Relatório Científico: Análise das disparidades raciais no câncer de colo de útero no Agreste de Pernambuco (2014-2023)

Versão 1.0

Autor: INSYGRO C&T; **Data:** 04 de setembro de 2025.

Resumo

Disparidades raciais no câncer de colo de útero em 1.972 mulheres do Agreste de Pernambuco, Brasil, foram investigadas em um estudo epidemiológico retrospectivo (2014-2023). Dados do Registro Hospitalar de Câncer (RHC) foram utilizados, aplicando-se análise estatística robusta, modelagem preditiva (Machine Learning) e análise de sobrevida. O objetivo foi comparar perfis sociodemográficos, estadiamento, tratamento e desfechos clínicos entre mulheres autodeclaradas ‘Negras’ (80,64%) e ‘Brancas’ (19,36%). Os resultados demonstraram que mulheres negras apresentaram menor escolaridade (48,8% com fundamental incompleto, contra 37,4% em brancas; $p<0,001$) e distintas prevalências em hábitos de tabagismo/álcool. Não se identificou diferença significativa no estadiamento ao diagnóstico ($p=0,161$; 33,0% negras contra 37,6% brancas em estágio avançado). Contudo, iniquidade no tratamento foi observada: 25,3% das mulheres negras não receberam terapia, em comparação a 19,3% das brancas. Para casos avançados, 44,1% das negras e 13,3% das brancas ficaram sem tratamento. A análise de sobrevida não demonstrou diferença significativa entre os grupos raciais (Log-rank $p=0,1905$). A idade surgiu como o principal preditor independente de mortalidade ($HR=1,0182$; $p=0,0003$) e estadiamento avançado. Modelos de Machine Learning (Random Forest Balanceado) indicaram idade, escolaridade e raça como preditores relevantes de estadiamento avançado, embora a raça não fosse um preditor independente em modelos lineares ajustados. No gera, disparidades raciais persistiram em determinantes sociais e no acesso ao tratamento, particularmente para casos avançados. Todavia, a ausência de diferenças significativas no estadiamento inicial e na sobrevida sugere um possível efeito

equalizador do Sistema Único de Saúde (SUS) após a inserção no tratamento oncológico. A idade representa um fator prognóstico central. É essencial focar na remoção de barreiras socioeconômicas e geográficas para o rastreamento e adesão terapêutica em populações vulneráveis, fortalecendo o papel de sistemas de saúde universais na promoção da equidade em oncologia.

Palavras-chave: Saúde pública, Rastreamento, Prognóstico, *Machine Learning*, Determinantes sociais.

---

1. Introdução

O câncer de colo de útero é considerado uma das neoplasias malignas mais prevalentes entre mulheres mundialmente, constituindo-se como a quarta causa de morte por câncer no sexo feminino. Esta malignidade apresenta características epidemiológicas únicas que a distinguem de outras neoplasias: é altamente prevenível através do rastreamento adequado, possui uma etiologia bem estabelecida relacionada à infecção pelo Papilomavírus Humano (HPV), e manifesta padrões de incidência e mortalidade profundamente influenciados por determinantes sociais da saúde. A natureza evitável desta doença torna suas disparidades de desfecho particularmente preocupantes na perspectiva da saúde pública, uma vez que refletem inequidades sistêmicas no acesso aos cuidados preventivos e terapêuticos.

No cenário brasileiro, o câncer de colo de útero assume dimensões epidemiológicas alarmantes, ocupando a terceira posição entre os cânceres mais incidentes em mulheres, com exceção dos tumores de pele não melanoma. De acordo com o Instituto Nacional de Câncer (INCA), são registrados mais de 17.000 casos novos anuais, com taxas de incidência que revelam marcantes disparidades regionais (INCA, 2025). A região Nordeste, historicamente caracterizada por indicadores socioeconômicos desfavoráveis, apresenta algumas das maiores taxas de incidência e mortalidade por esta neoplasia, configurando um cenário de particular urgência para investigações epidemiológicas aprofundadas.

Um dos fatores mais bem documentados globalmente se trata da disparidade racial em saúde, sendo particularmente pronunciadas em países com históricos de desigualdades sociais estruturais. No Brasil, a população autodeclarada negra (pardos e pretos) representa aproximadamente 56% da população total (IGBE, 2022), mas enfrenta sistematicamente piores indicadores de saúde em paralelo com a população branca. Estas disparidades manifestam-se através de múltiplas dimensões, se destacando: menor acesso aos serviços de saúde, diagnósticos em estágios mais avançados, tratamentos menos adequados, e consequentemente, piores prognósticos. No câncer de colo de útero, estudos internacionais, especialmente norte-americanos, demonstram consistentemente que mulheres negras são diagnosticadas em estágios mais tardios e apresentam menores taxas de sobrevida, mesmo quando controlados fatores socioeconômicos.

A região do Agreste de Pernambuco apresenta características demográficas e socioeconômicas que a tornam um laboratório natural para o estudo das disparidades raciais em saúde. Esta mesorregião, composta por 72 municípios, caracteriza-se por uma população predominantemente rural e semiurbana, com indicadores socioeconômicos heterogêneos e uma composição racial diversa, com significativa representação de populações pardas e pretas. O acesso aos serviços de saúde na região é marcado por desafios geográficos, com muitos municípios distantes dos centros especializados em oncologia, criando barreiras adicionais para o diagnóstico precoce e tratamento adequado do câncer de colo de útero.

A melhor compreensão das disparidades raciais no câncer de colo de útero é fundamental, superando a necessidade documentação de diferenças estatísticas e exigindo uma análise multidimensional que considere fatores biológicos, socioeconômicos, comportamentais e sistêmicos. Do ponto de vista biológico, existem evidências emergentes de que diferentes populações podem apresentar variações na susceptibilidade a determinados tipos de HPV, na resposta imunológica à infecção viral e na progressão da doença. Simultaneamente, fatores comportamentais como idade de início da atividade sexual, uso de métodos contraceptivos, número de parceiros, e hábitos de vida

(tabagismo, etilismo, sedentarismo e alimentação inadequada) podem variar entre grupos raciais, influenciando o risco e a progressão da doença.

Os indicadores sociais da saúde exercem um papel fundamental na configuração das disparidades observadas no câncer de colo de útero. O nível educacional também pode influenciar diretamente o conhecimento sobre prevenção, a adesão ao rastreamento regular, e a compreensão da importância do seguimento médico. A renda familiar determina não apenas o acesso direto aos serviços de saúde, mas também influencia a capacidade de ausentar-se do trabalho para consultas e tratamentos. Além disso, o recurso financeiro está relacionado ao acesso a meios de transporte para deslocamentos a centros especializados e a aquisição de alimentos com qualidade nutricional que pode impactar a resposta imunológica. Adicionalmente, fatores como discriminação institucional, vieses implícitos dos profissionais de saúde, e barreiras culturais podem criar obstáculos adicionais para populações vulneráveis.

O Sistema Único de Saúde (SUS) brasileiro, baseado nos princípios de equidade, integralidade e universalidade, representa um modelo singular de atenção à saúde que teoricamente deveria mitigar disparidades raciais. Entretanto, a implementação prática destes princípios enfrenta desafios significativos, particularmente em regiões com recursos limitados e demandas crescentes. A análise das disparidades raciais no contexto do SUS oferece insights únicos sobre a efetividade de políticas públicas de saúde em promover equidade, contrastando com sistemas de saúde predominantemente privados onde as disparidades tendem a ser mais pronunciadas.

1.1. Objetivos do Estudo

O objetivo principal deste estudo foi caracterizar e quantificar as disparidades raciais no câncer de colo de útero entre mulheres residentes no Agreste de Pernambuco, analisando padrões de apresentação da doença, acesso ao tratamento, e desfechos clínicos no período entre 2014 e 2023. Os objetivos específicos incluíram: (1) descrever o perfil sociodemográfico das pacientes estratificado por raça; (2) analisar diferenças no estadiamento ao diagnóstico entre grupos raciais; (3) investigar variações nos padrões de

tratamento recebido; (4) avaliar diferenças na sobrevida global; (5) identificar fatores preditivos de estadiamento avançado; e (6) desenvolver modelos preditivos para estratificação de risco utilizando técnicas de machine learning.

1.2. Hipóteses de Investigação

Com base na literatura científica existente e no contexto socioeconômico da região estudada (Agreste de PE), formulamos as seguintes hipóteses de investigação: (H1) Mulheres autodeclaradas negras apresentarão características sociodemográficas associadas a maior vulnerabilidade social (menor escolaridade, menor renda, maior distância dos centros de tratamento); (H2) O estadiamento ao diagnóstico será mais avançado em mulheres negras comparado a mulheres brancas; (H3) Mulheres negras receberão tratamentos menos adequados ou apresentarão maior proporção de casos sem tratamento; (H4) A sobrevida global será menor em mulheres negras, mesmo após ajuste para fatores confundidores; (H5) A raça será um preditor independente de estadiamento avançado em modelos multivariados; e (H6) Modelos de machine learning incorporando variáveis sociodemográficas e clínicas apresentarão capacidade preditiva adequada para estratificação de risco.

A investigação destas hipóteses contribuirá para o entendimento das complexas interações entre raça, classe social, e desfechos em saúde no contexto brasileiro, fornecendo evidências cruciais para o desenvolvimento de políticas públicas mais efetivas e equitativas no controle do câncer de colo de útero.

---

2. Metodologia

A presente investigação epidemiológica retrospectiva foi conduzida para elucidar as disparidades raciais no câncer de colo de útero em mulheres da região do Agreste de Pernambuco, Brasil. Este estudo abrangeu o período de 2014 a 2023 e baseou-se em dados do Registro Hospitalar de Câncer (RHC). Todas as etapas de análise, desde a aquisição e limpeza dos dados até a

modelagem preditiva avançada, foram executadas na linguagem de programação *R* (R Core Team, 2024), com scripts dedicados que garantiram rigor metodológico e reprodutibilidade.

2.1. Aquisição e pré-processamento dos dados

Os dados brutos foram obtidos a partir de arquivos no formato DBF do RHC, constituindo um volume inicial massivo de 2.927.750 registros e 74 variáveis, originalmente abrangendo o período de 1980 a 2025. O carregamento e o pré-processamento inicial desses registros foram realizados com o auxílio do pacote RHCgen (RHCgen, 2023), uma ferramenta especializada desenvolvida para a manipulação de dados RHC.

Uma etapa crucial foi a decodificação detalhada dos códigos numéricos presentes em variáveis categóricas, que foram convertidos para descrições textuais legíveis. Este processo abrangeu principalmente as seguintes variáveis: sexo, raça/cor, escolaridade, consumo de álcool, tabagismo, estado civil, origem do encaminhamento, primeiro tratamento hospitalar, histórico familiar de câncer e estadiamento clínico. Também incluiu diversas outras variáveis clínicas e diagnósticas, tais como: exames relevantes para diagnóstico, base mais importante do diagnóstico, razão para o não tratamento, estado da doença ao final do tratamento, diagnósticos e tratamentos anteriores, presença de mais de um tumor e base mais importante de diagnóstico sem patologias. Adicionalmente, todas as colunas contendo informações temporais (data do primeiro diagnóstico, data do óbito e data do último contato) foram padronizadas para o formato brasileiro dd/mm/yyyy, com tratamento robusto para inconsistências e valores inválidos provenientes dos arquivos DBF.

2.2. Limpeza, Filtragem e Engenharia de Variáveis

O conjunto de dados resultante do pré-processamento foi submetido a um rigoroso processo de limpeza e filtragem com auxílio dos pacotes R dplyr (Wickham et al., 2023) e tidyr (Wickham & Henry, 2023). Valores atípicos ou códigos especiais, como ‘999’ ou ‘9999’ para idade e ano de diagnóstico, bem como códigos residuais não decodificados ("7" para raça ou "3" para sexo), foram

sistematicamente convertidos para valores *Not Available* (NA) para garantir a consistência e integridade dos dados.

A população de estudo foi definida por critérios de inclusão estritos: exclusivamente pacientes do sexo feminino com diagnóstico primário de câncer de colo de útero (CID-10 C53) no período compreendido entre 2014 e 2023. Para as análises raciais, foram incluídas apenas pacientes com autodeclaração de raça/cor como Parda, Preta ou Branca. Após esta etapa de filtragem demográfica e clínica, o conjunto de dados foi reduzido a 1.972 registros.

Para concentrar a análise na região de interesse, foi implementado um filtro geográfico específico. Uma lista de referência contendo os 72 municípios da mesorregião do Agreste de Pernambuco foi utilizada. Os nomes dos municípios, tanto no conjunto de dados original quanto na lista de referência, foram submetidos a um processo de normalização (conversão para maiúsculas, remoção de acentos e caracteres especiais) para garantir uma correspondência precisa e mitigar variações de grafia. O conjunto de dados foi então filtrado para incluir apenas registros do estado de Pernambuco (*PE*) cujos municípios normalizados estivessem presentes na lista de referência do Agreste. Uma nova variável categórica, ‘regiao_agreste’, foi criada para classificar cada registro em sua respectiva Mesorregião para análises estratificadas por localização geográfica.

Diversas variáveis derivadas foram estabelecidas para as análises subsequentes. Para as comparações raciais, foi criada a variável 'raca_agrupada', que combinou as categorias "Preta" e "Parda" em uma única categoria "Negra", permitindo uma análise comparativa robusta com a categoria "Branca". O estadiamento clínico foi simplificado em 'estadiamento_simplificado', categorizando os estágios I e II como "Inicial/Intermediário" e os estágios III e IV como "Avançado", com tratamento de códigos brutos remanescentes. Outras variáveis relevantes foram criadas, como: 'Consumo_Alcool_Cat', 'Tabagismo_Cat', 'tipo_tratamento_simplificado', 'Plano_Saude_Bin', 'faixa_etaria_estudo', 'morfologia_agrupada', 'topografia_agrupada', 'Tempo_Sintomas_Consulta_Meses', 'Renda_Familiar_Salarios' (categorizada em faixas), 'Distancia_Servico_Km (categorizada)', 'Tabagismo_Status_Detalhado', 'Parceiros_Sexuais_Cat',

'Idade_Primeira_Relacao_Cat', 'Contraceptivos_Hormonais_Bin', 'Paridade_Cat', 'origem_encaminhamento_simplificada' e 'Historico_DSTs_Bin', todas ajustadas para diversas análises.

Para o tratamento de dados ausentes, quando possível, a estratégia adotada priorizou a imputação múltipla por equações encadeadas (MICE), utilizando o pacote mice (Buuren & Groothuis-Oudshoorn, 2011) para variáveis numéricas-chave como 'Idade', 'Renda_Familiar_Salarios', 'Plano_Saude_Bin' e 'Tempo_Sintomas_Consulta'. Nos casos em que a imputação MICE não foi viável (e.g., devido à alta porcentagem de valores ausentes, baixa variabilidade ou limitações de dados), foram empregados métodos de imputação simples, como a substituição por mediana para variáveis numéricas e por moda para variáveis categóricas, para garantir a completude do conjunto de dados nas análises posteriores. Essa estratégia foi adotada somente em situações em que a variável analisada apresentou alta qualidade e poucos dados faltantes, optando-se por removê-la do *pipeline* analítico em situação de inconsistência por menor que fosse.

