183301010>. Acesso em: 10 nov. 2025.

JUNIOR, M.; FERREIRA, A. A. Uma análise da progressão dos alunos cotistas sob a primeira ação afirmativa brasileira no ensino superior: o caso da Universidade do Estado do Rio de Janeiro. **Ensaio: Avaliação e Políticas Públicas em Educação**, Fundação CESGRANRIO, v. 22, p. 31–56, 2014. ISSN 0104-4036, 1809-4465. DOI: <https://doi.org/10.1590/S0104-40362014000100003>. Acesso em: 10 nov. 2025.

KLITZKE, M.; CARVALHAES, F. FATORES ASSOCIADOS À EVASÃO DE CURSO NA UFRJ: UMA ANÁLISE DE SOBREVIVÊNCIA. **Educação em Revista**, Faculdade de Educação da Universidade Federal de Minas Gerais, v. 39, e37576, 2023. ISSN 0102-4698, 1982-6621. DOI: <https://doi.org/10.1590/0102-469837576>. Acesso em: 10 nov. 2025.

SILVA, G. P. d. Análise de evasão no ensino superior: uma proposta de diagnóstico de seus determinantes. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 18, p. 311–333, 2013. Editora: Publicação da Rede de Avaliação Institucional da Educação Superior (RAIES), da Universidade Estadual de Campinas (UNICAMP) e da Universidade de Sorocaba (UNISO). ISSN 1414-4077, 1982-5765. DOI: <https://doi.org/10.1590/S1414-40772013000200005>. Acesso em: 10 nov. 2025.

GRAUNT, J. **Natural and Political Observations Made upon the Bills of Mortality**. [*S. l.: s. n.*], 1662. Disponível em: <https://collections.nlm.nih.gov/catalog/nlm:nlmuid-2356017R-bk>. Acesso em: 11 nov. 2025.

HALLEY, E. An Estimate of the Degrees of the Mortality of Mankind, Drawn from Curious Tables of the Births and Funerals at the City of Breslaw; With an Attempt to Ascertain the Price of Annuities upon Lives. **Philosophical Transactions of the Royal Society of London**, Royal Society, v. 17, p. 596–610, 1693. DOI: <10.1098/rstl.1693.0007>. Acesso em: 11 nov. 2025.

HALD, A. **A History of Probability and Statistics and Their Applications before 1750**. New York: Wiley, 1990. 611 p. ISBN 978-0-471-72517-6.

MACEDO, M. A. d. S.; SILVA, F. d. F. d.; SANTOS, R. M. Análise do mercado de seguros no Brasil: uma visão do desempenho organizacional das seguradoras no ano de 2003. **Revista Contabilidade & Finanças**, v. 17, p. 88–100, 2006. Editora: Universidade de São Paulo, Faculdade de Economia, Administração, Contabilidade e Atuária, Departamento de Contabilidade e Atuária - Cidade Universitária. ISSN 1519-7077, 1808-057X. DOI: <https://doi.org/10.1590/S1519-70772006000500007>. Acesso em: 10 nov. 2025.

BOWERS, N. L. *et al.* **Actuarial Mathematics**. [*S. l.*]: Society of Actuaries, 1997. 753 p. ISBN 978-0-938959-46-5.

DIRICK, L.; CLAESKENS, G.; BAESENS, B. Time to default in credit scoring using survival analysis: a benchmark study. **Journal of the Operational Research Society**, v. 68, n. 6, p. 652–665, 1 jun. 2017. ISSN 1476-9360. DOI: <10.1057/s41274-016-0128-9>. Acesso em: 10 nov. 2025.

NARAIN, B. Survival analysis and the credit-granting decision. *In:* THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. (ed.). **Readings in Credit Scoring: Foundations, developments, and aims**. [*S. l.*]: Oxford University Press, 8 jul. 2004. ISBN 978-0-19-852797-8. DOI: <10.1093/oso/9780198527978.003.0022>. Acesso em: 10 nov. 2025.

CAMARGO, P. O. A evolução recente do setor bancário no Brasil. Editora UNESP, p. 322, 2009. DOI: <https://doi.org/10.7476/9788579830396>. Acesso em: 10 nov. 2025.

