



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de Presidente Prudente

THAII CÉU SANTOS

**IDENTIFICAÇÃO DE FATORES ASSOCIADOS A NEFROTOXICIDADE
ASSOCIADA AO USO DE FUMARATO DE TENOFOVIR DISOPROXIL EM
PACIENTES COM HIV: UMA ABORDAGEM USANDO MODELOS DE
REGRESSÃO LOGÍSTICA**

PRESIDENTE PRUDENTE

2025

THAIL CÉU SANTOS

**IDENTIFICAÇÃO DE FATORES ASSOCIADOS A NEFROTOXICIDADE
ASSOCIADA AO USO DE FUMARATO DE TENOFOVIR DISOPROXIL EM
PACIENTES COM HIV: UMA ABORDAGEM USANDO MODELOS DE
REGRESSÃO LOGÍSTICA**

Relatório Semestral para a Bolsa COPE-
CONECTA apresentado ao Curso de
Graduação em Estatística da FCT/Unesp.
Orientador(a): Prof. Dr. Ricardo Puziol de
Oliveira

PRESIDENTE PRUDENTE

2025

Sumário

1 INTRODUÇÃO	1
2 METODOLOGIA.....	5
2.1 Modelo de Regressão Logística	6
2.2 Regularização L1: Lasso na Regressão Logística	8
3 RESULTADOS PARCIAIS	9
3.1 Análise Descritiva	9
3.2 Modelo de Regressão Logística	18
3.3 Modelo de Regressão Logística (Aplicando Lasso)	26
4 CONSIDERAÇÕES PARCIAIS	35
REFERÊNCIAS	36

1 INTRODUÇÃO

O vírus da imunodeficiência humana (HIV, do inglês *human immunodeficiency virus*), segundo o Ministério da Saúde (2008) é um lentivírus responsável pela síndrome da imunodeficiência adquirida (AIDS, do inglês *acquired immunodeficiency syndrome*), caracterizada por uma deterioração progressiva do sistema imunológico. O HIV infecta, principalmente, os linfócitos T CD4+, além de macrófagos e células dendríticas. A infecção leva à redução gradual da quantidade de linfócitos T CD4+ por diversos mecanismos, como a apoptose de células não infectadas próximas, a destruição direta de células infectadas e a ação de linfócitos T citotóxicos CD8+, que reconhecem e eliminam células infectadas. Quando o número de linfócitos T CD4+ atinge níveis críticos, o organismo perde sua capacidade de defesa mediada por células, tornando-se mais vulnerável a infecções oportunistas. A transmissão do HIV ocorre por meio de fluidos corporais como sangue, sêmen, secreções vaginais e leite materno. O vírus está presente nesses fluidos tanto em partículas livres quanto em células do sistema imunológico infectadas. As principais formas de contágio incluem relações sexuais sem proteção, compartilhamento de seringas contaminadas e a transmissão vertical — da mãe para o filho — durante a gestação ou o aleitamento.

Embora a origem exata dos vírus HIV-1 e HIV-2 ainda não seja completamente conhecida, evidências apontam que uma ampla família de retrovírus relacionados a eles circula entre primatas não humanos na África Subsaariana. Estudos científicos sugerem que o HIV-1 e o HIV-2 passaram a infectar seres humanos relativamente recentemente, possivelmente entre as décadas de 1940 e 1950. Diversos retrovírus encontrados em primatas africanos demonstram grande similaridade com esses dois tipos de HIV. Dentre eles, destaca-se o vírus da imunodeficiência símia (SIV), que infecta uma subespécie de chimpanzés africanos e possui aproximadamente 98% de similaridade com o HIV-1, indicando uma provável origem comum. Essas evidências sustentam a hipótese de que o HIV tenha surgido na África (BRASIL, 2008).

Em diversos países africanos, aproximadamente 83% das pessoas vivendo com HIV — o que corresponde a cerca de 26 milhões de indivíduos — têm acesso à terapia antirretroviral. No entanto, especialistas e profissionais da saúde alertam para a necessidade de um uso mais racional e consciente desses medicamentos. O HIV/AIDS permanece como um dos principais desafios de saúde pública na África e em escala global. Estima-se que quase 40 milhões de pessoas convivam com o vírus

no mundo, sendo aproximadamente 26 milhões apenas na África Subsaariana. Apesar da redução de 42% nas taxas de novas infecções entre 2010 e 2022 e do aumento significativo no acesso à terapia antirretroviral, a quantidade absoluta de pessoas vivendo com HIV continua a crescer ano após ano, evidenciando a necessidade de estratégias mais eficazes e sustentáveis de prevenção e cuidado. (GIRO, 2023)

O Ministério da Saúde, por meio do Departamento de HIV, Aids, Tuberculose, Hepatites Virais e Infecções Sexualmente Transmissíveis da Secretaria de Vigilância em Saúde (Dathi/SVSA), divulgou o Boletim Epidemiológico de HIV e Aids e, de acordo com o documento, houve um aumento de 4,5% nos casos de HIV em 2023 em comparação a 2022, o que reflete uma ampliação da capacidade de diagnóstico dos serviços de saúde no Brasil. Até setembro de 2023, 770 mil pessoas vivendo com HIV estavam em tratamento antirretroviral – 5% a mais que o registrado em todo o ano de 2022. Dessas, 49 mil iniciaram o tratamento em 2023. Atualmente, quase 200 mil pessoas sabem que têm o HIV no Brasil, mas não se tratam. Ainda segundo o boletim, a taxa de mortalidade por Aids em 2023 foi de 3,9 óbitos por 100 mil habitantes, a menor registrada desde 2013. Quanto ao perfil epidemiológico, 70,7% dos casos notificados ocorreram em pessoas do sexo masculino, 63,2% em indivíduos pretos e pardos, e 53,6% em homens que fazem sexo com homens (HSH). A razão entre os sexos foi de 2,7 casos em homens para cada caso em mulheres. A faixa etária mais acometida foi a de 20 a 29 anos, responsável por 37,1% das notificações, sendo que, entre os homens, essa mesma faixa etária concentrou 41% dos casos registrados (BRASIL, 2024).

O Fumarato de Tenofovir Disoproxil (TDF), um inibidor nucleotídico da transcriptase reversa (ITRN), é amplamente utilizado como componente de primeira linha na terapia antirretroviral (TARV) em países com recursos limitados, como a Índia e diversas nações da África Subsaariana. Desde sua aprovação pela Food and Drug Administration (FDA) dos Estados Unidos, em outubro de 2001, o TDF foi incorporado às diretrizes terapêuticas de vários países como medicamento de primeira escolha para o tratamento da infecção pelo HIV. (SANTOS et al., 2023)

Na Índia, o número de pacientes tratados com TDF tem aumentado progressivamente desde a implementação da política “Testar e Tratar”, adotada pelo Programa Nacional de Controle da Aids (NACP). Essa diretriz recomenda o início imediato do tratamento antirretroviral, independentemente da contagem de linfócitos

CD4 ou do estágio clínico da infecção pelo HIV, priorizando o regime baseado em TDF para pacientes virgens de tratamento, exceto nos casos de contraindicação. Administrado por via oral, o TDF apresenta boa eficácia e segurança relativa. No entanto, estudos têm demonstrado seu potencial nefrotóxico, principalmente devido à lesão das células dos túbulos proximais renais, o que pode culminar em lesão renal aguda (LRA), doença renal crônica (DRC) ou, em casos mais graves, na síndrome de Fanconi — parcial ou completa. A droga é amplamente distribuída no organismo, com maiores concentrações observadas no rim e no fígado. A toxicidade renal está associada ao acúmulo intracelular de TDF nos túbulos renais (SANTOS et al., 2023).

Comparado a outros ITRNs, como Abacavir (ABC), Lamivudina (3TC) e Zidovudina (AZT), o TDF tem sido relacionado a uma redução mais acentuada na taxa de filtração glomerular estimada (TFGe), além de disfunção tubular proximal. A deterioração da função renal pode se manifestar como LRA, DRC ou como uma queda da TFGe em relação aos valores basais, mesmo quando ainda se encontram dentro dos limites fisiológicos (SANTOS et al., 2023).

Diversos fatores estão associados à nefrotoxicidade induzida pelo TDF, incluindo: idade, sexo, estado nutricional (índice de massa corporal - IMC), tempo de uso da medicação, níveis basais de creatinina, uso concomitante de inibidores da transcriptase reversa não nucleosídeos (ITRNN), contagem basal de CD4, estágio clínico da infecção pelo HIV segundo a OMS, presença prévia de disfunção renal, histórico de uso de antirretrovirais nefrotóxicos, administração de outros medicamentos nefrotóxicos e a existência de comorbidades (SANTOS et al., 2023).