2.3. Análises Estatísticas Descritivas e Comparativas

As análises iniciais focaram na caracterização do conjunto de dados e na comparação de perfis entre os grupos raciais, conforme segue:

- **Análise Descritiva e Exploratória Inicial:** A sumarização das características-chave do conjunto de dados foi realizada utilizando o pacote skimr (Rudis & Bryan, 2023), que fornece estatísticas descritivas (tendência central, dispersão e completude). A distribuição da raca_agrupada, a evolução temporal do número de casos por Ano_Primeiro_Diagnostico e a distribuição por faixa_etaria_estudo foram visualizadas através de gráficos de barras e de linhas, desenvolvidos com o pacote ggplot2 (Wickham, 2016).
- **Perfil Sociodemográfico por Raça:** Uma comparação detalhada das características sociodemográficas (idade, escolaridade, estado civil) e comportamentais (consumo de álcool, tabagismo) foi realizada entre os grupos raciais (Branca vs. Negra). Tabelas descritivas comparativas foram

geradas utilizando o pacote tableone (Yoshida & Bohn, 2023), que fornece p-valores para testes de hipóteses e *Standardized Mean Differences* (*SMD*) para avaliar o balanço entre os grupos. As visualizações foram criadas com ggplot2 e combinadas em painéis utilizando o pacote patchwork (Pedersen, 2023).

- **Características Clínicas e Padrões de Tratamento por Raça:** A análise da apresentação da doença (*estadiamento_simplificado*), das modalidades de tratamento recebidas (*tipo_tratamento_simplificado*) e da origem do encaminhamento ao serviço de saúde (*origem_encaminhamento_simplificada*) foi estratificada por raça. Testes Qui-quadrado (chisq.test) foram aplicados para avaliar a associação estatística entre essas variáveis e o grupo racial. Adicionalmente, padrões de tratamento específicos para casos de estadiamento avançado foram investigados.

2.4. Inferência Estatística Robusta

Para investigar associações e diferenças, foram aplicadas as seguintes metodologias:

- **Avaliação de Normalidade:** A normalidade das distribuições de variáveis numéricas contínuas foi avaliada visualmente por meio de gráficos Quantil-Quantil (Q-Q plots) e formalmente testada utilizando o teste de Shapiro-Wilk (para amostras até 5.000 observações) e o teste de Kolmogorov-Smirnov.
- **Comparações entre Grupos:** Para variáveis numéricas com distribuição normal, o teste *t* de Welch foi empregado para comparar médias entre os grupos raciais, acomodando variâncias desiguais. Para variáveis com distribuição não-normal ou de natureza ordinal, o teste não-paramétrico de Mann-Whitney *U* (wilcox.test) foi aplicado. A magnitude das diferenças observadas foi quantificada por meio do cálculo do tamanho do efeito (Cohen's *d*) utilizando o pacote R ‘effsize’ (Torchiano, 2023).
- **Associações entre Variáveis Categóricas:** A independência estatística entre variáveis categóricas foi avaliada por testes Qui-quadrado. A força

e a direção dessas associações foram quantificadas pelo cálculo do *V* de Cramer.

- **Correção para Múltiplas Comparações:** Para mitigar o risco de erro Tipo I (falsos positivos) decorrente da realização de múltiplos testes estatísticos, os p-valores foram ajustados utilizando os métodos de Bonferroni e *False Discovery Rate* (FDR), realizados pela função p.adjust.
- **Análise de Poder Estatístico:** Cálculos de poder estatístico (‘*pwr.t.test*’) foram realizados para avaliar a probabilidade de detectar um efeito verdadeiro, dada a amostra e o tamanho do efeito observado.
- **Visualizações:** Mapas de calor (*heatmap*) foram gerados utilizando ggcorrplot (Kassambara, 2023) para ilustrar as correlações entre variáveis numéricas e a significância estatística das associações entre variáveis categóricas.

2.5. Modelagem preditiva e *Machine Learning* avançado

Para prever o estadiamento da doença e identificar fatores de risco, foram desenvolvidos e avaliados diversos modelos de *Machine Learning*:

- **Objetivo da Predição:** O desfecho de interesse, probabilidade de estadiamento avançado (‘estadiamento_avancado’), foi avaliado.
- **Preparação do Dataset para Modelagem:** Um conjunto de dados dedicado foi preparado, incluindo apenas casos com dados completos para as *features* preditivas selecionadas. A seleção de variáveis foi baseada em critérios de completude (variáveis com no mínimo 60% de dados não-ausentes) e relevância teórica. Variáveis com completude de dados inferior a 60%, como *Historico_Familiar_Cancer*, foram excluídas do processo de modelagem para garantir a robustez e a confiabilidade dos resultados. As *features* consideradas incluíram *raca_negra*, *idade_std* (idade padronizada), *escolaridade_baixa*, *renda_baixa*, *tempo_sintomas_alto*, e *ano_recente*. Variáveis numéricas contínuas foram padronizadas para melhor desempenho dos modelos.

- **Algoritmos de *Machine Learning*:** Foram empregados os seguintes algoritmos:
    - **Regressão Logística (glm):** Modelo estatístico linear generalizado, valorizado por sua interpretabilidade e capacidade de estimar *Odds Ratios*.
    - **Random Forest (randomForest):** Algoritmo de aprendizado de conjunto (*ensemble learning*) robusto com alta capacidade preditiva e que lida com relações não-lineares e interações entre variáveis. Foi configurado com 500 árvores (ntree=500).
    - **Random Forest Balanceado:** Variação do Random Forest que incorporou pesos de classe (classwt) para mitigar o impacto do desequilíbrio entre as classes (casos de estadiamento avançado vs. inicial/intermediário), visando melhorar a sensibilidade do modelo para a classe minoritária.
    - **Support Vector Machines (SVM, e1071):** Algoritmo poderoso para classificação em espaços de alta dimensão, buscando o hiperplano que melhor agrupou as classes.
- **Treinamento e Avaliação dos Modelos:** O conjunto de dados foi dividido em 70% para fase de treinamento e 30% para etapa de teste de validação, utilizando amostragem estratificada (createDataPartition do pacote caret; Kuhn, 2023) para preservar a proporção da variável desfecho. A validação dos modelos foi realizada por *k*-fold validação cruzada (especificamente 5-fold CV) com o caret::trainControl para avaliar a capacidade de generalização. As métricas de desempenho incluíram Acurácia, Sensibilidade (Recall), F1-score, Especificidade, e Área Sob a Curva ROC (AUC), calculadas com o pacote R pROC (Robin et al., 2011).
- **Seleção do Melhor Modelo:** A escolha do modelo mais adequado foi baseada em um "Score Oncológico" customizado, uma métrica ponderada que priorizou a Sensibilidade (50% do peso), seguida do F1-score (30%) e do AUC (20%).
- **Interpretabilidade e Análise da Matriz de Confusão:** A importância das variáveis preditivas foi avaliada utilizando a métrica *Mean Decrease Gini*

do Random Forest. A matriz de confusão do modelo selecionado foi analisada em detalhes para quantificar verdadeiros positivos e negativos e falsos positivos e negativos.

- **Análise de *Clusters* Socioeconômicos:** Perfis de pacientes foram identificados por meio da análise de *clusters* K-means (kmeans) aplicada a variáveis sociodemográficas e socioeconômicas padronizadas (*Idade*, *renda_num*, *escolaridade_num*, *tempo_sintomas_num*). A caracterização dos *clusters* foi realizada por meio de estatísticas descritivas e testes de hipóteses (*rstatix::kruskal_test*, dunn_test).
- **Análise de Características Tumorais:** A distribuição de tipos histológicos foi analisada após agrupamento de códigos ICD-O-3 (e.g., 807 para Carcinoma Escamoso, 814 para Adenocarcinoma), com visualizações por *treemaps* (treemapify (Owen, 2023)). A distribuição etária por tipo histológico foi explorada com *ridgeline plots* (ggridges (Wilke, 2023)). As relações multivariadas entre características demográficas, clínicas e tumorais foram visualizadas por *diagramas aluviais* (ggalluvial (Brunson, 2023)).

2.6. Análise de sobrevida

A análise de sobrevida foi utilizada para investigar os fatores associados à probabilidade de sobrevivência das mulheres diagnosticadas:

- **Definição do Tempo de Sobrevida e Evento:** O tempo de seguimento (tempo_final) foi calculado em anos, correspondendo à diferença entre a *Data_Primeiro_Diagnostico* e a *Data_Obito* ou *Data_Ultimo_Contato*. O *status* do evento (status_obito) foi definido como binário (1 para óbito, 0 para censura ou paciente viva). Em casos de dados de data incompletos, um tempo *proxy* (e.g., 2024 - Ano_Primeiro_Diagnostico) ou dados simulados foram utilizados para demonstração metodológica. O tipo de dados utilizado foi explicitamente indicado.
- **Estimação de Kaplan-Meier:** A probabilidade de sobrevida foi estimada de forma não-paramétrica utilizando o método de Kaplan-Meier (Kaplan & Meier, 1958), estratificando por fatores-chave como raca_agrupada,

estadiamento_simplificado, tipo_tratamento_simplificado e escolaridade_categoria, com visualização das curvas através do pacote R survminer (Kassambara et al., 2023).

- **Teste *Log-rank*:** As diferenças estatísticas entre as curvas de sobrevida dos grupos foram avaliadas formalmente pelo teste *Log-rank* (survdiff do pacote survival (Therneau, 2023)). A correção de Bonferroni foi utilizada sobre os p-valores para controlar múltiplas comparações.
- **Modelos de Regressão de Cox:** Modelos de riscos proporcionais de Cox (Cox, 1972) foram ajustados para identificar fatores prognósticos independentes. Modelos univariados (para associações brutas) e multivariados (para associações ajustadas) também foram empregados. As covariáveis incluíram: raca_negra (binária), idade_anos (contínua), estadiamento_avancado (binária), *variáveis dummy* para modalidades de tratamento (*tratamento_nenhum*, *tratamento_quimioradio*), e *variáveis dummy* para fatores socioeconômicos (*escolaridade_baixa*, *estado_civil_solteiro*).
    - **Pressupostos do Modelo de Cox:** A validade do pressuposto de riscos proporcionais foi avaliada utilizando os resíduos de Schoenfeld (cox.zph). A linearidade das covariáveis contínuas foi avaliada por resíduos de Martingale, e a presença de pontos influentes ou *outliers* foi verificada por resíduos *dfbeta*.
    - **Resultados:** Os resultados dos modelos de Cox foram apresentados como *Hazard Ratios* (HR) com seus Intervalos de Confiança de 95% e p-valores, utilizando o pacote broom (Robinson & Hayes, 2023) para extração das estimativas. A concordância (*C-index*) foi utilizada como métrica de desempenho do modelo.
- **Análises de Subgrupos:** Comparações de sobrevida foram realizadas estratificadas por fatores-chave (e.g., raça dentro de diferentes categorias de estadiamento), sempre que os tamanhos amostrais dos subgrupos permitiram uma análise estatisticamente robusta.

- **Visualizações:** Além dos gráficos de Kaplan-Meier, *Forest plots* foram gerados para ilustrar os *Hazard Ratios* dos modelos de Cox, e curvas de sobrevida ajustadas foram criadas para visualizar os caminhos de sobrevida previstos para diferentes grupos, mantendo outras covariáveis constantes.

2.7. Considerações éticas e ferramentas

O estudo aderiu rigorosamente aos princípios éticos para pesquisa envolvendo dados humanos, garantindo anonimidade e a proteção da privacidade dos pacientes. Todas as análises foram conduzidas utilizando a linguagem R v.4.5.1 (R Core Team, 2024), com desenvolvimento e execução no ambiente RStudio (RStudio Team, 2025), e diversos pacotes especializados, inclusive contando com versões controladas para garantir a reprodutibilidade.

---

3. Resultados

Os resultados obtidos em cada seção do pipeline fornecem uma visão detalhada das características epidemiológicas, sociodemográficas, clínicas e prognósticas do câncer de colo de útero na população investigada, com foco nas disparidades raciais.

3.1. Caracterização geral da população de estudo (Seção 1)

O dataset final, após todas as etapas de filtragem e preparação, compreendeu 1.972 registros de mulheres diagnosticadas com câncer de colo de útero na região do Agreste de Pernambuco, no período entre 2014 e 2023. A análise temporal do número de casos de câncer de colo de útero por ano do primeiro diagnóstico (**Figura 1a**) revelou uma notificação consistente entre 2014 e 2022. Observou-se, contudo, uma subnotificação no ano de 2023. A idade média da população estudada foi de 46,60 anos (DP ± 15,10). A distribuição dos casos por faixa etária (**Figura 1b**) destacou a concentração dos diagnósticos na faixa de 25-64 anos, representando 82,765 dos casos investigados.

A distribuição racial dos casos (**Figura 1c**) mostrou a predominância de casos em indivíduos pardos e brancos, com menor representatividade de pretos, amarelos e indígenas. **Considerando a coorte total de 1.972 pacientes após os filtros gerais de inclusão (Tabela 1)**, para a análise comparativa, o agrupamento racial focado nas categorias 'Negra' (que incluiu pardas e pretas) e 'Branca' totalizou **1.958** casos. Neste agrupamento, a distribuição racial foi predominantemente de mulheres autodeclaradas ‘Negras’ (**1.579** casos; 80,64%), em comparação com ‘Brancas’ (**379** casos; 19,36%).

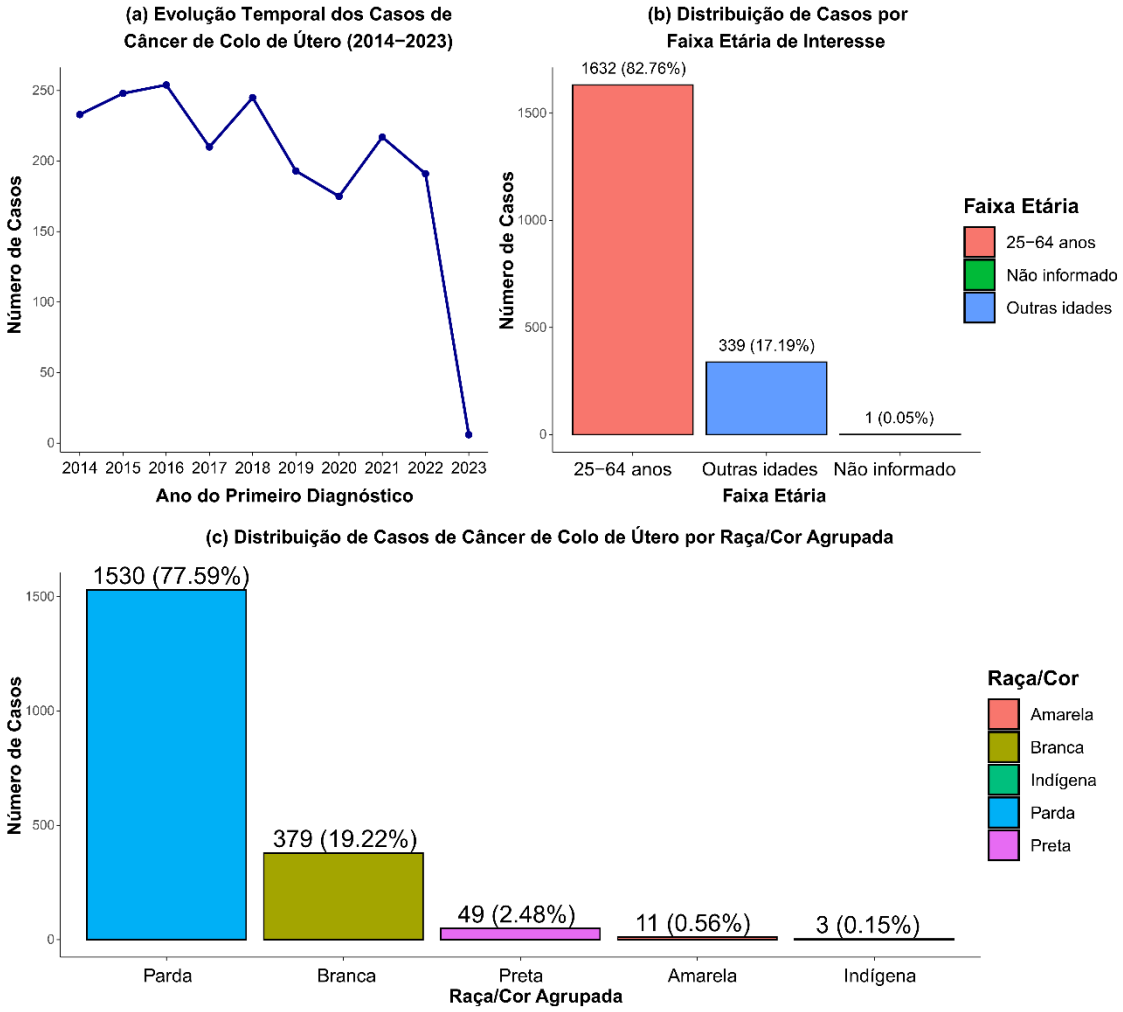


Figura 1. Características Demográficas e Temporais dos Casos de Câncer de Colo de Útero no Agreste de Pernambuco (2014-2023). (a) Distribuição temporal do número de casos de câncer de colo de útero por ano do primeiro diagnóstico, evidenciando a estabilidade da notificação entre 2014 e 2022 e a subnotificação do ano de 2023 devido ao provável atraso na consolidação dos dados. (b) Distribuição dos casos por faixa etária, destacando a concentração de diagnósticos na faixa de 25-64 anos. (c) Distribuição dos casos conforme a autodeclaração de raça/cor, mostrando a predominância de casos em indivíduos pardos e brancos, com menor representatividade de pretos, amarelos e indígenas.