PAULA, L. F. d.; OREIRO, J. L.; BASILIO, F. A. C. Estrutura do setor bancário e o ciclo recente de expansão do crédito: o papel dos bancos públicos federais. **Nova Economia**, Nova Economia, v. 23, p. 473–520, 2013. ISSN 0103-6351, 1980-5381. DOI: <https://doi.org/10.1590/S0103-63512013000300001>. Acesso em: 10 nov. 2025.

BELLOTTI, T.; CROOK, J. Credit scoring with macroeconomic variables using survival analysis. **Journal of the Operational Research Society**, v. 60, n. 12, p. 1699–1707, 1 dez. 2009. ISSN 0160-5682. DOI: <10.1057/jors.2008.130>. Acesso em: 10 nov. 2025.

MILHAUD, X.; DUTANG, C. Lapse tables for lapse risk management in insurance: a competing risk approach. **European Actuarial Journal**, v. 8, n. 1, p. 97–126, mar. 2018. Editora: Springer. DOI: <10.1007/s13385-018-0165-7>. Acesso em: 10 nov. 2025.

ELING, M.; KIESENBAUER, D. What Policy Features Determine Life Insurance Lapse? An Analysis of the German Market. **Journal of Risk and Insurance**, v. 81, n. 2, p. 241–269, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6975.2012.01504.x. ISSN 1539-6975. DOI: <10.1111/j.1539-6975.2012.01504.x>. Acesso em: 10 nov. 2025.

KAPLAN, E. L.; MEIER, P. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457–481, 1 jun. 1958. ISSN 0162-1459. DOI: <10.1080/01621459.1958.10501452>. Acesso em: 13 ago. 2025.

GREENWOOD, M. **A Report on the Natural Duration of Cancer**. [*S. l.*]: H.M. Stationery Office, 1926. (Reports on public health and medical subjects).

FLEMING, T. R.; HARRINGTON, D. P. Wiley Series in Probability and Statistics. *In:* COUNTING Processes and Survival Analysis. [*S. l.*]: John Wiley & Sons, Ltd, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118150672.scard. p. 431–438. ISBN 978-1-118-15067-2. DOI: <10.1002/9781118150672.scard>. Acesso em: 10 nov. 2025.

NELSON, W. Hazard Plotting for Incomplete Failure Data. **Journal of Quality Technology**, Taylor & Francis, v. 1, n. 1, p. 27–52, 1 jan. 1969. _eprint: https://doi.org/10.1080/00224065.1969.11980344. ISSN 0022-4065. DOI: <10.1080/00224065.1969.11980344>. Acesso em: 10 nov. 2025.

AALEN, O. Nonparametric Inference for a Family of Counting Processes. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 4, p. 701–726, jul. 1978. ISSN 0090-5364, 2168-8966. DOI: <10.1214/aos/1176344247>. Acesso em: 10 nov. 2025.

ANDERSEN, P. K. *et al.* **Statistical Models Based on Counting Processes**. New York, NY: Springer US, 1993. (Springer Series in Statistics). ISBN 978-0-387-94519-4 978-1-4612-4348-9. DOI: <10.1007/978-1-4612-4348-9>. Acesso em: 10 nov. 2025.

SWEET, A. On the hazard rate of the lognormal distribution. **IEEE Transactions on Reliability**, v. 39, n. 3, p. 325–328, ago. 1990. ISSN 1558-1721. DOI: <10.1109/24.103012>. Acesso em: 27 nov. 2025.

FELLER, W. **An Introduction to Probability Theory and Its Applications**. 3rd. New York, NY: John Wiley & Sons, 1968. v. 1. 510 p. ISBN 978-0-471-25711-0.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **J. Mach. Learn. Res.**, v. 13, p. 281–305, null 1 fev. 2012. ISSN 1532-4435.

HARRELL, F. E. *et al.* Evaluating the yield of medical tests. **JAMA**, v. 247, n. 18, p. 2543–2546, 14 maio 1982. ISSN 0098-7484.

GRAF, E. *et al.* Assessment and comparison of prognostic classification schemes for survival data. **Statistics in Medicine**, v. 18, n. 17, p. 2529–2545, 15 set. 1999. ISSN 0277-6715. DOI: <10.1002/(sici)1097-0258(19990915/30)18:17/18¡2529::aid-sim274¿3.0.co;2-5>.

STEYERBERG, E. W. *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. **Epidemiology (Cambridge, Mass.)**, v. 21, n. 1, p. 128–138, jan. 2010. ISSN 1531-5487. DOI: <10.1097/EDE.0b013e3181c30fb2>.