Os métodos de regressão tornaram-se uma ferramenta essencial em qualquer análise de dados voltada a descrever a relação entre uma variável resposta e uma ou mais variáveis explicativas. Frequentemente, a variável de desfecho é discreta, assumindo dois ou mais valores possíveis. Nesses casos, o modelo de regressão mais utilizado é o da regressão logística.

A regressão logística, tem como o mesmo objetivo de qualquer outro modelo de regressão: encontrar o modelo mais adequado, parcimonioso e clinicamente interpretável, que descreva a relação entre uma variável dependente (ou resposta) e um conjunto de variáveis independentes (ou explicativas), também chamadas de covariáveis. O exemplo mais comum de modelagem é a regressão linear, em que se assume que a variável de desfecho é contínua. A principal distinção entre a regressão logística e a linear está no tipo de variável resposta: enquanto a regressão linear lida

com desfechos contínuos, a regressão logística é empregada quando a variável resposta é binária ou dicotômica (HOSMER; LEMESHOW; STURDIVANT, 2013).

Essa diferença entre os dois modelos se reflete tanto na estrutura matemática do modelo quanto em seus pressupostos estatísticos. No entanto, uma vez compreendida essa distinção, os procedimentos analíticos da regressão logística seguem, em geral, os mesmos princípios adotados na análise por regressão linear (HOSMER; LEMESHOW; STURDIVANT, 2013). No entanto, em contextos com grande número de variáveis ou colinearidade entre os preditores, o modelo pode se tornar instável ou superajustado. Para mitigar esses problemas, métodos de regularização são aplicados, como o Lasso (Least Absolute Shrinkage and Selection Operator) amplamente discutido por Hastie, Tibshirani e Friedman (2009) como uma das abordagens mais eficientes para seleção automática de variáveis e redução de variância em modelos generalizados.

Com base nestas informações, o presente estudo propõe evidenciar e reforçar a identificação dos fatores associados a nefrotoxicidade associada ao TDF, por meio de modelos de Regressão Logística em conjunto ao método Lasso. A utilização conjunta da regressão logística e do método Lasso se justifica porque o desfecho analisado é binário — presença ou ausência de nefrotoxicidade — tornando a regressão logística o modelo adequado para estimar probabilidades (HOSMER; LEMESHOW; STURDIVANT, 2013). Entretanto, como há muitas variáveis clínicas, laboratoriais e demográficas potencialmente associadas ao TDF, o modelo pode sofrer instabilidade ou sobreajuste. Nesse contexto, o Lasso atua como um método de regularização que penaliza coeficientes, seleciona automaticamente as variáveis mais relevantes, reduz colinearidade e torna o modelo mais simples, robusto e clinicamente interpretável. Assim, a combinação desses métodos permite identificar, com maior precisão, os fatores realmente associados ao risco de nefrotoxicidade.

Com isso, o presente trabalho também atinge as duas metas da ODS: Meta 3 – “Saúde e bem-estar”, com a avaliação de fatores e efeitos adversos das ART, de forma que novas ferramentas de análise possam ser úteis em decisões clínicas, e a Meta 4 – “Educação Qualidade”, atingida ao aprender como organizar e depurar os dados, apresentar e interpretar dos resultados da área da saúde, trazendo novos horizontes que possam auxiliar nas decisões clínicas.

2 METODOLOGIA

Para o estudo, será considerado a base de dados disponibilizado no diretório Dryad , onde contém dados reais de 203 pacientes com informações médicas citadas na tabela a seguir. Há, no total, 26 variáveis:

Tabela 1 – Nome das variáveis, suas unidades de medida e descrição delas

Variável	Unidade	Definição
id	—	Identificador (Número sequencial)
sex	m/f	Sexo do participante (1 = masculino, 0 = feminino)
map	mmHg	Pressão arterial média
cd4	células/uL	Contagem basal de CD4
egfr1	mL/min/1.73m ³	Taxa de filtração glomerular estimada (EGFR) basal
egfr3	mL/min/1.73m ³	EGFR estimada após 3 meses
nephtox	yes/no	Desenvolveu nefrotoxicidade (0 = não, 1 = sim)
egfr10ct	mL/min/1.73m ³	EGFR dividido por 10
scr10ct	umol/L	Creatinina sérica dividida por 10
vl1	cópias/mL	Carga viral basal
scr1	umol/L	Creatinina sérica basal
akd35	yes/no	Queda > 35% na função renal (0 = não, 1 = sim)
akd50	yes/no	Aumento > 50% na creatinina (0 = não, 1 = sim)
scr120_1	yes/no	Creatinina basal > 120 (0 = não, 1 = sim)
egfr60_1	yes/no	EGFR basal < 60 (0 = não, 1 = sim)
viremia	viremia/no viremia	Carga viral < 50 (0 = sem viremia, 1 = com viremia)
supressed suppressed/unsuppressed		Carga viral > 1000 (0 = suprimida, 1 = não suprimida)

Variável	Unidade	Definição
chol1	mmol/L	Colesterol sérico basal
uacr1	mg/g	Razão albumina/creatinina na urina (basal)
phcr1	mol/mol	Razão fosfato/creatinina na urina (basal)
time	meses	Duração do acompanhamento até o desfecho
lnuacr1	—	Log natural da razão albumina/creatinina
lnvl1	—	Log natural da carga viral
glu1	mmol/L	Glicemia em jejum
bmi1	kg/m ²	Índice de massa corporal
age1	anos	Idade do paciente

Com base nestas variáveis, e alinhado ao objetivo do estudo, será considerada como variável resposta a nefrotoxicidade (nephtox). Para a análise, será considerado modelos de regressão logística para avaliar quais fatores estão mais associados a presença de nefrotoxicidade no paciente. Será considerada técnicas de regularização, em particular, o LASSO, com intuito de deixar o modelo adotado mais robusto e coerente com a literatura sobre a temática proposta.

2.1 Modelo de Regressão Logística

A regressão logística é um modelo estatístico amplamente utilizado para analisar situações em que a variável dependente é dicotômica, ou seja, assume apenas dois valores distintos, normalmente representados como 0 e 1. Esse tipo de regressão tem como objetivo modelar a probabilidade de ocorrência de um evento de interesse em função de um ou mais preditores, também chamados de variáveis explicativas ou covariáveis (HOSMER; LEMESHOW; STURDIVANT, 2013).

Na regressão logística múltipla, considera-se um vetor de covariáveis $x = (x_1, x_2, \dots, x_p)$, associado à probabilidade $\pi(x) = P(Y = 1 | x)$. O modelo é definido pela seguinte relação funcional:

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

A expressão acima representa a função logit, ou seja, o logaritmo da razão de chances (odds) da ocorrência do evento. A transformação logit garante que a relação entre a variável resposta e os preditores seja linear no log-odds, respeitando a natureza binária da variável resposta (HOSMER; LEMESHOW; STURDIVANT, 2013). A função inversa da logit é a função logística, que fornece a probabilidade estimada:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}$$

Os coeficientes β_p são estimados por método de máxima verossimilhança, cuja função é dada por:

$$\ell(\beta) = \sum_{i=1}^n [y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))]$$

Esse processo requer algoritmos iterativos, como Newton-Raphson ou IRLS (Iteratively Reweighted Least Squares), pois não existe uma solução analítica fechada para os coeficientes (HOSMER; LEMESHOW; STURDIVANT, 2013).

Cada coeficiente β_p tem uma interpretação direta: representa a mudança no logaritmo da razão de chances de $Y = 1$ para um aumento unitário em x_p , mantendo as demais variáveis constantes. A razão de chances associada à variável x_p é dada por:

$$OR_j = e^{\beta_p}$$

Apesar da robustez do modelo logístico múltiplo, sua aplicação em contextos de alta dimensionalidade (quando o número de variáveis explicativas é grande ou maior que o número de observações) pode levar a problemas como super ajuste, instabilidade nas estimativas e dificuldade de interpretação. Para lidar com esses desafios, há algumas técnicas de regularização, sendo o Lasso (Least Absolute Shrinkage and Selection Operator) uma das mais usadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2 Regularização L1: Lasso na Regressão Logística

O Lasso introduz uma penalização do tipo L1 à função de log-verossimilhança, forçando alguns coeficientes a se tornarem exatamente zero e, assim, realizando seleção de variáveis de forma embutida. A função objetivo passa a ser:

$$\max_{\beta_0, \beta} \sum_{i=1}^n \left[y_i(\beta_0 + \mathbf{x}_i^T \beta) - \log(1 + e^{\beta_0 + \mathbf{x}_i^T \beta}) \right] - \lambda \sum_{j=1}^p |\beta_j|$$

Nesta formulação:

- λ é o parâmetro de regularização, que controla o grau de penalização imposto aos coeficientes.
- A penalização incide apenas sobre os coeficientes β_j , não sobre o intercepto β_0 .