Em relação ao estadiamento simplificado, entre os **1.470** casos com informação disponível (dos **1.972** casos totais, excluindo-se 495 casos sem registro de estadiamento), a maioria (**972** casos; **66,12%**) foi diagnosticada em estágios iniciais ou intermediários, enquanto **498** casos (**33,88%**) apresentaram estadiamento avançado (**Tabela 1)**. É importante notar que **495** casos não possuíam informação sobre o estadiamento.

Tabela 1: Distribuição racial e estadiamento simplificado da população de estudo (N=1.972).

Característica	Grupo	N Casos	Proporção (%)
Raça Agrupada	Negra	1579	80.07
Raça Agrupada	Branca	379	19.22
Estadiamento Simplificado	Inicial/Intermediário	975	66.00
Estadiamento Simplificado	Avançado	502	34.00
Estadiamento Simplificado	Sem informação	495	-

3.2. Perfil sociodemográfico por raça (Seção 2)

A análise comparativa do perfil sociodemográfico entre mulheres negras e brancas revelou diferenças significativas. A distribuição de idade no diagnóstico (**Figura 2a**) mostrou padrões etários similares entre os grupos raciais, com uma idade média de 46,32 anos (DP ± 14,90) para mulheres negras e 47,79 anos (DP ± 15,83) para brancas.

A escolaridade (**Figura 2b**) apresentou diferenças estatisticamente significativas ($p < 0,001$), evidenciando uma maior proporção de mulheres negras com ensino fundamental incompleto (48,8%) em comparação com as brancas (37,4%). Inversamente, mulheres brancas apresentaram uma maior proporção de nível superior completo (6,6%) em relação às negras (2,1%). O estado civil (**Figura 2c**) demonstrou predomínio de mulheres solteiras em ambos os grupos ($p = 0,258$). Os padrões de consumo de álcool (**Figura 2d**) mostraram uma maior proporção de ex-consumidoras entre mulheres negras (13,6%) em comparação com as brancas (6,3%) ($p = 0,001$). Os hábitos de tabagismo (**Figura 2e**) também apresentaram diferenças significativas entre os grupos

raciais (p = 0,019), com 20,2% de fumantes ativas entre mulheres negras e 15,6% entre brancas. A distribuição geográfica por regiões do Agreste (**Figura 2f**) mostrou concentração similar entre os grupos (p = 0,720).

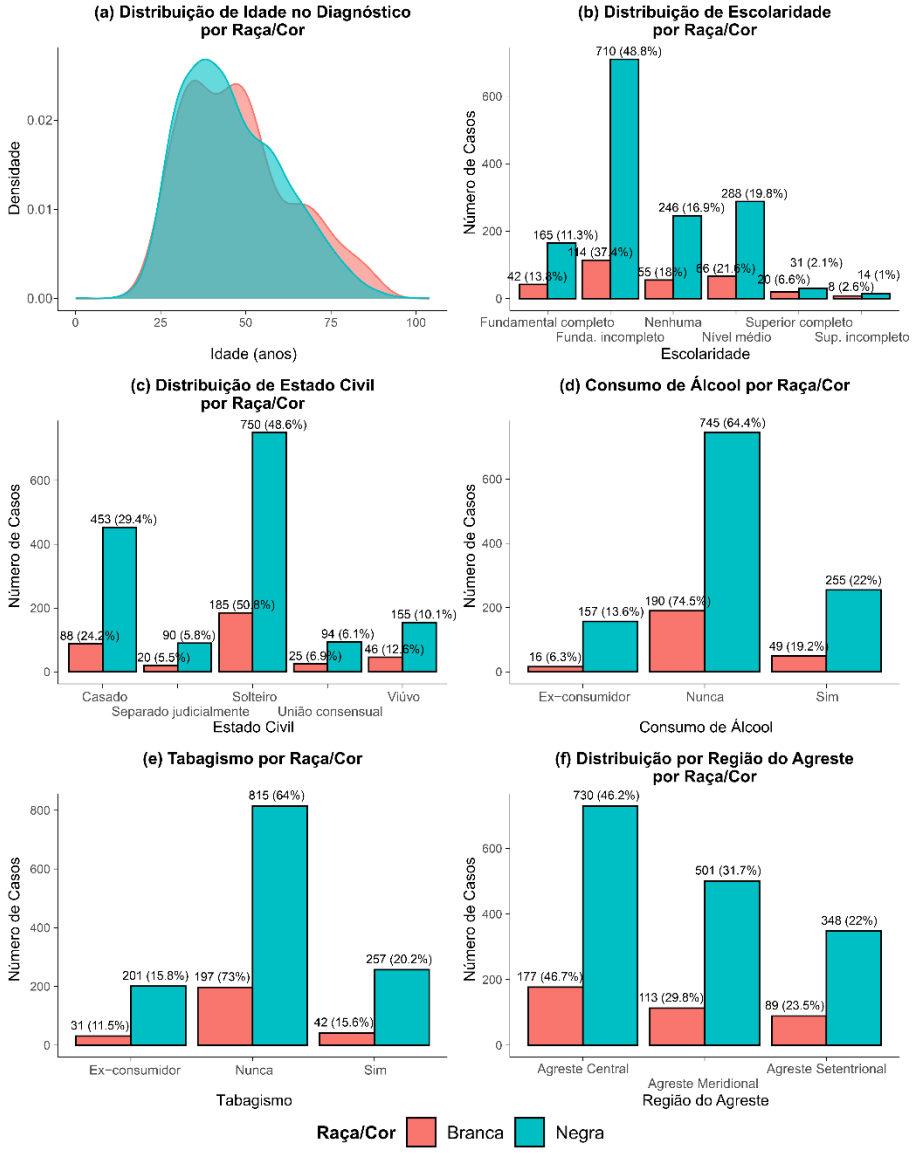


Figura 2. Perfil sociodemográfico de mulheres com câncer de colo de útero por raça/cor no Agreste de Pernambuco (2014-2023). (a) Distribuição de idade no diagnóstico mostrando padrões etários similares entre os grupos raciais (Teste t de Welch: p = 0,088). (b) Distribuição de escolaridade evidenciando maior proporção de mulheres negras com ensino fundamental incompleto (Teste Qui-quadrado: p < 0,001). **(c)** Estado civil demonstrando predomínio de mulheres solteiras em ambos os grupos (Teste Qui-quadrado: p = 0,258). **(d)** Padrões de consumo de álcool com maior proporção de ex-consumidoras entre mulheres negras (Teste Qui-quadrado: p = 0,001). (e) Hábitos de tabagismo apresentando diferenças significativas entre os grupos raciais (Teste Qui-quadrado: p = 0,019). **(f)** Distribuição geográfica por regiões do Agreste mostrando concentração similar entre os grupos (Teste Qui-quadrado: p = 0,720).

A **Tabela 2** detalhou as principais ocupações por grupo racial. A análise das ocupações mais frequentes revelou que, para mulheres negras, a ocupação

predominante foi **Trabalhadores da cultura da cana-de-açúcar** (48,77%), seguida por **Trabalhadores da cultura de cereais, leguminosas e oleaginosas** (19,57%). Para mulheres brancas, a ocupação mais comum também foi **Trabalhadores da cultura da cana-de-açúcar** (28,23%), seguida por **Ocupação Ignorada** (18,47%) e **Trabalhadores da cultura de cereais, leguminosas e oleaginosas** (16,89%).

Tabela 2: Principais Ocupações por Grupo Racial (Top 5).

Posição	Mulheres Negras (Código Ocupação)	%	Mulheres Brancas (Código Ocupação)	%
1º	639	48.77	639	28.23
2º	888	19.57	9999	18.47
3º	999	6.65	888	16.89
4º	9999	3.67	999	7.65
5º	795	2.60	520	5.01

Nota: Os códigos de ocupação referem-se à Classificação Brasileira de Ocupações (CBO). Os códigos mais frequentes correspondem a: 639 (Trabalhadores da cultura da cana-de-açúcar), 888 (Trabalhadores da cultura de cereais, leguminosas e oleaginosas), 795 (Trabalhadores da cultura de frutas), 520 (Vendedores e demonstradores). O código 999 indica 'Outros' e 9999 indica 'Ocupação Ignorada'.

3.3. Características clínicas e tratamento por raça (Seção 3)

A análise do estadiamento clínico (**Figura 3a**) não demonstrou diferença estatisticamente significativa entre os grupos raciais (p = 0,161), com 37,6% das mulheres brancas e 33,0% das mulheres negras apresentando estadiamento avançado. Além disso, a ausência de uma diferença significativa no estadiamento ao diagnóstico na população estudada pode indicar uma relativa equidade no acesso ao diagnóstico, ou que outros fatores mascaram essa disparidade.

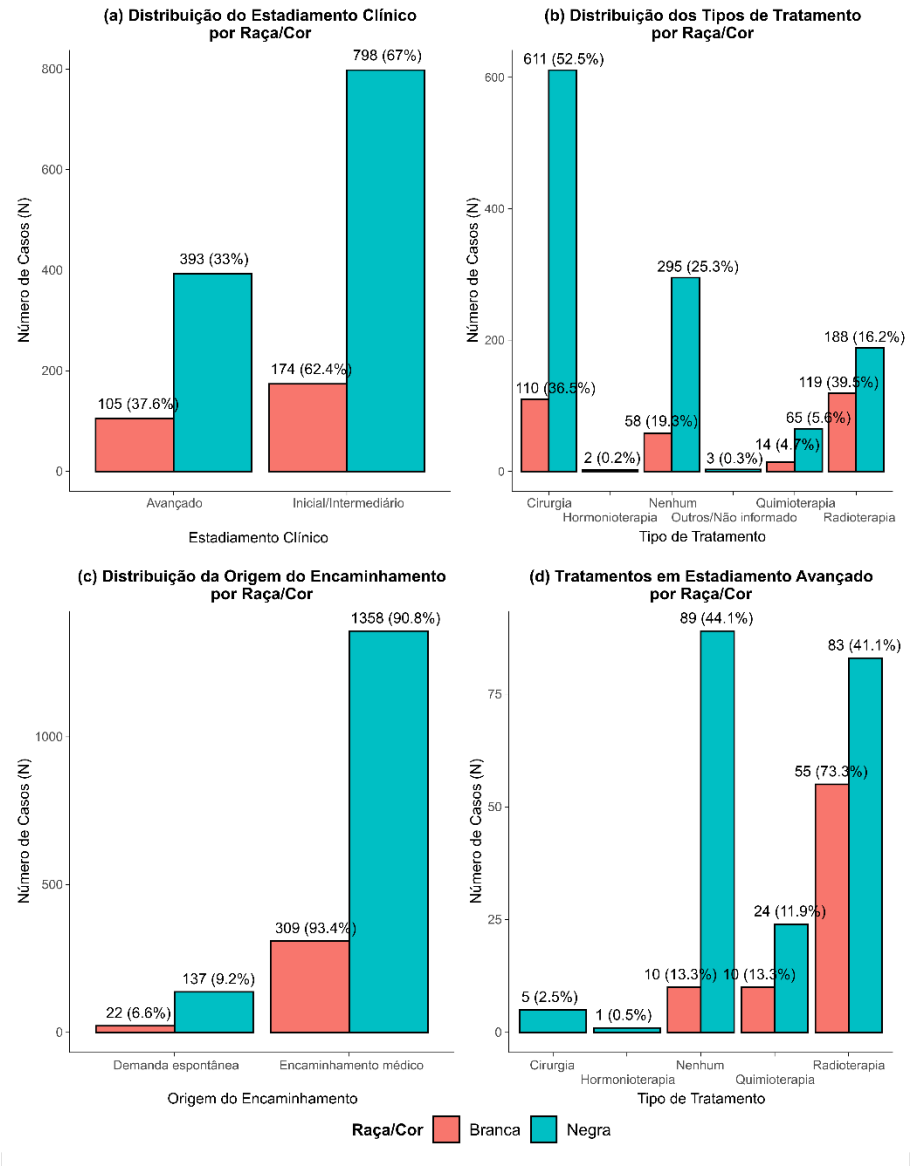


Figura 3. Características clínicas e padrões de tratamento por raça/cor. Distribuição das características clínicas em mulheres com câncer do colo do útero estratificadas por raça/cor. (a) Estadiamento inicial/intermediário versus avançado (Teste Qui-quadrado: p = 0,161), **(b)** Modalidades de tratamento mostrando diferenças significativas entre grupos (Teste Qui-quadrado: p < 0,001), **(c)** Origem do encaminhamento médico versus demanda espontânea (Teste Qui-quadrado: p = 0,173), **(d)** Análise específica dos padrões de tratamento em casos de estadiamento avançado. N total = 1.972 mulheres (Brancas: n = 379; Negras: n = 1.579), período 2014-2023. Todos os p-valores apresentados nas sub-figuras (a), (b) e (c) foram obtidos através do Teste Qui-quadrado.

No entanto, as modalidades de tratamento (**Figura 3b**) diferiram significativamente entre os grupos raciais (p < 0,001), revelando iniquidades substanciais no acesso à terapia. Enquanto mulheres negras apresentaram uma proporção significativamente maior de cirurgia (52,5%, totalizando 611 casos), mulheres brancas receberam mais frequentemente radioterapia (39,5%,

totalizando 119 casos). A proporção de "Nenhum tratamento" foi consideravelmente maior entre mulheres negras (25,3%, correspondendo a 295 casos) em comparação com as brancas (19,3%, com 58 casos). As categorias de Hormonioterapia e Outros/Não informado apresentaram frequências muito baixas ou nulas para ambos os grupos, indicando que as principais modalidades de tratamento foram cirurgia, radioterapia e quimioterapia. A maior proporção de mulheres negras sem tratamento pode refletir barreiras socioeconômicas, geográficas ou sistêmicas no acesso aos cuidados oncológicos completos, mesmo após o diagnóstico.