PENCINA, M. J.; D'AGOSTINO, R. B.; STEYERBERG, E. W. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. **Statistics in Medicine**, v. 30, n. 1, p. 11–21, 15 jan. 2011. ISSN 1097-0258. DOI: <10.1002/sim.4085>.

UNO, H. *et al.* On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. **Statistics in medicine**, v. 30, n. 10, p. 1105–1117, 10 maio 2011. ISSN 0277-6715. DOI: <10.1002/sim.4154>. Acesso em: 27 ago. 2025.

GERDS, T. A.; SCHUMACHER, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. **Biometrical Journal. Biometrische Zeitschrift**, v. 48, n. 6, p. 1029–1040, dez. 2006. ISSN 0323-3847. DOI: <10.1002/bimj.200610301>.

HOUWELINGEN, H. C. van. Validation, calibration, revision and combination of prognostic survival models. **Statistics in Medicine**, v. 19, n. 24, p. 3401–3415, 30 dez. 2000. ISSN 0277-6715. DOI: <10.1002/1097-0258(20001230)19:24¡3401::aid-sim554¿3.0.co;2-2>.

GUO, C. *et al.* **On Calibration of Modern Neural Networks**. [*S. l.*]: arXiv, 3 ago. 2017. DOI: <10.48550/arXiv.1706.04599>. Acesso em: 23 nov. 2025.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 out. 2001. ISSN 1573-0565. DOI: <10.1023/A:1010933404324>. Acesso em: 13 jun. 2025.

FISHER, A.; RUDIN, C.; DOMINICI, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. **Journal of Machine Learning Research**, v. 20, n. 177, p. 1–81, 2019. ISSN 1533-7928. Disponível em: <http://jmlr.org/papers/v20/18-760.html>. Acesso em: 10 nov. 2025.

STROBL, C. *et al.* Conditional variable importance for random forests. **BMC Bioinformatics**, v. 9, n. 1, p. 307, 11 jul. 2008. ISSN 1471-2105. DOI: <10.1186/1471-2105-9-307>. Acesso em: 10 nov. 2025.

HOOKER, G.; MENTCH, L.; ZHOU, S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. **Statistics and Computing**, v. 31, n. 6, p. 82, 29 out. 2021. ISSN 1573-1375. DOI: <10.1007/s11222-021-10057-z>. Acesso em: 10 nov. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **Painel Integrado de Monitoramento do Cuidado do HIV e da Aids**. Departamento de HIV, Aids, Tuberculose, Hepatites Virais e Infecções Sexualmente Transmissíveis. 2024. Disponível em: <https://www.gov.br/aids/pt-br/indicadores-epidemiologicos/painel-de-monitoramento/painel-integrado-de-monitoramento-do-cuidado-do-hiv>. Acesso em: 7 jun. 2025.

MELO, M. C. d. *et al.* Sobrevida de pacientes com aids e associação com escolaridade e raça/cor da pele no Sul e Sudeste do Brasil: estudo de coorte, 1998-1999. **Epidemiologia e Serviços de Saúde**, Secretaria de Vigilância em Saúde e Ambiente - Ministério da Saúde do Brasil, v. 28, e2018047, 8 abr. 2019. ISSN 1679-4974, 2237-9622. DOI: <https://doi.org/10.5123/S1679-49742019000100012>. Acesso em: 13 jun. 2025.

GUERRA, C. P. P.; SEIDL, E. M. F. Adesão em HIV/AIDS: estudo com adolescentes e seus cuidadores primários. **Psicologia em Estudo**, Universidade Estadual de Maringá, v. 15, p. 781–789, dez. 2010. ISSN 1413-7372, 1807-0329. Disponível em: <https://www.scielo.br/j/pe/a/sHz6CRWjc8BnYJHFdhp7m9h/?lang=pt>. Acesso em: 7 jun. 2025.

HHS PANEL ON ANTIRETROVIRAL GUIDELINES FOR ADULTS AND ADOLESCENTS. **Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV**. Rockville (MD): US Department of Health e Human Services (HHS), 12 set. 2024. (ClinicalInfo.HIV.gov [Internet]). Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK586306/>. Acesso em: 16 jun. 2025.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. 2024. Disponível em: <https://www.r-project.org/>. Acesso em: 7 jun. 2025.