A principal característica do Lasso é a geração de modelos esparsos, isto é, modelos que contêm apenas um subconjunto das variáveis originalmente consideradas. Isso contribui para uma melhor interpretação do modelo, aumento da estabilidade das estimativas e prevenção do sobre ajuste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O valor ideal de λ é geralmente escolhido por meio de validação cruzada. A interpretação dos coeficientes não nulos em modelos com Lasso permanece a mesma da regressão logística tradicional, como também a obtenção deles.

3 RESULTADOS PARCIAIS

3.1 Análise Descritiva

Primeiramente, realizamos uma análise descritiva completa das variáveis presentes no banco de dados, contemplando tanto variáveis qualitativas quanto quantitativas, com o objetivo de caracterizar o perfil dos pacientes e comparar os grupos com e sem nefrotoxicidade. Para as variáveis qualitativas, examinamos a distribuição das frequências e proporções em cada categoria, enquanto, para as variáveis quantitativas, avaliamos medidas de tendência central e dispersão, além da verificação da normalidade dos dados para orientar a escolha dos testes estatísticos. Em seguida, procedemos com a criação dos modelos (Modelo Completo Padronizado, Modelo Completo e Modelo Completo Padronizado StepWise), avaliando suas estatística e as ODDS. Após isso, avaliamos as variáveis que foram significativas e fizemos o diagnóstico do modelo. Em seguida, temos os mesmos passos para o modelo Lasso, e comparação ao modelo escolhido na primeira etapa.

Começando com as variáveis qualitativas (Na base, existem 8):

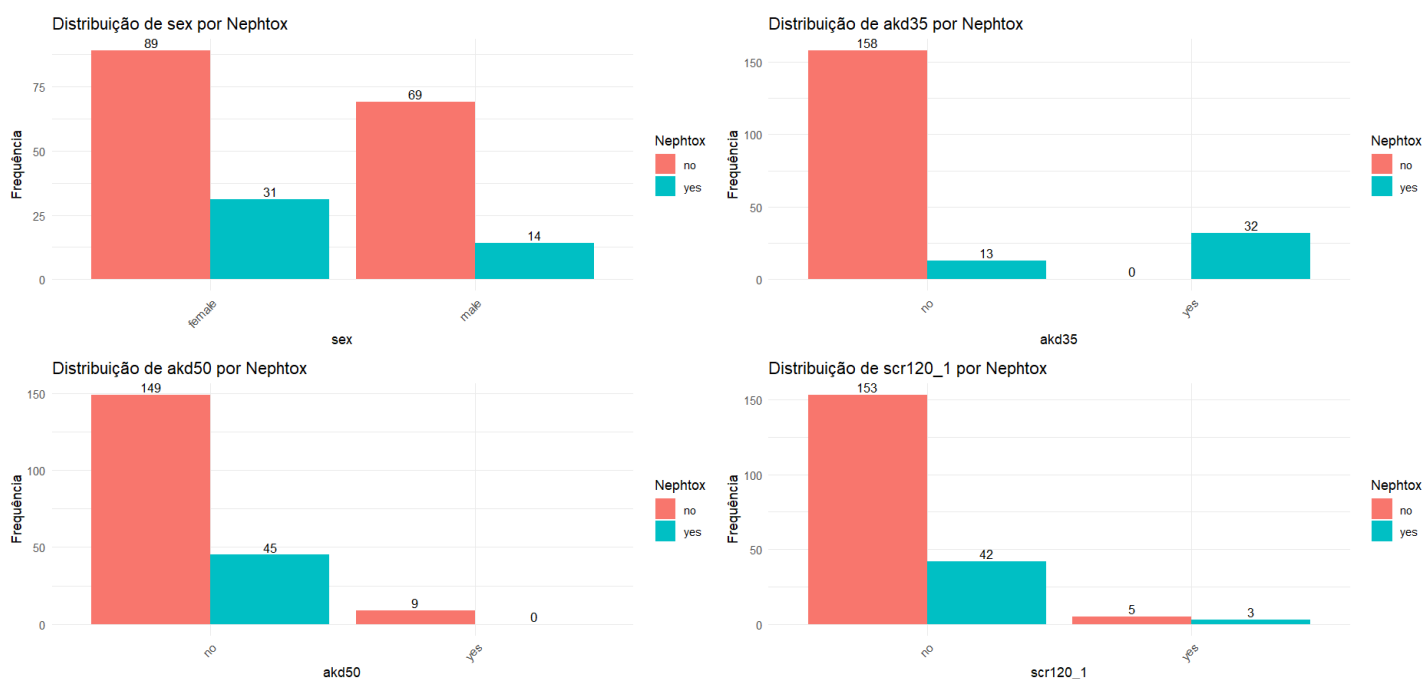
Tabela 2 - Características clínicas dos pacientes (em relação as variáveis qualitativas) por Nefrotoxicidade.

Variável	Categoria	Não (n=158)	Sim (n=45)
Sexo	Feminino	89 (56,3%)	31 (68,9%)
	Masculino	69 (43,7%)	14 (31,1%)
AKD35	Não	158 (100%)	13 (28,9%)
	Sim	0 (0%)	32 (71,1%)
AKD50	Não	149 (94,3%)	45 (100%)
	Sim	9 (5,7%)	0 (0%)
SCR120_1	Não	153 (96,8%)	42 (93,3%)
	Sim	5 (3,2%)	3 (6,7%)
eGFR60_1	Não	154 (97,5%)	43 (95,6%)
	Sim	4 (2,5%)	2 (4,4%)
Viremia	Supressed	30 (19,0%)	8 (17,8%)
	Unsupressed	128 (81,0%)	37 (82,2%)
Supressed	Supressed	25 (15,8%)	7 (15,6%)
	Unsupressed	133 (84,2%)	38 (84,4%)
BMI1	Normal	94 (59,5%)	23 (51,1%)
	Obeso	9 (5,7%)	5 (11,1%)
	Overweight	46 (29,1%)	8 (17,8%)
	Underweight	9 (5,7%)	9 (20,0%)

A análise das variáveis categóricas em relação à nefrotoxicidade mostra que, de maneira geral, a distribuição das categorias é relativamente equilibrada entre os grupos, com algumas exceções. Quanto ao sexo, a maior parte dos indivíduos é do sexo feminino, sendo 89 no grupo sem nefrotoxicidade e 31 no grupo com nefrotoxicidade, enquanto os homens representam 69 e 14, respectivamente, indicando proporções semelhantes entre os grupos. Em relação a AKD35, todos os casos positivos estão no grupo com nefrotoxicidade, sugerindo uma associação clara, enquanto em AKD50 a maioria dos casos positivos está no grupo sem nefrotoxicidade.

Para os indicadores laboratoriais categorizados, como SCR120_1 e eGFR60_1, a maior parte dos indivíduos se encontra nas categorias “Não”, embora haja alguns casos positivos nos dois grupos. Quanto à viremia e à supressão viral, a maioria dos pacientes apresenta status não suprimido, com distribuição relativamente semelhante entre os grupos, indicando que a supressão viral não difere significativamente. Por fim, em relação ao índice de massa corporal (BMI1), a maioria dos indivíduos está na faixa normal, seguida das categorias overweight e obeso, com proporções ligeiramente diferentes entre os grupos devido ao tamanho do grupo com nefrotoxicidade ser menor.