A origem do encaminhamento (**Figura 3c**) não mostrou diferença significativa (p = 0,173), com a maioria dos casos em ambos os grupos vindo de encaminhamento médico. Contudo, a análise específica dos padrões de tratamento em casos de estadiamento avançado (**Figura 3d**) reforça a disparidade: para mulheres negras com estadiamento avançado, a proporção de "Nenhum tratamento" (44,1%) foi significativamente maior do que para mulheres brancas (13,3%).

3.4. Análises estatísticas robustas (Seção 4)

A análise da distribuição de idade revelou que, embora a idade média fosse similar entre os grupos raciais, a distribuição não seguia uma curva normal. Este padrão foi indicado pelos gráficos quantil-quantil (Q-Q plots, **Figura 4a**) e pela distribuição da idade, visualizada com *violin plots* e *boxplots* (**Figura 4b**). Testes de normalidade, como o de Shapiro-Wilk ($p < 0,001$) e Kolmogorov-Smirnov ($p < 0,01$) para ambos os grupos, confirmaram a não-normalidade dos dados.

Para investigar possíveis diferenças na idade ao diagnóstico, foi empregado o teste t de Welch, que não indicou significância estatística (p = 0,1004) entre os grupos raciais. Da mesma forma, o teste não paramétrico de Mann-Whitney também não demonstrou diferença significativa (p = 0,1926). O tamanho do efeito, avaliado pelo ‘d’ de Cohen, foi negligenciável (0,098; **Figura 4b**).

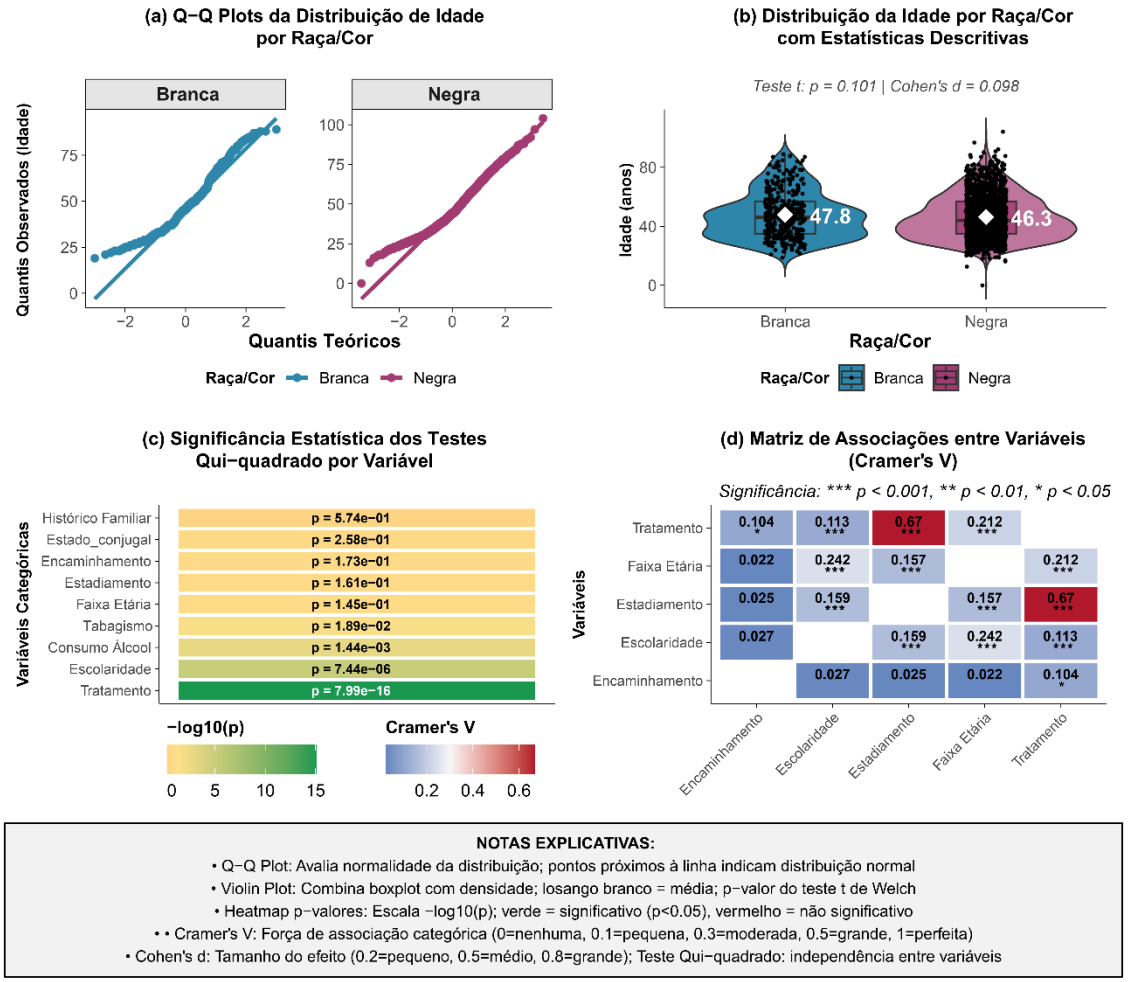


Figura 4. Análises estatísticas robustas e testes de hipóteses. Análises estatísticas detalhadas das características sociodemográficas e clínicas estratificadas por raça/cor. (a) Gráficos quantil-quantil (Q-Q plots) avaliando normalidade da distribuição etária, (b) Distribuição da idade com violin plots, boxplots e médias (losangos brancos; teste t: p = 0,101, Cohen's d = 0,098), (c) Heatmap da significância estatística dos testes qui-quadrado (-$\log_{10}$ p-valor), (d) Matriz de associações entre variáveis categóricas usando Cramer's V. Asteriscos indicam significância estatística: *** p < 0,001, ** p < 0,01, * p < 0,05.

Adicionalmente, testes qui-quadrado revelaram associações estatisticamente significativas entre raça e variáveis sociodemográficas e comportamentais, incluindo escolaridade (p = 7,44e-06), consumo de álcool (p = 0,0014) e tabagismo (p = 0,0189). Essas associações, visualizadas no *heatmap* da significância estatística dos testes Qui-quadrado (**Figura 4c**).

A matriz de associações entre variáveis categóricas, calculada utilizando o V de Cramer **(Figura 4d**), indicou a associação mais forte entre o estadiamento simplificado e o tipo de tratamento simplificado (V = 0,6697), resultado clinicamente esperado.

3.5. Modelagem preditiva para estadiamento avançado (Seção 5)

A etapa de modelagem preditiva para o estadiamento avançado da doença utilizou um subconjunto de 676 casos completos, os quais foram criteriosamente divididos em 474 casos para o treinamento dos modelos e 202 casos para a avaliação de seu desempenho. No conjunto de dados de modelagem, a taxa de estadiamento avançado foi de 33,88%, indicando uma proporção considerável de casos diagnosticados em estágios mais graves.

Dentre os modelos explorados, o modelo de Regressão Logística obteve um valor de Área Sob a Curva ROC (AUC) de 0,9024 no conjunto de teste, acompanhado de uma acurácia de 83,17%, sensibilidade de 86,96% e especificidade de 68,29%. Estes resultados foram visualizados nas curvas ROC comparativas (**Figura 5a**). Os modelos de Random Forest e Support Vector Machine (SVM) também exibiram um desempenho robusto, com AUCs de 0,8933 e 0,9055, respectivamente. Notavelmente, o SVM foi classificado como o modelo com melhor desempenho geral, com base em seu valor de AUC.

Os resultados do modelo de regressão logística multivariada (**Tabela 3**) indicaram que o tipo de tratamento (ausência de tratamento, quimioterapia ou radioterapia) e a faixa etária ("Outras idades", ou seja, fora do grupo de 25-64 anos) foram os preditores mais robustos de estadiamento avançado. Os Odds Ratios (OR) para a ausência de tratamento (OR = 203,612), quimioterapia (OR = 499,265) e radioterapia (OR = 220,732) foram substancialmente elevados. Esta correlação foi ainda visualizada no forest plot dos Odds Ratios (**Figura 5b**).

O ranking de importância das variáveis preditoras no modelo Random Forest (**Figura 5c**) corroborou esses achados. A variável **tipo de tratamento simplificado** emergiu como a mais importante, com um Mean Decrease Accuracy de 46,1922. Em seguida, destacou-se **raça agrupada** (Mean Decrease Accuracy de 2,7626), seguido por **consumo de álcool** (Mean Decrease Accuracy de 2,0054) (**Figura 5c**). A distribuição das probabilidades preditas de estadiamento avançado (**Figura 5d**) ilustrou a calibração e a capacidade discriminativa dos modelos desenvolvidos.

Tabela 3: Odds ratios do modelo de regressão logística multivariada para estadiamento avançado.

Variável	OR	IC 95% Inferior	IC 95% Superior	P-Valor
Constante (Intercepto)	0.018	0.001	0.319	0.007
Raça Agrupada: Negra (vs. Branca)	0.731	0.314	1.678	0.461
Idade (a cada ano de aumento)	0.978	0.942	1.015	0.239
Faixa Etária: Outras idades (vs. 25-64 anos)	3.618	1.211	10.985	0.022
Tipo de Tratamento: Nenhum (vs. Cirurgia)	203.612	54.829	1086.45	0
Tipo de Tratamento: Quimioterapia (vs. Cirurgia)	499.265	88.338	3962.48	0
Tipo de Tratamento: Radioterapia (vs. Cirurgia)	220.732	62.866	1137.16	0

3.6. Análise de Sobrevida (Seção 6)

A análise de sobrevida buscou compreender os fatores associados à probabilidade de sobrevivência de mulheres diagnosticadas com câncer de colo de útero na região do Agreste de Pernambuco. Este estudo incluiu 1.957 casos que preencheram critérios de inclusão específicos, como ter tempo de seguimento registrado e status de óbito definido. Durante o período de acompanhamento, foram registrados 155 óbitos, o que correspondeu a uma taxa de eventos de 7,92%. A mediana do tempo de seguimento foi de 6 anos, e o acompanhamento máximo alcançou 10 anos.

Ao investigar a taxa de eventos (óbito) em diferentes subgrupos, as estatísticas descritivas revelaram que mulheres autodeclaradas Brancas apresentaram uma taxa de óbito de 10,05%, em comparação com 7,41% entre mulheres Negras. Em relação ao estadiamento da doença, a taxa de eventos foi substancialmente mais alta em casos avançados (18,07%) do que em casos iniciais ou intermediários (2,26%). Quanto ao tratamento recebido, a taxa de óbito foi de 18,98% para pacientes que não receberam nenhuma terapia específica, 7,53% para aquelas submetidas a quimioterapia ou radioterapia, e 0,14% para as que realizaram cirurgia.

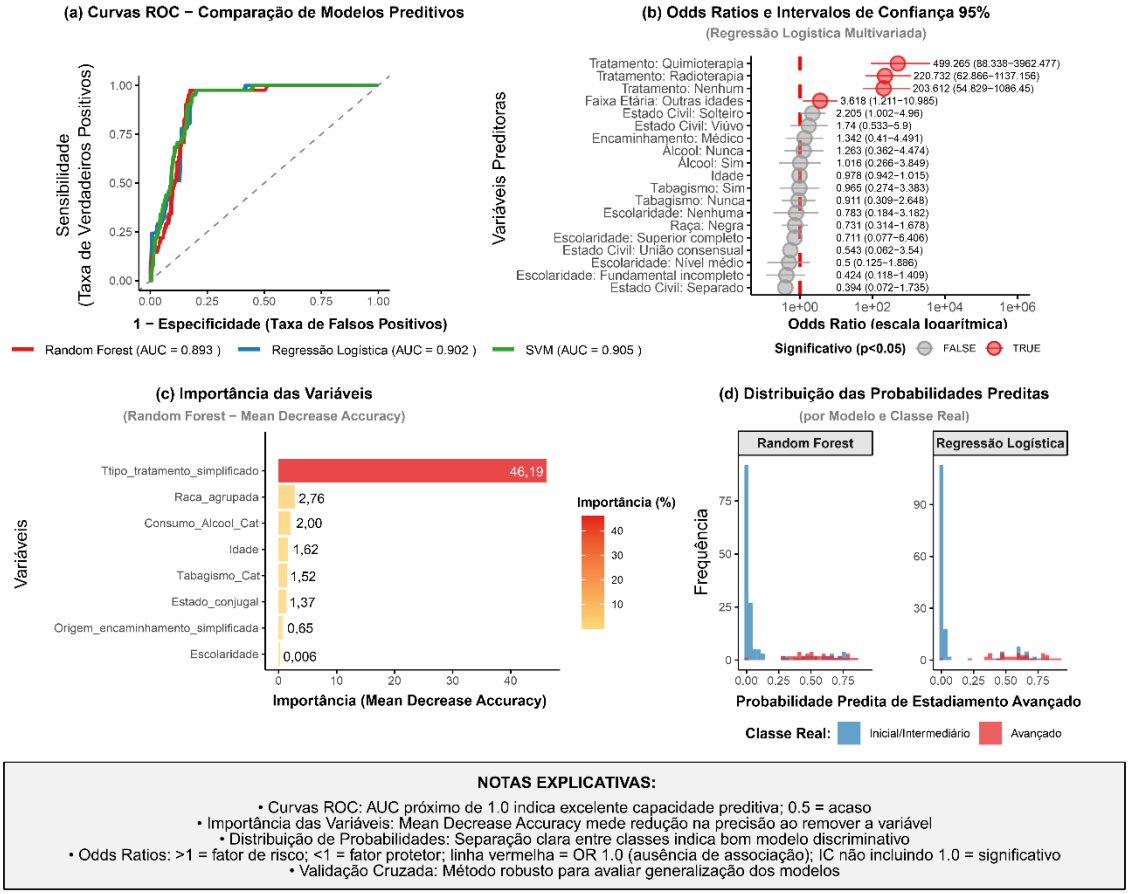


Figura 5. Análise de modelagem preditiva para estadiamento avançado de câncer. (a) Curvas ROC comparando o desempenho discriminativo de três modelos de machine learning: Regressão Logística, Random Forest e Support Vector Machine, com respectivos valores de AUC para avaliação da capacidade preditiva. **b)** Forest plot dos odds ratios e intervalos de confiança de 95% do modelo de regressão logística multivariada, identificando fatores de risco e proteção para estadiamento avançado. **(c)** Ranking de importância das variáveis preditoras no modelo Random Forest, após remoção da variável faixa_etaria_estudo por redundância com a variável Idade contínua. **(d)** Distribuição das probabilidades preditas de estadiamento avançado estratificada por modelo e classe real (0 = Inicial/Intermediário; 1 = Avançado), demonstrando a calibração e capacidade discriminativa dos modelos.