POSIT SCIENCE. **Hello, Quarto**. Quarto Documentation. 2023. Disponível em: <https://quarto.org/docs/get-started/hello/rstudio.html>. Acesso em: 7 jun. 2025.

KLUYVER, T. *et al.* Jupyter Notebooks—a publishing format for reproducible computational workflows. *In:* IOS Press. [*S. l.: s. n.*], 1 jun. 2016. p. 87–90. DOI: <10.3233/978-1-61499-649-1-87>. Acesso em: 28 nov. 2025.

APACHE SOFTWARE FOUNDATION. **Apache Parquet**. Apache Parquet. 2024. Disponível em: <https://parquet.apache.org/>. Acesso em: 13 jun. 2025.

MOTHERDUCK. **Why Choose the Parquet Table File Format**. MotherDuck. 2023. Disponível em: <https://motherduck.com/learn-more/why-choose-parquet-table-file-format/>. Acesso em: 13 jun. 2025.

FAJARDO, O. **Pyreadr**. [*S. l.: s. n.*], dez. 2018. DOI: <10.5281/zenodo.7110170>. Acesso em: 28 nov. 2025.

CRAMERI, F. **Scientific colour maps**. Em colaboração com Grace E. Shephard. [*S. l.*]: Zenodo, 5 out. 2023. Disponível em: <https://zenodo.org/record/1243862>. Acesso em: 7 jun. 2025.

CRAMERI, F.; SHEPHARD, G. E.; HERON, P. J. The misuse of colour in science communication. **Nature Communications**, Nature Publishing Group, v. 11, n. 1, p. 5444, 28 out. 2020. ISSN 2041-1723. DOI: <10.1038/s41467-020-19160-7>. Acesso em: 17 jun. 2025.

TSITSULIN, A. **Optimal qualitative colour palettes**. Anton Tsitsulin – Blog. Section: blog. 2019. Disponível em: <http://tsitsul.in/blog/coloropt/>. Acesso em: 7 jun. 2025.

CHALITA, L. V. A. S.; COLOSIMO, E. A.; DEMÉTRIO, C. G. B. Likelihood Approximations and Discrete Models for Tied Survival Data. **Communications in Statistics - Theory and Methods**, v. 31, n. 7, p. 1215–1229, 23 jul. 2002. ISSN 0361-0926. DOI: <10.1081/STA-120004920>. Acesso em: 7 jun. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **Indicadores e Dados Básicos do HIV/AIDS nos Municípios Brasileiros**. DATHI – Indicadores HIV/AIDS. 2022. Disponível em: <https://indicadores.aids.gov.br//>. Acesso em: 7 jun. 2025.

MARTINS, P. d. S. **Imputação de dados faltantes**. 2017. TCC – Universidade Federal Fluminense, Niterói. Accepted: 2020-09-02T20:04:44Z. Disponível em: <https://app.uff.br/riuff/handle/1/14853>. Acesso em: 6 jun. 2025.

STEKHOVEN, D. J. **missForest: Nonparametric Missing Value Imputation using Random Forest**. [*S. l.: s. n.*], 14 abr. 2022. Disponível em: <https://cran.r-project.org/web/packages/missForest/index.html>. Acesso em: 8 jul. 2025.

STEKHOVEN, D. J.; BÜHLMANN, P. MissForest—non-parametric missing value imputation for mixed-type data. **Bioinformatics**, v. 28, n. 1, p. 112–118, 2012. DOI: <10.1093/bioinformatics/btr597>. Acesso em: 13 jun. 2025.

ALBU, E. *et al.* **missForestPredict – Missing data imputation for prediction settings**. [*S. l.*]: arXiv, 2 jul. 2024. DOI: <10.48550/arXiv.2407.03379>. Acesso em: 13 jun. 2025.

WALJEE, A. K. *et al.* Comparison of imputation methods for missing laboratory data in medicine. **BMJ Open**, British Medical Journal Publishing Group, v. 3, n. 8, e002847, 1 ago. 2013. ISSN 2044-6055, 2044-6055. DOI: <10.1136/bmjopen-2013-002847>. Acesso em: 16 jun. 2025.

BIANCHI, B. Application of the MissForest algorithm for imputation in the Survey on Income and Living Conditions. *In:* EXPERT MEETING ON STATISTICAL DATA EDITING. **United Nations Economic Commission for Europe – Conference of European Statisticians**. Virtual: UNECE, 2022. Disponível em: <https://unece.org/sites/default/files/2022-10/SDE2022_S4_Switzerland_Bianchi_AD.pdf>.