Figura 1 – Gráfico de barras da relação das variáveis com a nefrotoxicidade

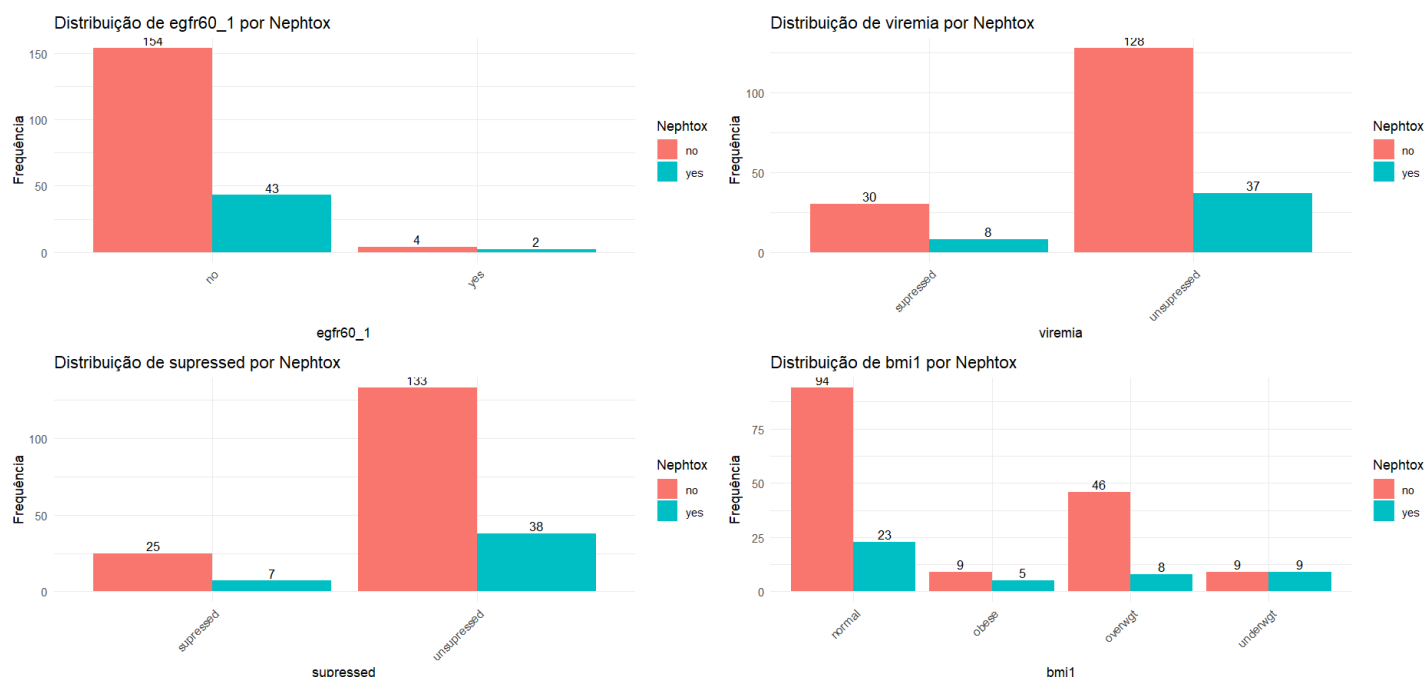


A Figura 1, em relação ao “sexo”, observa-se que a maioria dos pacientes com nefrotoxicidade é do sexo feminino, enquanto apenas 14 são do sexo masculino. Entre os pacientes sem nefrotoxicidade, a maioria também é do sexo feminino.

Ao observar a variável “akd35”, que indica uma queda superior a 35% na função renal, 32 dos pacientes com nefrotoxicidade apresentaram essa queda, enquanto entre os pacientes sem nefrotoxicidade, nenhum apresentou “akd35”. Essa tendência se repete na variável “akd50”, onde todos os 45 pacientes com nefrotoxicidade não apresentaram aumento superior a 50% na creatinina, enquanto 9 dos pacientes sem nefrotoxicidade apresentaram essa elevação.

Em relação à variável “scr120_1”(nível elevado de creatinina sérica aos 120 minutos), nota-se que a maioria dos pacientes com nefrotoxicidade não apresentaram elevação, e apenas 3 sim. Já entre os sem nefrotoxicidade, 153 não apresentaram elevação e 5 sim.

Figura 2 – Gráfico de barras da relação das variáveis com a nephtoxicidade



Com relação a Figura 2, à variável “egfr60_1” indica que a maioria dos pacientes com nefrotoxicidade apresentou função renal reduzida (observando a categoria “no”), enquanto apenas 2 apresentaram uma taxa de filtração glomerular estimada superior a 60 (categoria “yes”).

A variável “viremia” mostra uma tendência semelhante: entre os pacientes com nefrotoxicidade, a maioria apresentava viremia não suprimida, enquanto apenas 8 tinham viremia suprimida. Já entre os pacientes sem nefrotoxicidade, a maior parte também estava na categoria “unsuppressed”, mas a proporção de pacientes com viremia suprimida é relativamente maior.

De maneira similar, a variável “supressed”, mostra que a maioria dos pacientes com nefrotoxicidade não apresentava supressão viral, enquanto apenas 7 tinham supressão, reforçando a observação anterior. Já no grupo sem nefrotoxicidade, 25 pacientes estavam suprimidos e 133 não.

Por fim, a distribuição do IMC (“bmi1”) aponta que a maior parte dos pacientes com nefrotoxicidade tinha IMC normal, seguido por casos distribuídos entre as categorias “underweight”, “overwgt” e “obese”, com valores semelhantes. No grupo

sem nefrotoxicidade, há predominância de pacientes com IMC normal, ou com sobrepeso.

Partindo para as variáveis quantitativas (16 variáveis):

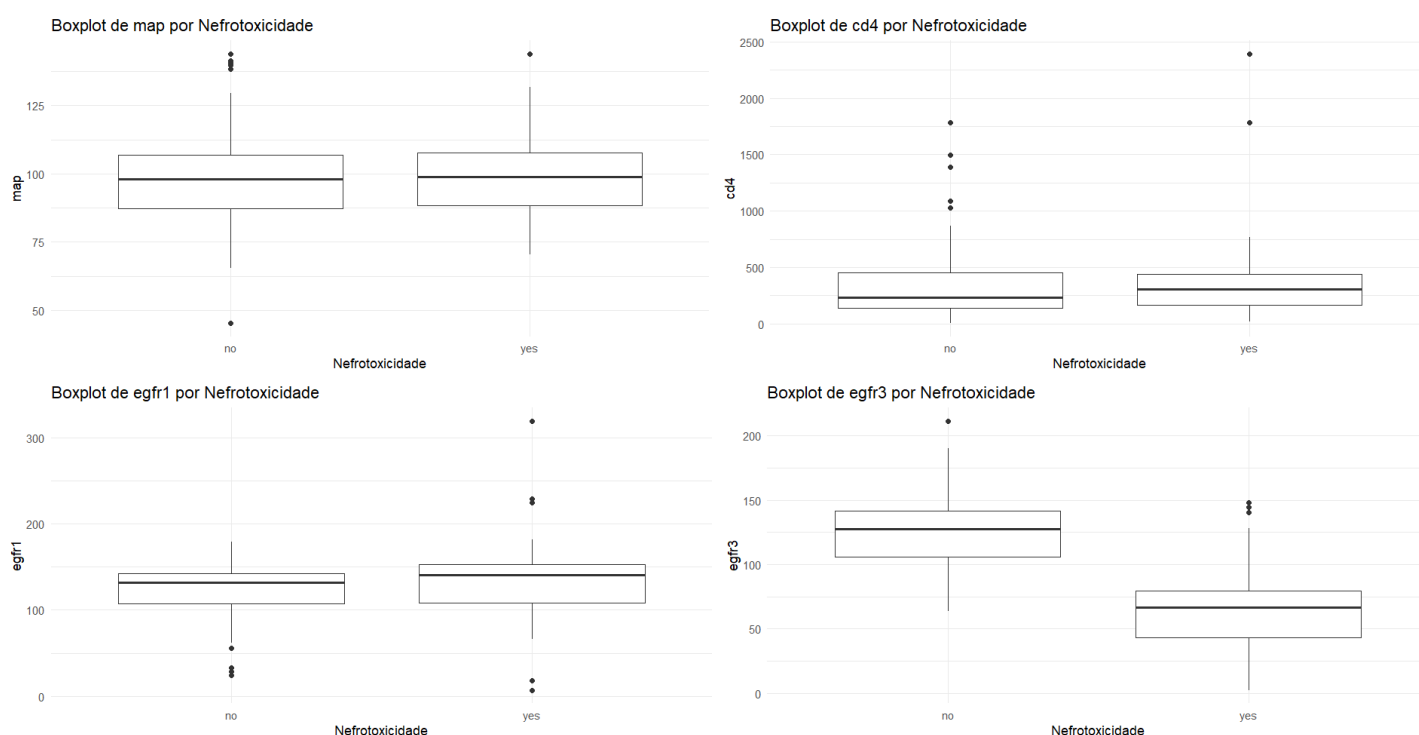
Tabela 3 - Características clínicas dos pacientes (em relação as variáveis qualitativas) por Nefrotoxicidade.