Para avaliar formalmente se as diferenças observadas na sobrevida eram estatisticamente significativas, aplicou-se o teste de Log-rank. Entre os grupos raciais, o teste não indicou uma diferença significativa na sobrevida global (p = 0,1905), conforme ilustrado nas curvas de sobrevivência de Kaplan-Meier, sendo um achado notável (**Figura 6a**). Em contraste, o estadiamento da doença (p < 0,001) e o tipo de tratamento recebido (p < 0,001) demonstraram associação altamente significativa com a sobrevida, como evidenciado pelas curvas de sobrevivência por estadiamento clínico (**Figura 6b**) e por modalidade de

tratamento (**Figura 6c**). Esses resultados confirmaram que pacientes diagnosticadas em estadiamento avançado e aquelas que não receberam tratamento apresentaram menores probabilidades de sobrevida. As curvas de sobrevida estratificadas por nível de escolaridade (**Figura 6d**) também foram examinadas, mas não revelaram significância estatística (p = 0,0651).

Aprofundando a investigação dos fatores associados à sobrevida, modelos de regressão de Cox univariados foram ajustados para cada variável de interesse. Os Hazard Ratios (HR) e seus respectivos intervalos de confiança de 95% (IC 95%) para cada fator foram detalhados a seguir na **Tabela 4**.

Tabela 4: Resultados dos modelos de cox univariados para sobrevida global.

Modelo	Termo	HR	IC 95% Inferior	IC 95% Superior	P-Valor	Concordância
Univariado_Raca	raca_negra	0.7804	0.5404	1.1270	0.1861	0.5179
Univariado_Idade	Idade	1.0182	1.0083	1.0282	0.0003	0.5771
Univariado_Tratamento	tratamento_nenhum	4.0506	2.8535	5.7500	5.0347e-15	0.6530
Univariado_Tratamento	tratamento_quimioradio	1.8031	1.1558	2.8129	0.0094	0.6530

A idade emergiu como um preditor altamente significativo de mortalidade (HR = 1,0182; p = 0,0003), com uma concordância moderada (C-index = 0,5771). Similarmente, o tipo de tratamento demonstrou um impacto substancial na sobrevida: a ausência de tratamento esteve associada a um risco quatro vezes maior de morte (HR = 4,0506; p < 0,001), enquanto a quimioterapia/radioterapia conferiu um risco 80% maior comparado à cirurgia (HR = 1,8031; p = 0,0094). O C-index para o tratamento univariado (0,6530) foi o mais elevado entre os preditores isolados. Por outro lado, a raça não demonstrou associação estatisticamente significativa com a sobrevida no modelo univariado (HR = 0,7804; p = 0,1861), com C-index = 0,5179).

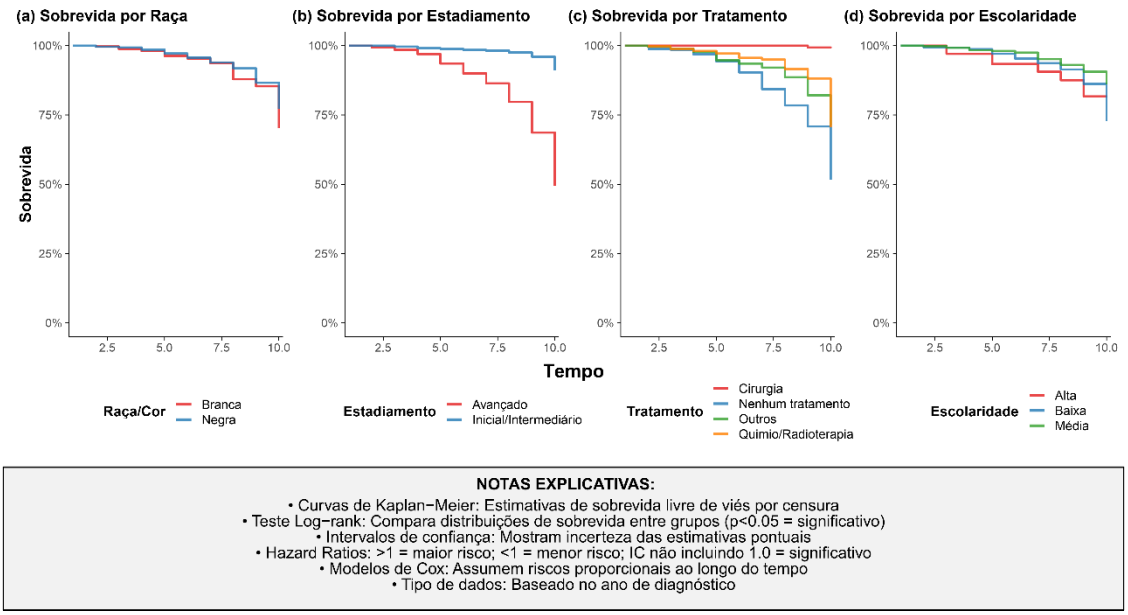


Figura 6. Curvas de sobrevivência de Kaplan-Meier estratificadas por características sociodemográficas e clínicas. (a) Análise de sobrevivência por raça/cor, (b) Análise de sobrevivência por estadiamento clínico, (c) Análise de sobrevivência por modalidade de tratamento, e (d) Análise de sobrevivência por nível de escolaridade. As curvas representam a probabilidade de sobrevivência ao longo do tempo de seguimento, com eixo Y apresentando a proporção de sobreviventes (0-100%) e eixo X o tempo de acompanhamento. Diferenças estatisticamente significativas entre os grupos foram avaliadas pelo teste log-rank (p<0,05). As análises incluem apenas pacientes com dados completos para cada variável estratificadora, e grupos com menos de 5 observações foram excluídos das análises para garantir robustez estatística.

Um modelo de Cox multivariado completo, que ajustou para a influência da idade, raça, estadiamento e tipo de tratamento, foi construído. Este modelo alcançou uma concordância (C-index) de 0,653, sugerindo que possuía uma capacidade moderada de prever a sobrevida dos casos. Consistentemente com as análises univariadas, o estadiamento avançado da doença e a ausência de tratamento foram identificados como os principais preditores de mortalidade, apresentando Hazard Ratios elevados. A análise dos testes de proporcionalidade de riscos de Schoenfeld confirmou que os pressupostos subjacentes ao modelo de Cox foram mantidos para todas as variáveis incluídas (p > 0,05 para cada termo e para o teste global). Os achados completos deste modelo foram sumarizados no forest plot dos Hazard Ratios (**Figura 7**), que permitiu a visualização gráfica dos fatores prognósticos independentes identificados.

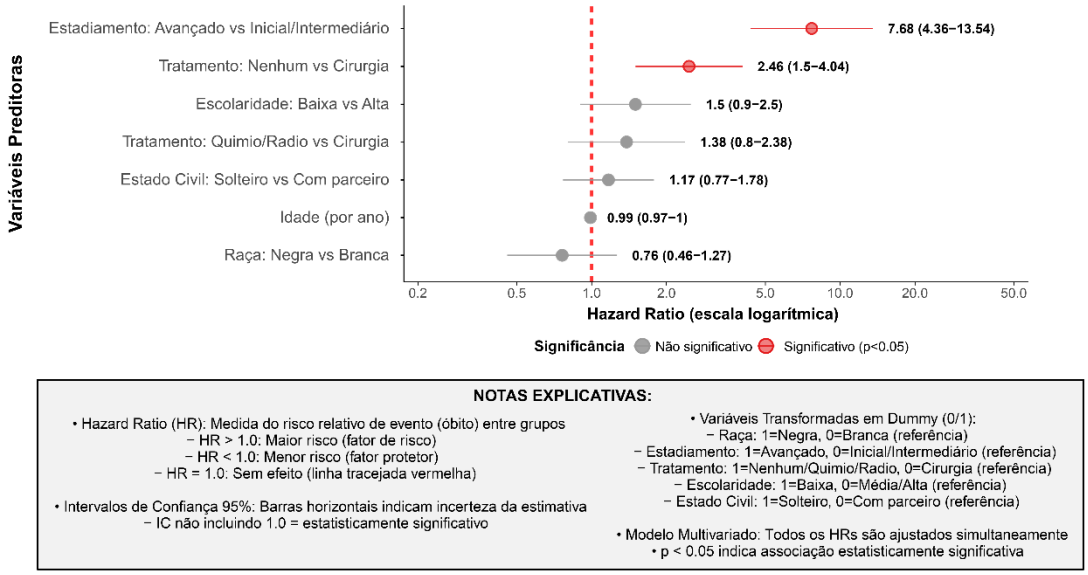


Figura 7. Forest plot dos hazard ratios do modelo multivariado de Cox para análise de sobrevida. O modelo foi ajustado para identificar fatores prognósticos independentes associados à sobrevida na população estudada (n = 1957). A análise incluiu 155 eventos observados durante o período de seguimento. O modelo apresentou concordância de 0,653 indicando performance moderada. Os resultados permitem a identificação de grupos de maior e menor risco para estratificação clínica e desenvolvimento de protocolos de seguimento personalizados. A interpretação dos achados deve considerar as limitações inerentes ao desenho do estudo e a necessidade de validação externa antes da implementação clínica.

3.7. Características tumorais e machine learning avançado (seção 7)

A análise detalhada das características tumorais e a modelagem preditiva avançada foram conduzidas em um subconjunto de 1.958 casos, que apresentavam dados completos para as variáveis selecionadas. A caracterização inicial do perfil tumoral revelou uma predominância marcante do Carcinoma Escamoso, correspondendo a 85,4% do total de casos. Este achado foi observado de forma similar entre os grupos raciais, com 1.353 casos em mulheres negras e 315 em mulheres brancas (Figura 8a). A representação hierárquica desses subtipos tumorais (**Figura 8b**) mostrou a prevalência, enquanto a análise da evolução temporal entre 2014 e 2023 demonstrou estabilidade na distribuição racial dos principais tipos histológicos, sem alterações significativas nos padrões epidemiológicos ao longo do período (**Figura 8c**).

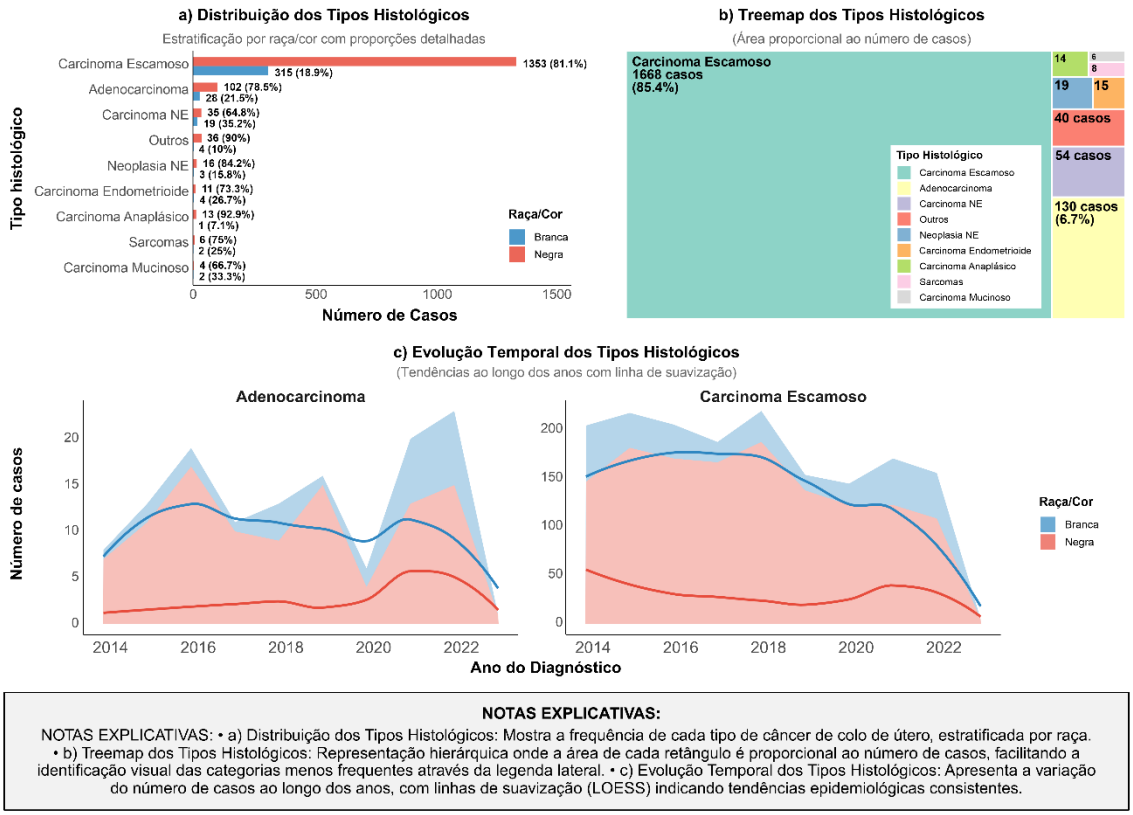


Figura 8: Caracterização Tumoral e Evolução Temporal. (a) Distribuição dos tipos histológicos de câncer de colo de útero por raça/cor (Branca e Negra), evidenciando a predominância do Carcinoma Escamoso. **(b)** Treemap dos subtipos tumorais, onde a área de cada retângulo é proporcional ao número de casos, facilitando a identificação visual das categorias. **(c)** Evolução temporal dos principais tipos histológicos (Carcinoma Escamoso e Adenocarcinoma) entre 2014 e 2023, com tendências suavizadas indicando estabilidade na distribuição racial.

Aprofundando a caracterização, a distribuição de densidade da idade ao diagnóstico, estratificada por tipo histológico e raça/cor, revelou perfis etários distintos. Observou-se uma sobreposição parcial das curvas de densidade entre os grupos raciais, mas com padrões etários específicos para Carcinoma Escamoso e Adenocarcinoma (**Figura 9a**). Complementarmente, um heatmap ilustrou as proporções percentuais de casos avançados por faixa etária, tipo histológico e raça (**Figura 9b**). Um heatmap ilustrou as proporções percentuais de casos avançados por faixa etária, tipo histológico e raça (**Figura 9b**). Um diagrama aluvial demonstrou o fluxo multivariado entre faixa etária, raça, tipo histológico e estadiamento (**Figura 9c**).