AZUR, M. J. *et al.* Multiple imputation by chained equations: what is it and how does it work? **International Journal of Methods in Psychiatric Research**, v. 20, n. 1, p. 40–49, 24 fev. 2011. ISSN 1049-8931. DOI: <10.1002/mpr.329>. Acesso em: 6 jun. 2025.

MENEZES, E. G. *et al.* Fatores associados à não adesão dos antirretrovirais em portadores de HIV/AIDS. **Acta Paulista de Enfermagem**, Escola Paulista de Enfermagem, Universidade Federal de São Paulo, v. 31, p. 299–304, jun. 2018. ISSN 0103-2100, 1982-0194. DOI: <https://doi.org/10.1590/1982-0194201800042>. Acesso em: 7 jun. 2025.

AZEVEDO, A. L. M. d. S. **IBGE - Educa — Jovens**. IBGE Educa Jovens. 2022. Disponível em: <https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18319-cor-ou-raca.html>. Acesso em: 16 jun. 2025.

GRANGEIRO, A.; ESCUDER, M. M. L.; CASTILHO, E. A. d. A epidemia de AIDS no Brasil e as desigualdades regionais e de oferta de serviço. **Cadernos de Saúde Pública**, v. 26, p. 2355–2367, dez. 2010. Editora: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz. ISSN 0102-311X, 1678-4464. DOI: <https://doi.org/10.1590/S0102-311X2010001200014>. Acesso em: 16 jun. 2025.

CUNHA, A. P. d.; CRUZ, M. M. d.; PEDROSO, M. Análise da tendência da mortalidade por HIV/AIDS segundo características sociodemográficas no Brasil, 2000 a 2018. **Ciência & Saúde Coletiva**, v. 27, p. 895–908, 11 mar. 2022. Editora: ABRASCO - Associação Brasileira de Saúde Coletiva. ISSN 1413-8123, 1678-4561. DOI: <https://doi.org/10.1590/1413-81232022273.00432021>. Acesso em: 16 jun. 2025.

MIRANDA, M. D. M. F. *et al.* Adherence to antiretroviral therapy by adults living with HIV/aids: a cross-sectional study. **Revista Brasileira de Enfermagem**, v. 75, n. 2, e20210019, 2022. ISSN 1984-0446, 0034-7167. DOI: <10.1590/0034-7167-2021-0019>. Acesso em: 7 jun. 2025.

KINGMA, D. P.; BA, J. **Adam: A Method for Stochastic Optimization**. [*S. l.*]: arXiv, 30 jan. 2017. DOI: <10.48550/arXiv.1412.6980>. Acesso em: 14 out. 2025.

SRIVASTAVA, N. *et al.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **Journal of Machine Learning Research**, v. 15, n. 56, p. 1929–1958, 2014. ISSN 1533-7928. Disponível em: <http://jmlr.org/papers/v15/srivastava14a.html>. Acesso em: 10 nov. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **No Brasil, raça-cor ainda é um determinante social que afeta negativamente a saúde das pessoas negras**. Departamento de HIV, Aids, Tuberculose, Hepatites Virais e Infecções Sexualmente Transmissíveis. 2024. Disponível em: <https://www.gov.br/aids/pt-br/assuntos/noticias/2024/novembro/no-brasil-raca-cor-ainda-e-um-determinante-social-que-afeta-negativamente-a-saude-das-pessoas-negras>. Acesso em: 13 jun. 2025.

BRASIL. MINISTÉRIO DA SAÚDE. **Manual de rotinas para assistência de adolescentes vivendo com HIV/Aids**. Edição: Maria Letícia Santos Cruz. Brasília, DF: Ministério da Saúde, 2006. 176 p. (Série Manuais, 69). Accepted: 2016-12-15T02:15:28Z Publisher: Ministério da Saúde. ISBN 85-334-1290-8. Disponível em: <https://lume.ufrgs.br/handle/10183/150087>. Acesso em: 7 jun. 2025.

BUCKLEY, J. P. *et al.* Evolving methods for inference in the presence of healthy worker survivor bias. **Epidemiology**, Cambridge, Mass., v. 26, n. 2, p. 204–212, mar. 2015. ISSN 1531-5487. DOI: <10.1097/EDE.0000000000000217>.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. **Bayesian Survival Analysis**. New York, NY: Springer, 2001. (Springer Series in Statistics). ISBN 978-1-4419-2933-4 978-1-4757-3447-8. DOI: <10.1007/978-1-4757-3447-8>. Acesso em: 19 jun. 2025.