Variável	Não (média \pm dp) (158)	Sim (média \pm dp) (45)
MAP (mmHg)	98,4 \pm 16,6	101,0 \pm 16,7
CD4 (cél/mm ³)	334 \pm 304	372 \pm 420
eGFR1 (mL/min)	124 \pm 28,3	132 \pm 53,2
eGFR3 (mL/min)	124 \pm 25,6	66,5 \pm 38,0
eGFR10ct	12,4 \pm 2,83	13,2 \pm 5,32
SCR10ct	6,90 \pm 2,95	7,74 \pm 9,90
VL1	410.818 \pm 1.301.222	228.731 \pm 438.140
SCR1 (μ mol/L)	69,0 \pm 29,5	77,4 \pm 99,0
CHOL (mmol/L)	4,13 \pm 1,29	3,84 \pm 1,26
UACR1	8.122 \pm 71.993	14.008 \pm 90.984
PHCR (mg/L)	3,61 \pm 11,2	5,10 \pm 17,2
Time	3,19 \pm 0,998	3,01 \pm 0,989
lnUACR	4,94 \pm 1,77	5,12 \pm 1,85
lnVL	9,68 \pm 4,30	9,69 \pm 4,06
GLU (mmol/L)	4,43 \pm 0,882	4,32 \pm 1,03
AGE (anos)	35,1 \pm 8,09	36,7 \pm 9,89

A análise das variáveis numéricas em relação à nefrotoxicidade mostra que, de forma geral, os valores médios foram bastante semelhantes entre os grupos com e sem nefrotoxicidade. A pressão arterial média foi de 98,4 \pm 16,6 mmHg no grupo sem nefrotoxicidade e de 101,0 \pm 16,7 mmHg no grupo com nefrotoxicidade. O número de células CD4 apresentou média de 334 \pm 304 no grupo sem nefrotoxicidade e 372 \pm 420 no grupo com nefrotoxicidade, indicando grande variabilidade entre os indivíduos. A taxa de filtração glomerular (eGFR) apresentou pequenas diferenças entre os grupos, assim como a creatinina sérica (SCR), com médias de 69,0 \pm 29,5 μ mol/L no

grupo sem nefrotoxicidade e $77,4 \pm 99,0$ $\mu\text{mol/L}$ no grupo com nefrotoxicidade. Outras variáveis laboratoriais, como colesterol, proteína C reativa e relação albumina/creatinina urinária, também apresentaram médias ligeiramente diferentes entre os grupos, mas com alta dispersão, o que sugere ausência de diferenças consistentes. A carga viral log-transformada (lnVL) e a glicemia foram praticamente iguais entre os grupos. Por fim, a idade média foi levemente superior no grupo com nefrotoxicidade ($36,7 \pm 9,89$ anos) em comparação ao grupo sem nefrotoxicidade ($35,1 \pm 8,09$ anos).

Figura 3 – Boxplot da relação das variáveis com a nefrotoxicidade



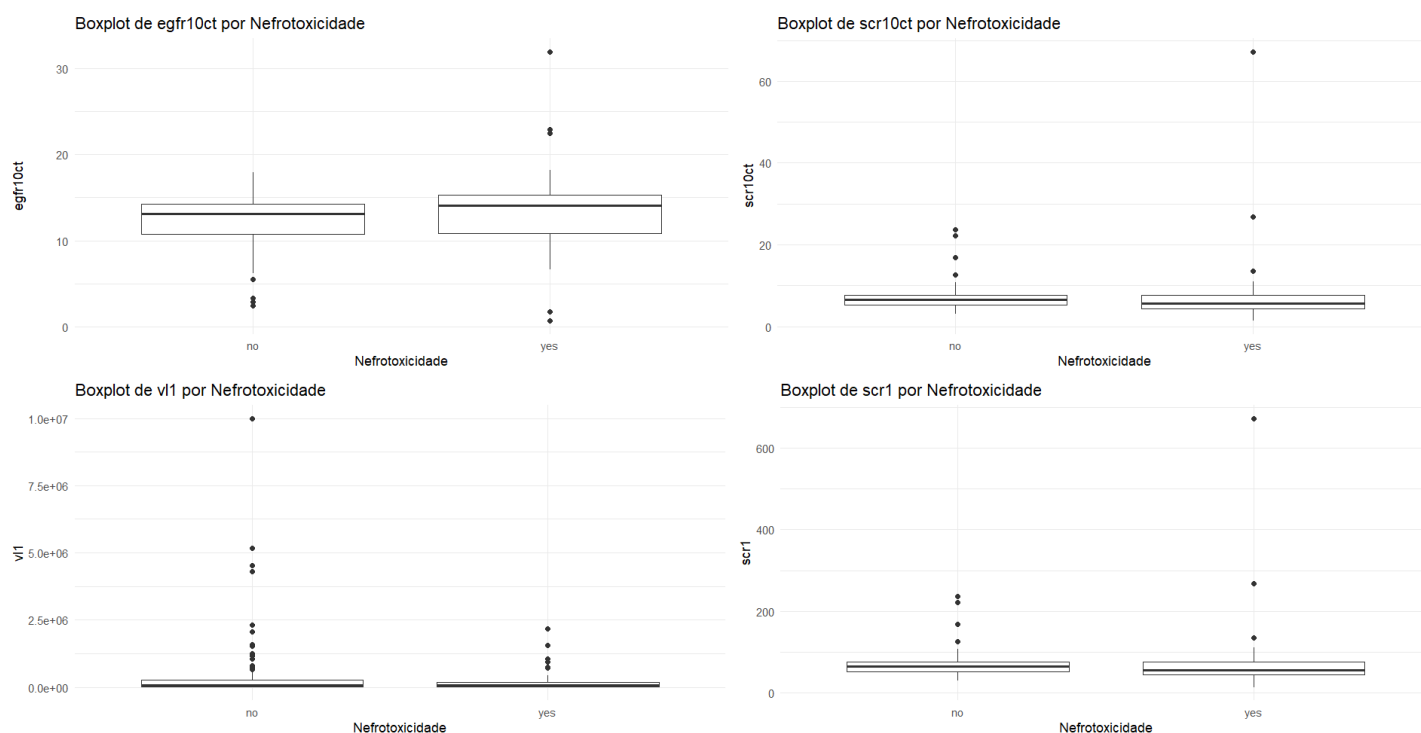
Na Figura 3, na variável “map” (pressão arterial média), observa-se que os indivíduos com e sem nefrotoxicidade apresentam distribuições semelhantes, com medianas próximas e amplitudes similares. No entanto, há presença de alguns outliers nos dois grupos, o que indica variações pontuais de pressão.

Para a variável “cd4”, os valores são bastante dispersos em ambos os grupos, com presença de muitos outliers, especialmente entre os indivíduos sem nefrotoxicidade. Apesar disso, as medianas são próximas.

A variável “egfr1” (taxa de filtração glomerular estimada no tempo 1) apresenta distribuição levemente mais baixa no grupo com nefrotoxicidade, com uma mediana ligeiramente inferior e maior variabilidade dos valores.

Já na variável “egfr3” (filtração glomerular em outro momento), a diferença entre os grupos é mais evidente. Indivíduos com nefrotoxicidade apresentam valores medianos mais baixos e maior assimetria negativa, com outliers puxando a distribuição para valores extremamente baixos. Isso reforça a hipótese de comprometimento da função renal ao longo do tempo no grupo com nefrotoxicidade.

Figura 4 – Boxplot da relação das variáveis com a nefrotoxicidade



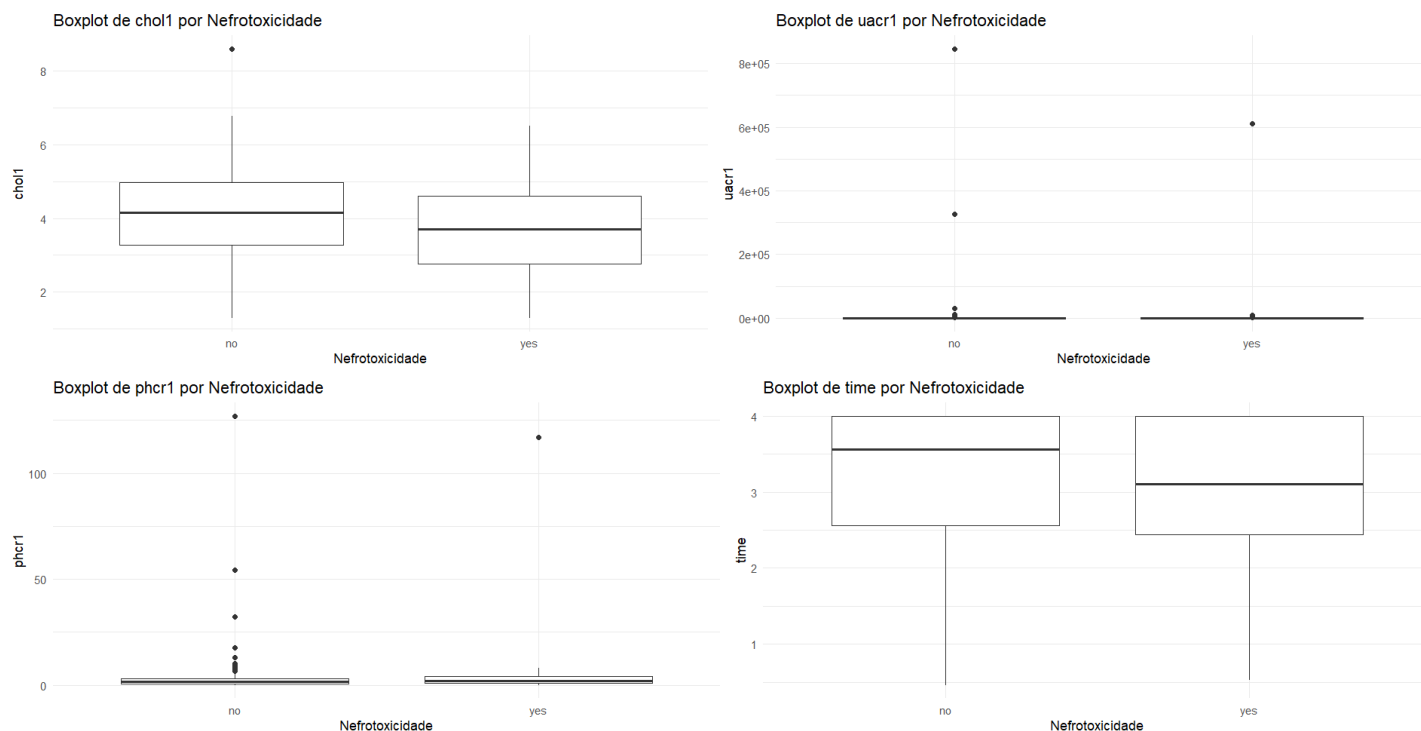
A Figura 4, na variável “egfr10ct”, que representa a taxa de filtração glomerular em um determinado momento, apresenta distribuição semelhante nos dois grupos, com medianas próximas e dispersão parecida. Apesar disso, há outliers em ambos os grupos, indicando variabilidade individual.