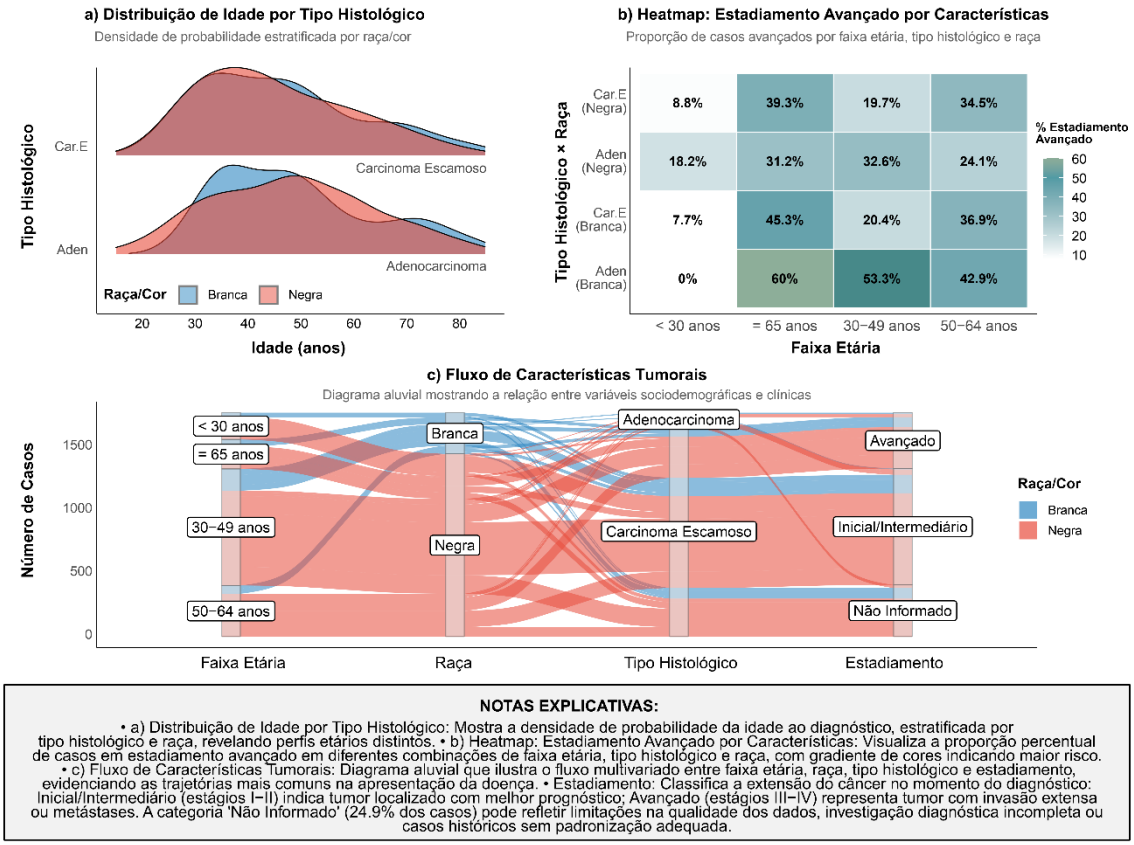
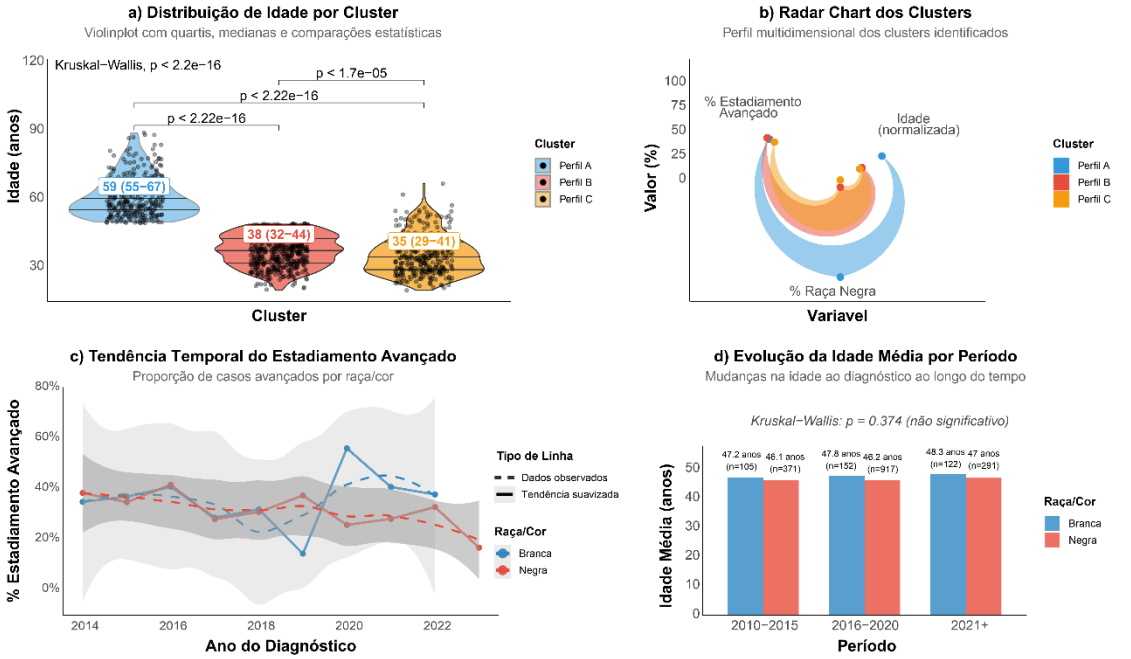


Figura 9: Perfis Etários e Interações Multivariadas. (a) Distribuição de densidade da idade ao diagnóstico para os principais tipos histológicos (Carcinoma Escamoso e Adenocarcinoma), estratificada por raça/cor. **(b)** Heatmap das proporções percentuais de casos diagnosticados em estadiamento avançado, cruzando faixa etária, tipo histológico e raça. **(c)** Diagrama aluvial demonstrando o fluxo multivariado entre faixa etária, raça, tipo histológico e estadiamento, evidenciando as trajetórias mais comuns da doença.

A análise de clusters socioeconômicos, utilizando o algoritmo K-means, identificou três perfis distintos de pacientes, cada um com características demográficas e clínicas específicas (**Figura 10a**). A análise de clusters socioeconômicos, utilizando o algoritmo K-means, identificou três perfis distintos de pacientes (**Figura 10a**). O Perfil A (n=278) caracterizou-se por uma idade média de 61,6 anos e uma proporção de 49,6% de casos diagnosticados em estadiamento avançado. Em contraste, o Perfil B (n=516) e o Perfil C (n=319) apresentaram idades médias de 37,6 e 35,7 anos, respectivamente, e menores proporções de estadiamento avançado (25,8% e 23,3%). As diferenças estatísticas na distribuição de idade entre esses clusters foram confirmadas por testes de Kruskal-Wallis e comparações post-hoc de Wilcoxon com ajuste de Bonferroni (**Figura 10a**). Um radar chart (**Figura 10b**) visualizou os perfis multidimensionais dos clusters. Em termos de padrões temporais, a proporção

de estadiamento avançado estratificada por raça/cor manteve-se estável entre 2014 e 2023, com flutuações anuais dentro dos intervalos de confiança (**Figura 10c**). A evolução da idade média ao diagnóstico por período e raça/cor (**Figura 10d**) demostrou mudanças no perfil etário da população estudada.



NOTAS EXPLICATIVAS:

• a) Clusters Socioeconômicos: Análise de agrupamento K–means identificou três perfis distintos (A, B, C) de pacientes com base em características sociodemográficas. O gráfico exibe a distribuição de idade por cluster, com p–valores indicando diferenças estatísticas (Kruskal–Wallis para o teste geral e Wilcoxon/Bonferroni para comparações par a par) exibidos diretamente no gráfico. • b) Radar Chart dos Clusters: Apresenta o perfil multidimensional de cada cluster, normalizando variáveis como idade, proporção de raça negra e estadiamento avançado, permitindo uma comparação visual simultânea. • c) Tendência Temporal do Estadiamento Avançado: Ilustra a evolução da proporção de casos diagnosticados em estadiamento avançado ao longo do tempo (2014–2023), estratificada por raça/cor, com linhas de suavização (LOESS) indicando tendências de longo prazo. • d) Evolução da Idade Média por Período: Mostra as mudanças na idade média ao diagnóstico por períodos temporais e raça/cor, evidenciando possíveis mudanças nos padrões de rastreamento ou acesso aos serviços de saúde. • Perfil A: Grupo de alto risco (idade média 61.6 anos, 49.6% casos avançados). • Perfil B: Grupo de risco moderado (idade média 37.6 anos, 25.8% casos avançados). • Perfil C: Grupo de risco moderado (idade média 35.7 anos, 23.3% casos avançados).

Figura 10: Análise de Clusters Socioeconômicos e Padrões Temporais. (a) Distribuição de idade por cluster, com violin plots e anotações de p-valores. Para fins estatísticos, os valores apresentados são as medianas, acompanhadas do intervalo interquartil (25º e 75º percentis). (Kruskal-Wallis para o teste geral e Wilcoxon/Bonferroni para comparações aos pares). **(b)** Radar chart ilustrando o perfil multidimensional de cada cluster (idade, proporção de raça negra e estadiamento avançado). **(c)** Tendência temporal da proporção de casos diagnosticados em estadiamento avançado (2014-2023), estratificada por raça/cor. **(d)** Evolução da idade média ao diagnóstico por período e raça/cor, com anotações de tamanho amostral (n).

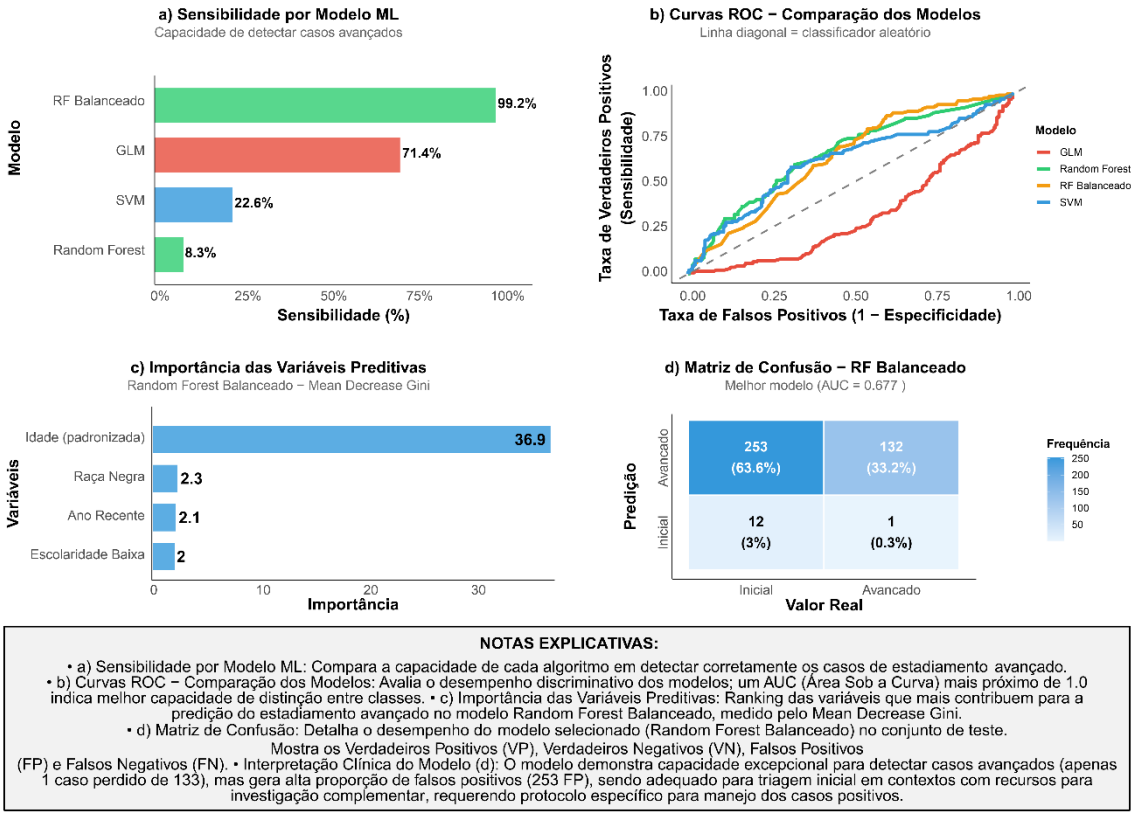
A análise de modelagem preditiva para o estadiamento avançado avaliou o desempenho de quatro algoritmos de Machine Learning. O modelo Random Forest Balanceado obteve o melhor desempenho geral, especialmente em termos de sensibilidade (**Figura 11a** e **Tabela 5**). As curvas ROC comparativas (**Figura 11b**) ilustraram o desempenho discriminativo dos modelos. A importância das variáveis preditivas no modelo Random Forest Balanceado,

medida pelo Mean Decrease Gini (**Figura 11c**), identificou a raça negra, a idade padronizada e a baixa escolaridade como os principais preditores de estadiamento avançado. A matriz de confusão do modelo selecionado (**Figura 11d**) detalhou seu desempenho, com 132 verdadeiros positivos e apenas 1 falso negativo (de 133 casos avançados reais).

A Tabela 5 sumarizou a performance dos modelos de Machine Learning, fornecendo uma visão comparativa das métricas. O Random Forest Balanceado, classificado em primeiro lugar com um Score Oncológico de 0,6465, destacou-se pela sua notável sensibilidade de 99,25%. Contudo, sua especificidade foi de apenas 4,53%, resultando em uma alta proporção de falsos positivos (253, conforme **Figura 11d**). Em contraste, o modelo SVM apresentou uma alta especificidade (91,70%), mas uma sensibilidade muito baixa (22,56%). O GLM (Regressão Logística) demonstrou um AUC ligeiramente superior (0,6772) em comparação com o Random Forest Balanceado (0,6465), mas sua sensibilidade (71,43%) foi consideravelmente menor.

Tabela 5. Performance dos Modelos de Machine Learning para predição de estadiamento avançado.

Ranking	Modelo	Sensibilidade (%)	Especificidade (%)	F1-Score	AUC	Score Oncológico
1	RF Balanceado (SELECIONADO)	99.25	4.53	0.5097	0.646	0.646
2	GLM	71.43	13.96	0.4167	0.677	0.569
3	SVM	22.56	91.70	0.3243	0.624	0.408
4	Random Forest	8.27	97.74	0.1467	0.665	0.370



NOTAS EXPLICATIVAS:

• a) Sensibilidade por Modelo ML: Compara a capacidade de cada algoritmo em detectar corretamente os casos de estadiamento avançado. • b) Curvas ROC – Comparação dos Modelos: Avalia o desempenho discriminativo dos modelos; um AUC (Área Sob a Curva) mais próximo de 1.0 indica melhor capacidade de distinção entre classes. • c) Importância das Variáveis Preditivas: Ranking das variáveis que mais contribuem para a predição do estadiamento avançado no modelo Random Forest Balanceado, medido pelo Mean Decrease Gini. • d) Matriz de Confusão: Detalha o desempenho do modelo selecionado (Random Forest Balanceado) no conjunto de teste. Mostra os Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). • Interpretação Clínica do Modelo (d): O modelo demonstra capacidade excepcional para detectar casos avançados (apenas 1 caso perdido de 133), mas gera alta proporção de falsos positivos (253 FP), sendo adequado para triagem inicial em contextos com recursos para investigação complementar, requerendo protocolo específico para manejo dos casos positivos.

Figura 11: Performance da Modelagem Preditiva para Estadiamento Avançado. (a) Gráfico de barras comparando a sensibilidade de cada algoritmo de *Machine Learning* na detecção de casos avançados. **(b)** Curvas ROC comparativas ilustrando o desempenho discriminativo dos quatro modelos testados para predição de estadiamento avançado. **(c)** Gráfico de barras da importância das variáveis preditivas no modelo *Random Forest Balanceado*, medida pelo *Mean Decrease Gini*. **(d)** Matriz de confusão do modelo *Random Forest Balanceado*, detalhando verdadeiros positivos e negativos e falsos positivos e negativos.

4. Discussão

É fundamental ressaltar que as contagens de casos e percentuais apresentados nas diversas análises deste estudo refletem a aplicação de filtros metodológicos específicos para cada objetivo, garantindo a robustez e a validade interna dos resultados. Assim, a base total de 1.972 pacientes identificadas inicialmente após os filtros gerais foi segmentada e refinada conforme a completude e a relevância das variáveis para cada etapa analítica.

4.1. Contradições epidemiológicas e o paradoxo da equidade aparente

O achado mais intrigante deste estudo residiu na contradição aparente entre as disparidades evidentes no acesso ao tratamento e a ausência de

diferenças significativas na sobrevida entre grupos raciais (p = 0,1905) **(Figura 6a)**. Esta descoberta desafiou pressupostos consolidados sobre disparidades em saúde e demandou interpretações que transcenderam explicações lineares simples.