HOFFMAN, M. D.; GELMAN, A. **The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo**. [*S. l.*]: arXiv, 18 nov. 2011. DOI: <10.48550/arXiv.1111.4246>. Acesso em: 7 dez. 2025.

CARPENTER, B. *et al.* Stan: A Probabilistic Programming Language. **Journal of Statistical Software**, v. 76, p. 1–32, 11 jan. 2017. ISSN 1548-7660. DOI: <10.18637/jss.v076.i01>. Acesso em: 7 dez. 2025.

SALVATIER, J.; WIECKI, T. V.; FONNESBECK, C. Probabilistic programming in Python using PyMC3. **PeerJ Computer Science**, PeerJ Inc., v. 2, e55, 6 abr. 2016. ISSN 2376-5992. DOI: <10.7717/peerj-cs.55>. Acesso em: 7 dez. 2025.

KUCUKELBIR, A. *et al.* Automatic differentiation variational inference. **J. Mach. Learn. Res.**, v. 18, n. 1, p. 430–474, 2017. ISSN 1532-4435.

SPARAPANI, R. A. *et al.* Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). **Statistics in Medicine**, v. 35, n. 16, p. 2741–2753, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6893. ISSN 1097-0258. DOI: <10.1002/sim.6893>. Acesso em: 7 dez. 2025.

CHIPMAN, H. A.; GEORGE, E. I.; MCCULLOCH, R. E. BART: Bayesian additive regression trees. **The Annals of Applied Statistics**, v. 4, n. 1, 1 mar. 2010. ISSN 1932-6157. DOI: <10.1214/09-AOAS285>. Acesso em: 7 dez. 2025.

HILL, J.; LINERO, A.; MURRAY, J. Bayesian Additive Regression Trees: A Review and Look Forward. **Annual Review of Statistics and Its Application**, Annual Reviews, v. 7, p. 251–278, Volume 7, 2020 7 mar. 2020. ISSN 2326-8298, 2326-831X. DOI: <10.1146/annurev-statistics-031219-041110>. Acesso em: 7 dez. 2025.

TAN, Y. V.; ROY, J. Bayesian additive regression trees and the General BART model. **Statistics in Medicine**, v. 38, n. 25, p. 5048–5069, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8347. ISSN 1097-0258. DOI: <10.1002/sim.8347>. Acesso em: 7 dez. 2025.

HE, J.; YALOV, S.; HAHN, P. R. **XBART: Accelerated Bayesian Additive Regression Trees**. [*S. l.*]: arXiv, 14 mar. 2019. DOI: <10.48550/arXiv.1810.02215>. Acesso em: 7 dez. 2025.

HE, S.; SANG, H.; ZHOU, Q. **GS-BART: Bayesian Additive Regression Trees with Graph-split Decision Rules**. [*S. l.*]: arXiv, 8 set. 2025. DOI: <10.48550/arXiv.2509.07166>. Acesso em: 7 dez. 2025.

SPARAPANI, R. *et al.* Nonparametric competing risks analysis using Bayesian Additive Regression Trees. **Statistical Methods in Medical Research**, v. 29, n. 1, p. 57–77, jan. 2020. ISSN 1477-0334. DOI: <10.1177/0962280218822140>.

SPARAPANI, R. A. *et al.* Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. **Biostatistics (Oxford, England)**, v. 21, n. 1, p. 69–85, 1 jan. 2020. ISSN 1468-4357. DOI: <10.1093/biostatistics/kxy032>.

HENDERSON, N. C. *et al.* Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. **Biostatistics (Oxford, England)**, v. 21, n. 1, p. 50–68, 1 jan. 2020. ISSN 1468-4357. DOI: <10.1093/biostatistics/kxy028>.

LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. [*S. l.*]: arXiv, 25 nov. 2017. DOI: <10.48550/arXiv.1705.07874>. Acesso em: 10 nov. 2025.

PEARL, J. **Causality**. 2. ed. Cambridge: Cambridge University Press, 2009. ISBN 978-0-521-89560-6. DOI: <10.1017/CBO9780511803161>. Acesso em: 10 nov. 2025.