Na variável “scr10ct” (creatinina sérica no mesmo ponto de “egfr10ct”), observa-se que os valores tendem a ser ligeiramente mais elevados nos indivíduos com nefrotoxicidade..

Para “vl1” (possivelmente carga viral), há muitos valores extremos em ambos os grupos, especialmente entre indivíduos sem nefrotoxicidade.

A variável “scr1” (creatinina sérica em um outro momento) também apresenta padrão semelhante ao de “scr10ct”.

Figura 5 – Boxplot da relação das variáveis com a nefrotoxicidade



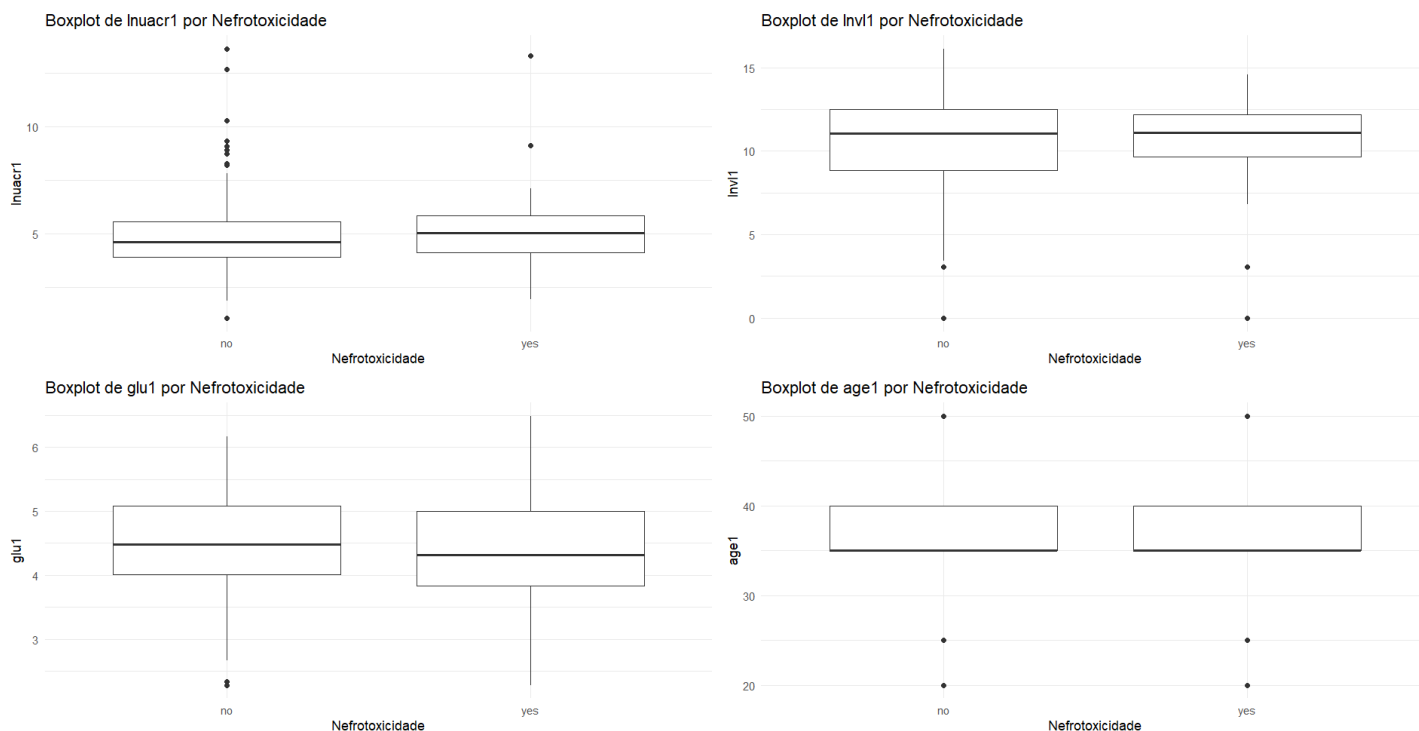
A Figura 5, na variável “chol1” (colesterol) apresenta distribuições semelhantes entre os grupos. A mediana é levemente menor no grupo com nefrotoxicidade.

Para “uacr1” (relação albumina/creatinina urinária), há muitos valores extremos, especialmente no grupo sem nefrotoxicidade. Apesar disso, a mediana é próxima de zero para ambos os grupos, indicando que a maioria dos indivíduos apresenta valores baixos dessa razão, dificultando a diferenciação entre os grupos.

A variável “phcr1”, também apresenta mediana próxima entre os grupos, com grande assimetria e presença de outliers, especialmente no grupo sem nefrotoxicidade.

Já “time” (tempo até evento ou de acompanhamento, presumivelmente), apresenta distribuição ligeiramente diferente: a mediana no grupo com nefrotoxicidade parece menor.

Figura 6 – Boxplot da relação das variáveis com a nefrotoxicidade



A Figura 6, na variável “lnuacr1”, que representa o logaritmo natural da razão albumina/creatinina urinária, mostra medianas próximas entre os dois grupos, com leve tendência de valores ligeiramente maiores no grupo com nefrotoxicidade. No entanto, há sobreposição significativa dos quartis e presença de outliers nos dois grupos, o que indica que a variável isoladamente pode não discriminar bem os grupos.

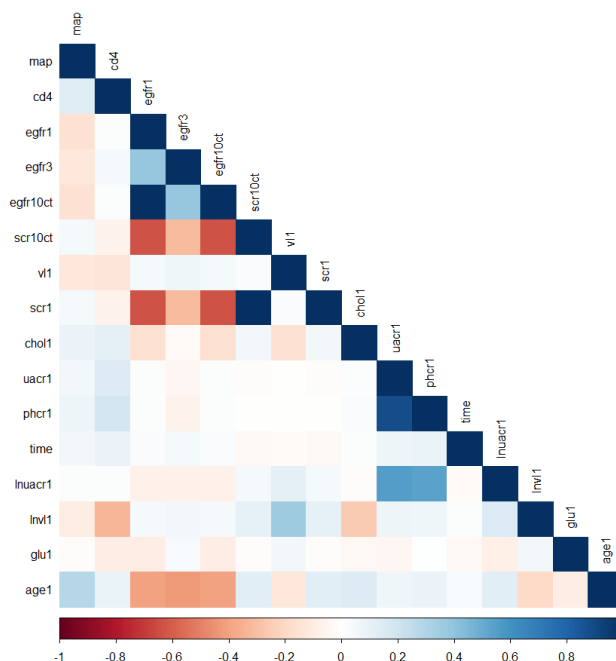
Para a variável “lnv1”, nota-se também uma distribuição bastante similar entre os grupos, embora o grupo com nefrotoxicidade apresente uma leve concentração maior de valores mais altos.

A variável “glu1” (glicose) mantém uma distribuição quase idêntica entre os grupos, com medianas e amplitudes próximas, reforçando a baixa capacidade discriminatória dessa variável.

Já “age1” (idade) também não apresenta diferenças visuais marcantes, com distribuições bastante próximas entre os dois grupos.