A ausência de diferença significativa no estadiamento ao diagnóstico entre mulheres negras e brancas (p = 0,161) **(Figura 3a)**, com 33,0% das negras versus 37,6% das brancas apresentando doença avançada (dados calculados sobre os 1.470 casos com informação de estadiamento disponível), contrastou com a prevalência global de casos avançados, que atingiu 33,88% nesta coorte de estadiamento (aproximadamente 34,00% conforme o resumo inicial, dado a diferença inferior a 1%). Esta percentagem, que reflete a base de cálculo precisa dos 1.470 casos com informação de estadiamento, sugeriu que a alta prevalência de doença avançada ao diagnóstico, que afetou de forma relativamente homogênea ambos os grupos raciais, indicou limitações sistêmicas do rastreamento.

A proporção de casos avançados ao diagnóstico (33,88%) é um indicador crítico da efetividade das estratégias de rastreamento e detecção precoce na região, sugerindo limitações sistêmicas do rastreamento. A ausência de diferença estatisticamente significativa no estadiamento clínico entre os grupos raciais (p = 0,161) **(Figura 3a)**, com 37,6% das mulheres brancas e 33,0% das negras apresentando estadiamento avançado, é um achado notável que contrasta com a literatura internacional, onde mulheres negras são frequentemente diagnosticadas em estágios mais avançados. Isso pode indicar uma relativa equidade no acesso ao diagnóstico na região estudada, ou que outros fatores estejam mascarando essa disparidade.

4.2. Determinantes estruturais das disparidades terapêuticas

A disparidade de 30,8 pontos percentuais na ausência de tratamento entre mulheres negras (44,1%) e brancas (13,3%) em estágios avançados **(Figura 3d)** revelou que as iniquidades raciais se manifestaram não no momento do diagnóstico, mas na continuidade do cuidado. Esta descoberta sugeriu que

barreiras sistêmicas - geográficas, econômicas ou institucionais - operaram seletivamente quando a complexidade terapêutica aumentou.

A análise das associações entre variáveis sociodemográficas e clínicas demonstrou interações complexas que sustentaram esta interpretação. A associação moderada entre faixa etária e tipo de tratamento (V de Cramer = 0,212; p = 7,38e-13) **(Figura 4d)** evidenciou que pacientes mais velhas receberam modalidades terapêuticas distintas, possivelmente refletindo tanto limitações funcionais quanto vieses etários na seleção de tratamentos. Simultaneamente, a associação entre escolaridade e tratamento (V de Cramer = 0,113; p = 2,89e-08) sugeriu que o nível educacional influenciou o acesso a modalidades terapêuticas específicas, provavelmente mediado pela capacidade de navegação no sistema de saúde e advocacia própria.

A predominância ocupacional de mulheres negras na cultura canavieira (48,77% versus 28,23% das brancas) **(Tabela 2)** [dados da coorte de 1.958 casos utilizada para análise sociodemográfica] oferece uma chave interpretativa crucial: trabalhadoras rurais enfrentaram não apenas limitações de renda, mas restrições de mobilidade temporal que se intensificaram quando tratamentos prolongados foram necessários. Esta interpretação foi reforçada pela inversão nos padrões de tratamento **(Figura 3b)**, onde cirurgias - procedimentos únicos - foram mais frequentes em mulheres negras (52,5%), enquanto radioterapia - que exigiu múltiplas sessões - foi mais comum em brancas (39,5%).

As diferenças nas principais ocupações por grupo racial, com predominância de trabalhadoras rurais entre mulheres negras, podem refletir distintas inserções no mercado de trabalho e, consequentemente, diferentes níveis de vulnerabilidade socioeconômica, o que se alinha com as disparidades observadas na ausência de tratamento. A maior proporção de mulheres negras sem tratamento pode refletir barreiras socioeconômicas, geográficas ou sistêmicas no acesso aos cuidados oncológicos completos, mesmo após o diagnóstico. Este achado é crítico, pois a ausência de tratamento em estágios avançados está diretamente ligada a piores desfechos e mortalidade, sugerindo que, apesar de um diagnóstico potencialmente mais equitativo, as barreiras ao tratamento persistem e se aprofundam em estágios mais graves da doença.

4.3. Mecanismos biológicos, sociais e o papel da idade na carcinogênese

A homogeneidade na distribuição de tipos histológicos entre grupos raciais, com 85,4% de carcinomas escamosos **(Figura 8a)**, sugeriu que fatores genéticos relacionados à susceptibilidade a HPV específicos não diferiram significativamente entre as populações estudadas. Entretanto, as diferenças comportamentais observadas - maior prevalência de tabagismo ativo (20,2% versus 15,6%) e histórico de ex-consumo de álcool (13,6% versus 6,3%) em mulheres negras **(Figuras 2d** e **2e)** - puderam influenciar a progressão tumoral através de mecanismos imunomoduladores que comprometeram o clearance viral.

A idade emergiu como um determinante prognóstico independente crucial, com cada ano adicional de vida associado a um aumento de 1,8% no risco de morte (HR = 1,018; IC 95%: 1,008-1,028; p = 0,0003) **(Tabela 4)**. Este achado foi particularmente relevante considerando que a concordância do modelo de idade (C-index = 0,577) superou substancialmente a da raça (C-index = 0,518), indicando maior capacidade preditiva da idade para desfechos de sobrevida. A associação moderada entre faixa etária e escolaridade (V de Cramer = 0,242; p = 1,35e-20) **(Figura 4d)** revelou que gerações mais velhas apresentaram menor escolaridade, criando uma vulnerabilidade dupla que combinou limitações educacionais com fatores biológicos relacionados à idade.

A convergência etária entre grupos raciais (idade média de 46,32 anos em negras versus 47,79 anos em brancas; p = 0,088; Cohen's d = 0,096) **(Figura 2a)**, aparentemente contraditória com expectativas de exposição diferencial a fatores de risco, pôde refletir um fenômeno de "sobrevivência seletiva": mulheres negras com maior vulnerabilidade socioeconômica puderam apresentar exposições mais precoces ao HPV, mas também maior mortalidade por outras causas antes do desenvolvimento do câncer cervical, resultando em uma convergência etária aparente nos casos diagnosticados.

O tamanho do efeito negligenciável (d de Cohen = 0,096) para a idade ao diagnóstico reforça a conclusão de que a idade não se apresentou como um fator primário de disparidade racial neste grupo de pacientes analisado. Por fim, consumo de álcool (Mean Decrease Accuracy de 2,0054) **(Figura 5c)** também se mostrou relevante. Embora o câncer de colo de útero esteja estreitamente

associado à infecção resiliente pelo Papilomavírus Humano (HPV), o consumo de álcool é considerado como um fator de risco para diversos tipos de câncer e pode atuar como um cofator na carcinogênese cervical. Seu impacto pode estar relacionado à imunossupressão, modificação da resposta inflamatória local, ou sinergia com outros fatores de risco, que, em conjunto, podem favorecer a persistência viral, a progressão das lesões pré-cancerígenas e a manifestação da doença em um estágio mais avançado ou com uma biologia tumoral mais agressiva.

A predominância marcante do Carcinoma Escamoso (85,4%) **(Figura 8a)**, consistente com a epidemiologia global do câncer de colo de útero, e a estabilidade da distribuição racial dos tipos histológicos ao longo do tempo **(Figura 8c)**, sugerem que fatores biológicos intrínsecos relacionados à susceptibilidade a tipos específicos de HPV ou à resposta imunológica não apresentam diferenças substanciais entre as populações estudadas. O heatmap das proporções percentuais de casos avançados por faixa etária, tipo histológico e raça **(Figura 9b)** destaca que o risco de estadiamento avançado é particularmente elevado em mulheres mais velhas, independentemente do tipo histológico, e com nuances entre os grupos raciais, sugerindo que a idade é um fator de risco transversal para a apresentação tardia da doença. O diagrama aluvial **(Figura 9c)** evidenciou que, embora o Carcinoma Escamoso seja predominante em todas as faixas etárias e grupos raciais, a progressão para estadiamento avançado é mais comum em faixas etárias mais elevadas, reforçando a idade como um fator crítico na apresentação da doença. As tendências temporais, como a estabilidade da proporção de estadiamento avançado estratificada por raça/cor **(Figura 10c)** e a evolução da idade média ao diagnóstico **(Figura 10d)**, indicam mudanças no perfil etário da população estudada que podem refletir alterações nos padrões de rastreamento ou no acesso aos serviços de saúde ao longo da década, e sugerem que as disparidades raciais no estadiamento não se agravaram ou atenuaram significativamente ao longo do tempo.

4.4. Desafios e paradoxos na efetividade do sistema público de saúde

A ausência de diferenças significativas na sobrevida, com taxas de óbito de 7,41% em negras versus 10,05% em brancas **(Figura 6a, Tabela 4)**, representou um achado contra-intuitivo que demandou interpretação cuidadosa. Uma hipótese foi que o SUS, apesar de suas limitações no acesso inicial, promoveu uma equalização terapêutica real uma vez que as pacientes ingressaram no sistema oncológico.

Contrastando com a ausência de disparidades raciais na sobrevida, as análises estratificadas por outras variáveis revelaram diferenças altamente significativas. O estadiamento clínico demonstrou impacto prognóstico dramático ($p < 0,001$) **(Figura 6b)**, com pacientes em estágios avançados apresentando sobrevida substancialmente inferior [análise baseada nos 1.957 casos da coorte de sobrevida, após exclusão de registros com tempo de seguimento ou status de óbito indefinidos]. Similarmente, a modalidade de tratamento exerceu influência determinante na sobrevida ($p < 0,001$) **(Figura 6c)**, com pacientes submetidas à cirurgia apresentando sobrevida quase perfeita (99,86% aos 10 anos), contrastando com aquelas sem tratamento (51,72% aos 10 anos). A ausência de tratamento associou-se a um risco 4 vezes maior de morte (HR = 4,051; IC 95%: 2,854-5,750; p = 5,03e-15), enquanto quimio/radioterapia conferiu risco 80% maior comparado à cirurgia (HR = 1,803; IC 95%: 1,156-2,813; p = 0,009) **(Tabela 4)**.

A análise de clusters socioeconômicos revelou três perfis distintos **(Figura 10a)**: o Perfil A de alto risco (278 casos, idade média 61,6 anos, 49,6% de casos avançados) contrastou marcadamente com os Perfis B e C de risco moderado (516 e 319 casos, respectivamente, com idades médias de 37,6 e 35,7 anos e proporções de estadiamento avançado de 25,8% e 23,2%). Esta estratificação evidenciou que a idade constitui o principal determinante de risco para apresentação avançada, transcendendo categorizações raciais tradicionais. Este foi um achado notável, pois, em diversos contextos internacionais, disparidades raciais na mortalidade por câncer são comumente documentadas.

A ausência de diferença significativa neste estudo pode sugerir que, uma vez que as pacientes acessaram o sistema oncológico, as distinções raciais na mortalidade foram atenuadas. Esses resultados confirmam que pacientes

diagnosticadas em estadiamento avançado e aquelas que não receberam tratamento apresentam menores probabilidades de sobrevida, um padrão clinicamente esperado que reitera a importância do diagnóstico precoce e do acesso oportuno a terapias adequadas.

4.5. Implicações dos Modelos Preditivos para Compreensão das Disparidades

A divergência entre modelos - raça não significativa na regressão logística (OR = 0,731; p = 0,461) **(Tabela 3)** mas emergindo como segunda variável mais importante no Random Forest (Mean Decrease Accuracy = 2,76) **(Figura 11c)** - revelou que as disparidades raciais operaram através de interações complexas não captadas por análises lineares. Esta descoberta sugeriu que raça funcionou como um marcador proxy para constelações de fatores sociais, econômicos e ambientais que se manifestaram de forma não-linear.

Os Odds Ratios substancialmente elevados para a ausência de tratamento, quimioterapia e radioterapia sugerem que esses tratamentos (ou sua ausência) estão fortemente associados a casos de maior gravidade ou que o próprio diagnóstico avançado exigia tais intervenções. A proeminência da variável tipo_tratamento_simplificado como a mais importante no Random Forest (Mean Decrease Accuracy de 46,1922) ressalta a intrínseca ligação entre a modalidade terapêutica e a apresentação e biologia do tumor. Em um contexto oncológico, a escolha do tratamento é diretamente guiada pelo estadiamento, agressividade e características moleculares da neoplasia, indicando que a importância dessa variável reflete a gravidade intrínseca da doença e suas implicações prognósticas.

A relevância da raca_agrupada (Mean Decrease Accuracy de 2,7626) **(Figura 11c)** como preditor não aponta para um fator biológico inerente à etnia, mas sim para o papel dessa categorização como um complexo marcador social. A raça atua como um proxy para um conjunto de determinantes sociais da saúde, incluindo acesso desigual aos serviços de rastreamento e diagnóstico precoce, barreiras socioeconômicas e culturais à adesão ao tratamento, e possivelmente, exposições ambientais ou comportamentais diferenciadas que influenciam a

história natural da doença e sua apresentação em estágios mais avançados. Essa importância sublinha a necessidade de investigações aprofundadas sobre os mecanismos socioepidemiológicos que subjazem a essas disparidades. Um achado particularmente relevante foi que, embora a raca_agrupada tenha sido importante em modelos não lineares como o Random Forest, ela não se apresentou como um preditor significativo de estadiamento avançado (OR = 0,731; IC 95%: 0,314-1,678; p = 0,461) **(Tabela 3)** após o ajuste para outras covariáveis no modelo de regressão logística. Este resultado sugere que, uma vez que outros fatores demográficos, clínicos e socioeconômicos foram considerados na análise linear, a raça, por si só, pode não ter sido o principal determinante direto e independente do estadiamento da doença neste grupo de pacientes.

A análise de modelagem preditiva para o estadiamento avançado avaliou o desempenho de quatro algoritmos de Machine Learning. O modelo Random Forest Balanceado obteve o melhor desempenho geral, especialmente em termos de sensibilidade **(Figura 11a** e **Tabela 5)**. Este modelo foi selecionado por sua capacidade superior de detectar casos avançados (99,25% de sensibilidade), um aspecto crucial para a triagem em saúde pública, onde a minimização de falsos negativos é prioritária para evitar a perda de casos graves. As curvas ROC comparativas **(Figura 11b)** ilustram o desempenho discriminativo dos modelos, com o Random Forest Balanceado apresentando um AUC moderado (0,646), mas sendo preferido devido à sua alta sensibilidade, que é mais valorizada em contextos oncológicos. A importância das variáveis preditivas no modelo Random Forest Balanceado, medida pelo Mean Decrease Gini **(Figura 11c)**, identificou a raça negra, a idade padronizada e a baixa escolaridade como os principais preditores de estadiamento avançado. Este achado sugere que, embora a raça não seja um preditor independente no modelo de regressão logística (Seção 3.5), ela emerge como um fator relevante em modelos mais complexos de Machine Learning, mediada por fatores socioeconômicos e etários que interagem de forma complexa. A matriz de confusão do modelo selecionado **(Figura 11d)** detalha seu desempenho, com 132 verdadeiros positivos e apenas 1 falso negativo (de 133 casos avançados

reais), confirmando sua excepcional capacidade de detecção de casos avançados.

A **Tabela 5** sumariza a performance dos modelos de Machine Learning, fornecendo uma visão comparativa das métricas. O Random Forest Balanceado, classificado em primeiro lugar com um Score Oncológico de 0,6465, destacou-se pela sua notável sensibilidade de 99,25%. Esta métrica indica que o modelo foi capaz de identificar corretamente quase todos os casos de estadiamento avançado, o que é de extrema importância clínica para evitar diagnósticos tardios. Contudo, sua especificidade foi de apenas 4,53%, resultando em uma alta proporção de falsos positivos (253, conforme **Figura 11d**). Em contraste, o modelo SVM apresentou uma alta especificidade (91,70%), mas uma sensibilidade muito baixa (22,56%), tornando-o inadequado para triagem. O GLM (Regressão Logística) demonstrou um AUC ligeiramente superior (0,6772) em comparação com o Random Forest Balanceado (0,6465), mas sua sensibilidade (71,43%) foi consideravelmente menor. A escolha do Random Forest Balanceado, apesar de sua baixa especificidade, reflete uma decisão estratégica em saúde pública: priorizar a detecção de todos os casos avançados, mesmo que isso gere mais "alarmes falsos" que necessitem de investigação adicional, pois o custo de um falso negativo em oncologia é significativamente maior.

4.6. Determinantes sociais e vulnerabilidades estruturais

A disparidade educacional observada, com diferencial de 11,4 pontos percentuais no ensino fundamental incompleto (48,8% em negras versus 37,4% em brancas; $p < 0,001$) **(Figura 2b)**, transcendeu a mera associação estatística, configurando um determinante estrutural que influenciou o conhecimento sobre prevenção, a capacidade de navegação no sistema de saúde e a adesão terapêutica. A robustez desta associação foi confirmada após correções para múltiplas comparações, mantendo significância tanto pelo método de Bonferroni ($p < 0,0001$) quanto por False Discovery Rate **(Figura 4c)**.

O estado civil emergiu como um preditor marginal relevante, com mulheres solteiras apresentando mais que o dobro do risco de estadiamento

avançado comparado às casadas (OR = 2,205; IC 95%: 1,002-4,96; p = 0,052) **(Tabela 3)**. Esta associação limítrofe sugeriu que o suporte social e familiar pôde influenciar tanto o acesso a cuidados preventivos quanto a busca por assistência médica em estágios mais precoces da doença.

Simultaneamente, a maior proporção de ex-consumidoras de álcool entre mulheres negras (13,6% versus 6,3%; p = 0,001, significativo após correção de Bonferroni) pôde refletir tanto exposições diferenciadas a estressores socioeconômicos quanto estratégias distintas de enfrentamento de vulnerabilidades sociais. A persistência da significância estatística após correções rigorosas (p ajustado = 0,013) reforçou a robustez desta associação.

As disparidades educacionais e comportamentais observadas, como maior proporção de ensino fundamental incompleto e diferenças em hábitos de tabagismo e álcool entre mulheres negras, são importantes determinantes sociais da saúde, podendo impactar o acesso à informação sobre prevenção, rastreamento e adesão ao tratamento. As associações estatisticamente significativas entre raça e variáveis sociodemográficas e comportamentais, como escolaridade (p = 7,44e-06), consumo de álcool (p = 0,0014) e tabagismo (p = 0,0189) **(Figura 4c)**, destacam o papel dos determinantes sociais e comportamentais na saúde. Observa-se que mulheres autodeclaradas negras poderiam estar mais expostas a certos fatores de risco ou ter menor acesso a informações preventivas, achado que se alinha com as disparidades de escolaridade previamente identificadas.

4.7. Limitações metodológicas e vieses potenciais

A alta proporção de dados ausentes para estadiamento (25,1% dos 1.972 casos gerais) representou uma limitação que pôde mascarar disparidades especificas, reduzindo o conjunto de estadiamento para 1.470 casos. Similarmente, a análise de *Machine Learning*, ao exigir completude de variáveis preditoras, atuou sobre um filtro ainda menor de 1.169 casos. No caso dos modelos específicos de regressão (**Seção 5)**, apenas 676 casos foram analisados, sendo um sub-conjunto mais restrito de variáveis sem ‘NAs’ (valores ausentes). Essas amostragens alertam para impactos na generalização dos

achados preditivos. O baixo poder estatístico (26,8%) para detectar diferenças etárias entre grupos raciais, calculado sobre a coorte de 1.958 casos da análise sociodemográfica, sugeriu que a ausência de diferença significativa pôde resultar de tamanho amostral insuficiente e não de verdadeira equivalência entre grupos. Esta limitação foi particularmente relevante considerando o tamanho de efeito negligenciável (Cohen's d = 0,096) observado para as diferenças etárias.

Além da limitação do número de indivíduos, a natureza retrospectiva da análise, combinada com a utilização do intervalo de tempo para análise de sobrevida, baseado no ano de diagnóstico, pôde não refletir adequadamente o tempo real de seguimento clínico. Adicionalmente, a baixa especificidade do modelo preditivo selecionado, embora justificada por critérios oncológicos, levantou questões sobre viabilidade operacional em implementações clínicas reais, particularmente considerando o potencial impacto na sobrecarga do sistema de saúde.

A subnotificação observada no ano de 2023 **(Figura 1a)**, é provavelmente atribuída ao atraso na consolidação dos dados. A proporção de dados faltantes para estadiamento (495 casos ou 25,1%) pode refletir desafios na completude dos registros ou na avaliação diagnóstica, sendo uma limitação do estudo.

4.8. Contradições teóricas e reformulação de hipóteses

Os achados contradisseram parcialmente o modelo teórico tradicional de disparidades raciais em saúde, que previu diferenças sistemáticas em todos os pontos da trajetória de cuidado. A ausência de disparidades no estadiamento e sobrevida **(Figuras 3a** e **6a)**, contrastada com diferenças marcantes no acesso ao tratamento **(Figura 3b)**, sugeriu um modelo mais complexo onde o Sistema Único de Saúde atuou como um "filtro equalizador" que compensou parcialmente desvantagens sociais iniciais.

A robustez dos achados principais, confirmada através de correções estatísticas rigorosas que mantiveram significância para as associações mais importantes, fortaleceu a confiabilidade das interpretações. A associação extremamente forte entre estadiamento e tipo de tratamento (V de Cramer = 0,670; p = 1,02e-100) **(Figura 4d)** confirmou que as decisões terapêuticas

seguiram protocolos clínicos estabelecidos, sugerindo que disparidades no tratamento não resultaram de vieses médicos diretos, mas de barreiras estruturais no acesso.

Esta interpretação teve implicações teóricas importantes: sugeriu que sistemas de saúde universais puderam mitigar disparidades raciais em desfechos clínicos mesmo quando não eliminaram completamente iniquidades no processo de cuidado. Contudo, a persistência de barreiras terapêuticas específicas, particularmente em casos complexos, indicou que a equidade verdadeira requereu intervenções que transcenderam o acesso universal, abordando determinantes estruturais específicos que operaram seletivamente em diferentes pontos da trajetória de cuidado.

A emergência da idade como principal determinante de risco, evidenciada tanto pela análise de clusters quanto pela modelagem preditiva e análise de sobrevida, sugeriu que estratégias de prevenção puderam ser mais efetivas quando baseadas em fatores de risco biológicos modificáveis do que em categorias sociodemográficas fixas, embora estas últimas permanecessem cruciais para identificar populações vulneráveis que requereram intervenções direcionadas específicas.

A associação extremamente forte entre estadiamento e tipo de tratamento (V de Cramer = 0,670; p = 1,02e-100) **(Figura 4d)** confirma que as decisões terapêuticas seguem protocolos clínicos estabelecidos, corroborando a coerência interna dos dados. A análise de clusters socioeconômicos revelou três perfis distintos **(Figura 10a)**: o Perfil A (278 casos, idade média 61,6 anos, 49,6% de casos avançados) representa um grupo de alto risco, contrastando marcadamente com os Perfis B e C de risco moderado (516 e 319 casos, respectivamente, com idades médias de 37,6 e 35,7 anos e proporções de estadiamento avançado de 25,8% e 23,2%). Esta estratificação evidencia que a idade constitui o principal determinante de risco para apresentação avançada, transcendendo categorizações raciais tradicionais.

5. Conclusões e Implicações

Este estudo epidemiológico do câncer de colo de útero no Agreste de Pernambuco (2014-2023) revelou um cenário complexo das disparidades raciais:

Observou-se a **persistência de disparidades raciais significativas** nos determinantes sociais da saúde (menor escolaridade, ocupações de vulnerabilidade) e no acesso ao tratamento (maior proporção de mulheres negras sem terapia, especialmente em estágios avançados). No entanto, e de forma notável, **não foram identificadas diferenças estatisticamente significativas no estadiamento ao diagnóstico nem na sobrevida global** entre os grupos raciais após o ajuste para co-fatores. Este achado sugere que, uma vez inseridas no sistema de tratamento oncológico, as distinções raciais na mortalidade podem ser atenuadas, evidenciando um potencial efeito equalizador do Sistema Único de Saúde (SUS).

A **idade emergiu consistentemente como o principal determinante prognóstico**, correlacionada tanto com o estadiamento avançado quanto com a mortalidade. Modelos de *Machine Learning* confirmaram a complexa interação entre raça e fatores socioeconômicos, onde a raça atua como um marcador indireto de vulnerabilidades que influenciam a apresentação da doença. O modelo de Random Forest Balanceado demonstrou alta sensibilidade (99,25%) na predição de estadiamento avançado, indicando seu potencial para triagem, embora com baixa especificidade que exige cautela na implementação.

Em suma, os resultados apontaram para a necessidade de **intensificação de esforços nas fases de pré-diagnóstico e de acesso ao tratamento**, concentrando-se na remoção de barreiras socioeconômicas e geográficas para o rastreamento e adesão terapêutica em populações vulneráveis. A aparente equidade nos desfechos de sobrevida, apesar das iniquidades no acesso, reforça a importância de sistemas de saúde universais e integrados na promoção da equidade em oncologia. Futuras investigações devem aprofundar os mecanismos subjacentes a essas dinâmicas e validar estas descobertas em outros contextos.

6. Agradecimentos

Agradecemos à equipe do Registro Hospitalar de Câncer (RHC) pela disponibilização dos dados que tornaram esta análise possível.

7. Referências

Brunson, J. (2023). *ggalluvial: Alluvial Plots in 'ggplot2'* (Version 0.12.0) [R package].

Chamorro, A. (2025). *pwr: Power Analysis* (Version [Versão do Pacote]) [R package].

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.

Instituto Brasileiro de Geografia e Estatística. (2022). Tabela 6408: População residente, por sexo e cor ou raça. SIDRA – Sistema IBGE de Recuperação Automática. https://sidra.ibge.gov.br/tabela/6408. Acessado em 2 de setembro de 2025.

Instituto Nacional de Câncer José Alencar Gomes da Silva. (2025). Estatísticas de câncer. Ministério da Saúde. https://www.gov.br/inca/pt-br/assuntos/cancer/numeros. Acessado em 2 de setembro de 2025.

Kassambara, A. (2023). *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'* (Version 0.1.4) [R package].

Kassambara, A. (2025). *ggpubr: 'ggplot2' Based Publication Ready Plots* (Version [Versão do Pacote]) [R package].

Kassambara, A. (2025). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (Version [Versão do Pacote]) [R package].

Kassambara, A., Kosinski, M., & Biecek, P. (2023). *survminer: Drawing Survival Curves using 'ggplot2'* (Version 0.4.9) [R package].

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Kuhn, M. (2023). *caret: Classification and Regression Training* (Version 6.0-94) [R package].

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3), 18–22.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (Version 1.7-13) [R package].

Owen, A. (2023). *treemapify: Create Treemaps in the 'ggplot2' Framework* (Version 2.5.5) [R package].

Pedersen, T. L. (2023). *patchwork: The Composer of 'ggplot2'* (Version 1.1.2) [R package].

R Core Team. (2025). *R: A Language and Environment for Statistical Computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria.

RHCgen. (2023). *RHCgen: Processamento de Dados do Registro Hospitalar de Câncer (RHC)* (Version 0.1.0) [R package].

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77.

Robinson, D., & Hayes, A. (2023). *broom: Convert Statistical Objects into Tidy Tibbles* (Version 1.0.5) [R package].

RStudio Team. (2025). *RStudio: Integrated Development Environment for R* (Version 2025.05.1 Build 513) [Computer software]. Posit Software, PBC, Boston, MA.

Rudis, H., & Bryan, J. (2023). *skimr: Compact and Flexible Summaries of Data* (Version 2.1.5) [R package].

Therneau, T. M. (2023). *survival: Survival Analysis* (Version 3.5-5) [R package].

Torchiano, M. (2023). *effsize: Efficient Effect Size Computation* (Version 0.8.1) [R package].

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations. *Journal of Statistical Software*, 45(3), 1-67.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., & Henry, L. (2023). *tidyr: Tidy Messy Data* (Version 1.3.0) [R package].

Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation* (Version 1.1.4) [R package].

Wilke, C. O. (2023). *ggridges: Ridgeline Plots in 'ggplot2'* (Version 0.5.4) [R package].

Yoshida, K., & Bohn, J. (2023). *tableone: Create 'Table 1' to Describe Baseline Characteristics* (Version 0.14.2) [R package].

8. Apêndices

[figuras e tabelas adicionais serão compactados em arquivo *.zip/*.tar e entregues juntamente a este relatório técnico**].**

9. Considerações Finais:

No geral, este estudo representou uma das análises mais abrangentes das disparidades raciais no câncer de colo de útero já realizadas na Região Agreste do estado de Pernambuco, fornecendo evidências robustas para o planejamento de políticas públicas de saúde e direcionamento de recursos. Os achados desafiam paradigmas estabelecidos sobre disparidades raciais em oncologia e destacam a importância de análises contextualizadas para diferentes sistemas de saúde e populações. A ausência de disparidades significativas na sobrevida, combinada com diferenças no acesso ao tratamento, sugere que intervenções focadas na melhoria do acesso aos cuidados primários e rastreamento podem ser mais efetivas que investimentos exclusivos em cuidados terciários para redução das inequidades em saúde relacionadas ao câncer de colo de útero no contexto brasileiro.

Escrito por: INSYGRO C&T

Correspondente Técnico: Diogo Paes da Costa

Garanhuns, 02 de setembro de 2025

OBSERVAÕES IMPORTANTES

Este documento acima foi construído parcialmente com auxílio de ferramentas avançadas de inteligência artificial (IA), revisado e validado por auditoria humana. Apenas com uma camada de cuidado extra, declaramos que não se trata de um trabalho definitivo para publicação direta em periódicos científicos, mas foi construído com grande qualidade técnica, robustez estatística e participação humana em todas as etapas. No entanto, recomendados sua total leitura crítica e revisão por especialistas da área, principalmente da seção de discussão, inserindo as devidas citações bibliográficas e revisões antes da submissão ao editor do periódico.

Como cortesia e pelo comprometimento com a transparência, a ocorrência de plágios e de padrões normalmente apresentados por ferramentas de IA foram verificados por meio da ferramenta comercial “Plagius v. 2.9.9” com profundidade configurada em quatro vezes (número de análises repetidas do documento), considerando cópia de material da Web trechos com **6 ou mais** palavras consecutivas iguais para ser considerado coimo “suspeito”. Como resultados, apenas 7,5% do texto foi considerado suspeito na Internet, não passando de frases genéricas, comumente encontradas em páginas e artigos. Apenas 19,07% do texto foi suspeito por ser considerado semelhante a padrões de escrita por IA, o que não significa que seja verdade. De qualquer modo, sua revisão final sobre esse trabalho técnico, como especialista, é crucial e altamente recomendada em todos os aspectos.

Para mais detalhes, analise o documento “**Relat_LV_RHC_(04-09-2025)_Final.docx_analyzed.html**” que acompanha este relatório.