Vamos analisar as correlações de cada variável:

Figura 7 – Matriz de correlação das variáveis



Na Figura 7, podemos observar correlações negativas fortes entre as variáveis “scr1” e “egfr1”/“egfr3”. A variável “scr10ct” também apresenta correlação negativa forte com “egfr10ct” e positiva com “scr1”. A variável “egfr10ct” também apresenta correlação positiva com a “egfr1”. Sendo assim, temos que tomar um cuidado a mais com elas na hora de aplicarmos o modelo.

3.2 Modelo de Regressão Logística

Para esta parte, iremos analisar 4 modelos: 2 modelos completos (um sendo padronizado e o outro não), para termos um panorama, e 2 modelos usando métodos de seleção de variáveis (um sendo padronizado e o outro não) que foram encontrados a partir dos modelos completos. A padronização foi feita para as variáveis quantitativas, sendo que é pego a média, subtrai pelo valor e divide pelo desvio padrão. Para os modelos completos, retiramos 3 variáveis (“egfr1”, “egfr3” e “scr1”), pois existe dentro da base de dados, variáveis que dão a mesma informação em formato log. Entretanto, para os modelos padronizados, retiramos as 3 ditas anteriormente, e outras 8 (“v11”, “akd35”, “akd50”, “scr120_1”, “egfr60_1”, “viremia”, “uacr1”, “time”). O método utilizado para seleção de variáveis foi o StepWise. Nestes métodos, ele tenta encontrar o melhor modelo com menor erro possível, avaliando seu AIC.

Vamos avaliar o modelo completo padronizado:

Tabela 3 – Estatísticas das variáveis (Modelo Completo Padronizado)

Variável	Estimativa (β)	Erro Padrão	Valor-p
(Intercepto)	-1.21128	1.02001	0.23502
sex (masculino)	-1.23819	0.46568	0.00784
map	0.26772	0.20260	0.18636
cd4	-0.05966	0.19304	0.75726
egfr10ct	1.07623	0.50698	0.03377
scr10ct	0.94092	0.64781	0.14637
supressed (não)	0.31047	1.11402	0.78048
chol1	-0.28949	0.20426	0.15641
phcr1	0.03479	0.21654	0.87235
lnuacr1	0.05528	0.23671	0.81535
lnvl1	-0.20688	0.42829	0.62907
glu1	0.09840	0.19723	0.61784
bmi1 (obeso)	0.73147	0.80898	0.36590
bmi1 (sobrepeso)	-0.56387	0.49978	0.25922
bmi1 (baixo peso)	1.40720	0.58665	0.01645
age1	0.61534	0.26025	0.01806

Na Tabela 2, os resultados mostraram que algumas variáveis foram significativamente associadas (marcadas em negrito), sabendo que nosso nível de significância é 0.05.

O sexo masculino apresentou um coeficiente negativo significativo, indicando que homens têm menor chance de desenvolver nefrotoxicidade em comparação às mulheres. Em termos de razão de chances, esse resultado representa uma redução de aproximadamente 71% na chance de nefrotoxicidade para indivíduos do sexo masculino.

A variável egfr10ct apresentou um coeficiente positivo significativo, o que indica que valores mais altos dessa variável estão associados a um aumento na chance de nefrotoxicidade. Especificamente, cada aumento de uma unidade padronizada em egfr10ct quase triplica a chance de ocorrência do desfecho.

Da mesma forma, o grupo de indivíduos classificados como abaixo do peso (bmi1 baixo peso) teve associação positiva significativa com nefrotoxicidade, indicando que pessoas com baixo peso têm cerca de quatro vezes mais chance de desenvolver o evento em comparação ao grupo de referência (provavelmente o grupo com IMC normal).

Além disso, a variável idade (age1) também foi significativamente associada ao desfecho, sugerindo que indivíduos mais velhos têm maior probabilidade de desenvolver nefrotoxicidade. Um aumento de um desvio-padrão na idade implica em um aumento de aproximadamente 85% na chance do desfecho.

Para o modelo completo:

Tabela 4 – Estatísticas das variáveis (Modelo Completo)

Variável	Estimativa (β)	Erro Padrão	Valor-p
(Intercepto)	-20.470	22,65	0.3661
sex (masculino)	-1.967	2,240	0.3800
map	0.0411	0,03005	0.1717
cd4	0.0015	0,0009783	0.1142
egfr10ct	0.4922	0,8987	0.5840
scr10ct	0.2043	1,147	0.8586
vl1	-0.000001478	0,000002228	0.5072
akd35 (sim)	832.9	1902	0.6615
akd50 (sim)	-33.92	10970	0.9975
scr120_1 (sim)	-390.0	4,745e+07	1.0000
egfr60_1 (sim)	408.3	4,745e+07	1.0000
viremia (não suprimida)	-2.232	2,357	0.3436
supressed (não)	3.347	2,912	0.2504
chol1	-0.8638	0,3988	0.0303
uacr1	-0.0013	0,001244	0.2833
phcr1	-0.1021	0,1709	0.5502
time	-0.5807	0,4342	0.1811

Variável	Estimativa (β)	Erro Padrão	Valor-p
Inuacr1	1.008	0,6615	0.1276
Invl1	0.0091	0,3821	0.9810
glu1	-0.5374	0,5277	0.3085
bmi1 (obeso)	0.5625	1,947	0.7727
bmi1 (sobrepeso)	0.7045	1,237	0.5690
bmi1 (baixo peso)	3.181	1,468	0.0302
age1	0.2039	0,1327	0.1244

Pela Tabela 4, a variável colesterol total (“chol1”) apresentou uma associação negativa significativa, indicando que valores mais altos de colesterol total estão associados a uma menor chance de ter nefrotoxicidade. Já a variável “baixo peso” (bmi1 baixo peso) também foi estatisticamente significativa, mas com associação positiva, o que sugere que indivíduos classificados como abaixo do peso apresentaram maior chance de se obter a nefrotoxicidade, comparados ao grupo de referência (peso normal). Para este modelo, pelo motivo dos erros estarem muito altos ou muito baixos, não iremos trabalhar com ele.

Vamos observar os modelos com o método de seleção StepWise:

Tabela 5 – Estatísticas das variáveis (Modelo Completo Padronizado)
StepWise

Variável	Estimativa (β)	Erro Padrão	p-valor
(Intercepto)	-0.9935	0.3060	0.00117
sexmale	-1.1434	0.4353	0.00862
egfr10ct	0.9335	0.4330	0.03110
scr10ct	0.7589	0.5300	0.15218
chol1	-0.2769	0.1984	0.16286
bmi1obese	0.8495	0.7017	0.22602
bmi1overwgt	-0.4768	0.4767	0.31725
bmi1 (baixo peso)	1.2591	0.5715	0.02758
age1	0.6581	0.2398	0.00606

Observando a Tabela 5, temos as mesmas interpretações da Tabela 3, com uma única diferença: O AIC do modelo da Tabela 2 é de 215.65, enquanto o do StepWise, 204.08.

Observando as ODDS:

Tabela 6 – Odds Ratio das variáveis (Modelo Completo padronizado) StepWise

Variável	OR (Razão de Chances)	IC 2.5%	IC 97.5%	Interpretação
(Intercepto)	0.370	0.202	0.674	-
sexmale	0.319	0.129	0.718	Homens têm menor chance de nefrotoxicidade do que mulheres
egfr10ct	2.543	1.297	6.866	A cada aumento de 10 unidades no eGFR, há maior chance de nefrotoxicidade
scr10ct	2.136	1.145	7.573	Aumento da creatinina sérica está associado à maior chance de nefrotoxicidade
chol1	0.758	0.508	1.110	Redução não significativa da chance com aumento do colesterol
bmi1obese	2.338	0.542	8.970	Obesidade parece aumentar chance, mas sem significância estatística
bmi1overwgt	0.621	0.232	1.531	Sobrepeso pode reduzir a chance, mas sem significância
bmi1underwgt	3.522	1.129	10.904	Baixo peso aumenta significativamente a chance de nefrotoxicidade
age1	1.931	1.235	3.189	Idade mais avançada aumenta chance de nefrotoxicidade

Na Tabela 6, a variável sexo masculino apresentou OR = 0.319 (IC95%: 0.129–0.718), indicando que homens possuem aproximadamente 68% menos chances de desenvolver nefrotoxicidade em comparação com mulheres, de forma estatisticamente significativa.

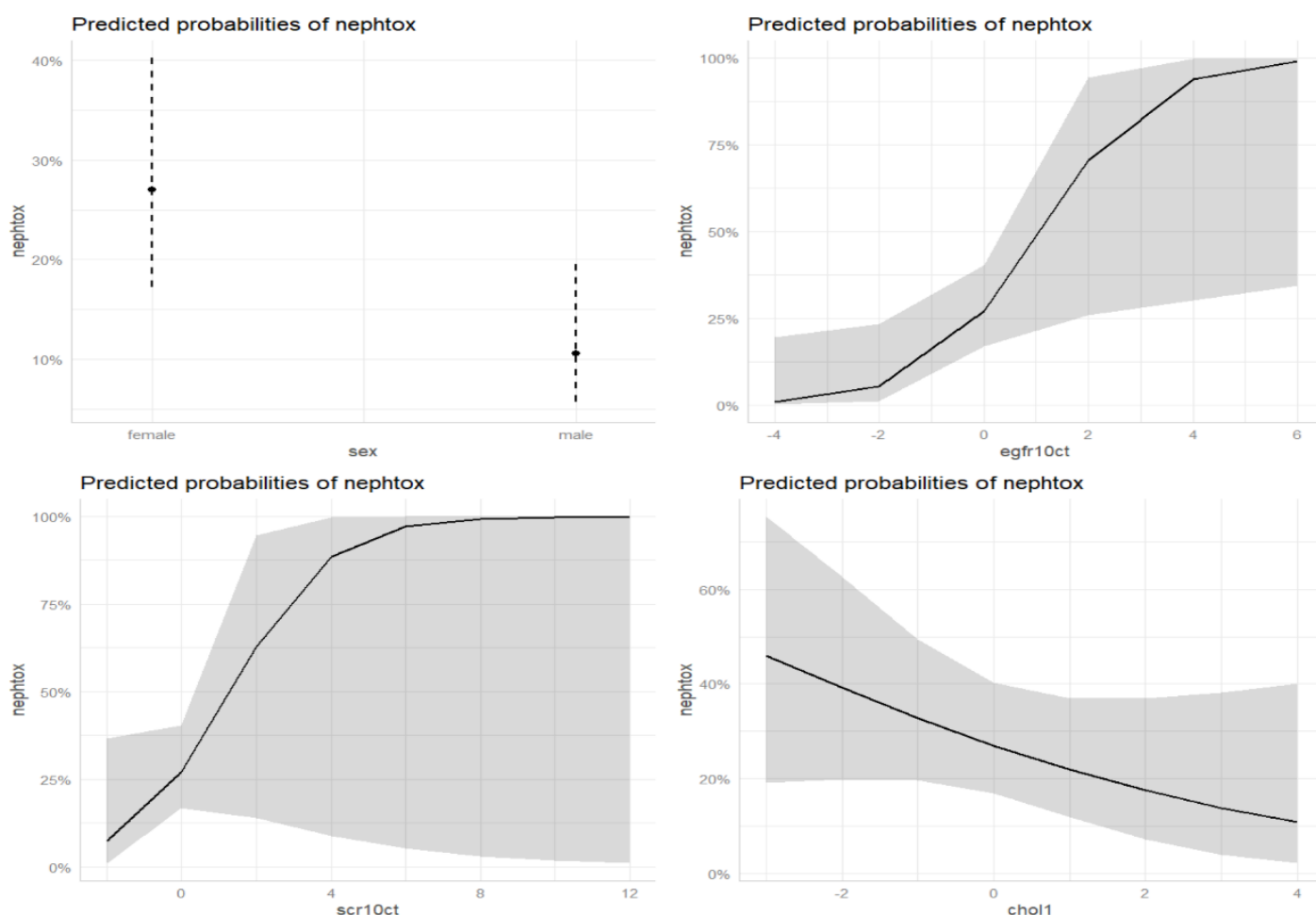
Tanto a taxa de filtração glomerular (“egfr10ct”) quanto a creatinina sérica (“scr10ct”) foram associadas a um maior risco de nefrotoxicidade. Para o “egfr10ct”, cada aumento de 10 unidades eleva a chance em mais de 2,5 vezes, e para “scr10ct”, o aumento da creatinina também está ligado a um risco significativamente maior.

A idade (“age1”) foi outro fator de risco importante: indivíduos mais velhos têm quase o dobro da chance de desenvolver nefrotoxicidade.

O baixo peso (“bmi1underwt”) se destacou como um fator de risco relevante: indivíduos com IMC abaixo do normal apresentaram mais de 3,5 vezes mais chance de apresentar nefrotoxicidade em relação aos de peso normal.

Vamos avaliar isso graficamente:

Figura 8 – Probabilidades Previstas de Nefrotoxicidade (Modelo padronizado StepWise)



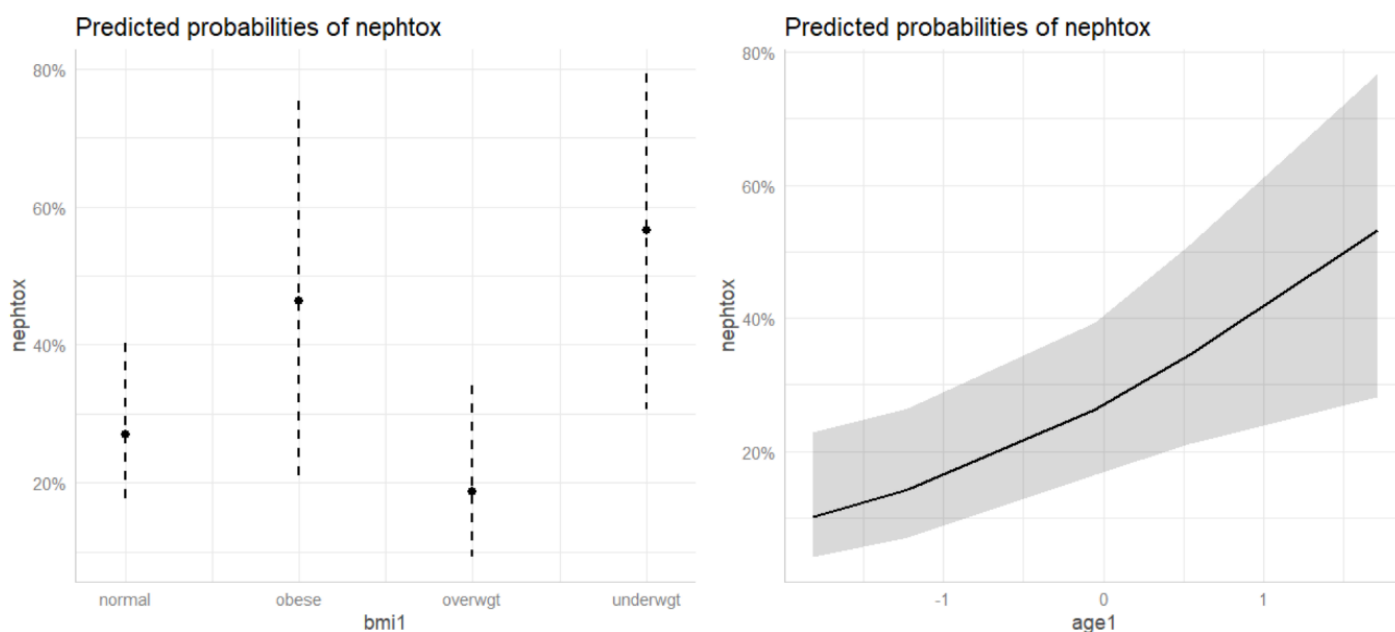
Na Figura 8, observa-se que o sexo feminino apresenta maior probabilidade de desenvolver nefrotoxicidade, com uma estimativa central perto de 30% e intervalo de confiança entre aproximadamente 20% e 40%, enquanto homens apresentam probabilidade menor, em torno de 10%, com intervalo entre abaixo de 10% e 20%. Isso indica uma associação entre o sexo feminino e maior risco de nefrotoxicidade.

No segundo gráfico, a variável “egfr10ct”, apresenta uma relação direta com a probabilidade de nefrotoxicidade: quanto maior o valor de “egfr10ct”, maior a probabilidade estimada. A curva cresce progressivamente, atingindo valores próximos de 100% de probabilidade conforme “egfr10ct” aumenta.

O terceiro gráfico mostra a relação entre “scr10ct” (creatinina sérica padronizada) e a probabilidade de nefrotoxicidade. Verifica-se que à medida que scr10ct aumenta, a probabilidade de nefrotoxicidade sobe rapidamente até atingir um platô próximo de 100%.

Por fim, o quarto gráfico apresenta a relação entre “chol1” (colesterol padronizado) e a probabilidade de nefrotoxicidade. Nesse caso, há uma tendência de diminuição da probabilidade com o aumento do colesterol. Ou seja, níveis mais altos de colesterol parecem estar associados a menor risco de nefrotoxicidade.

Figura 9 – Probabilidades Previstas de Nefrotoxicidade (Modelo padronizado StepWise)

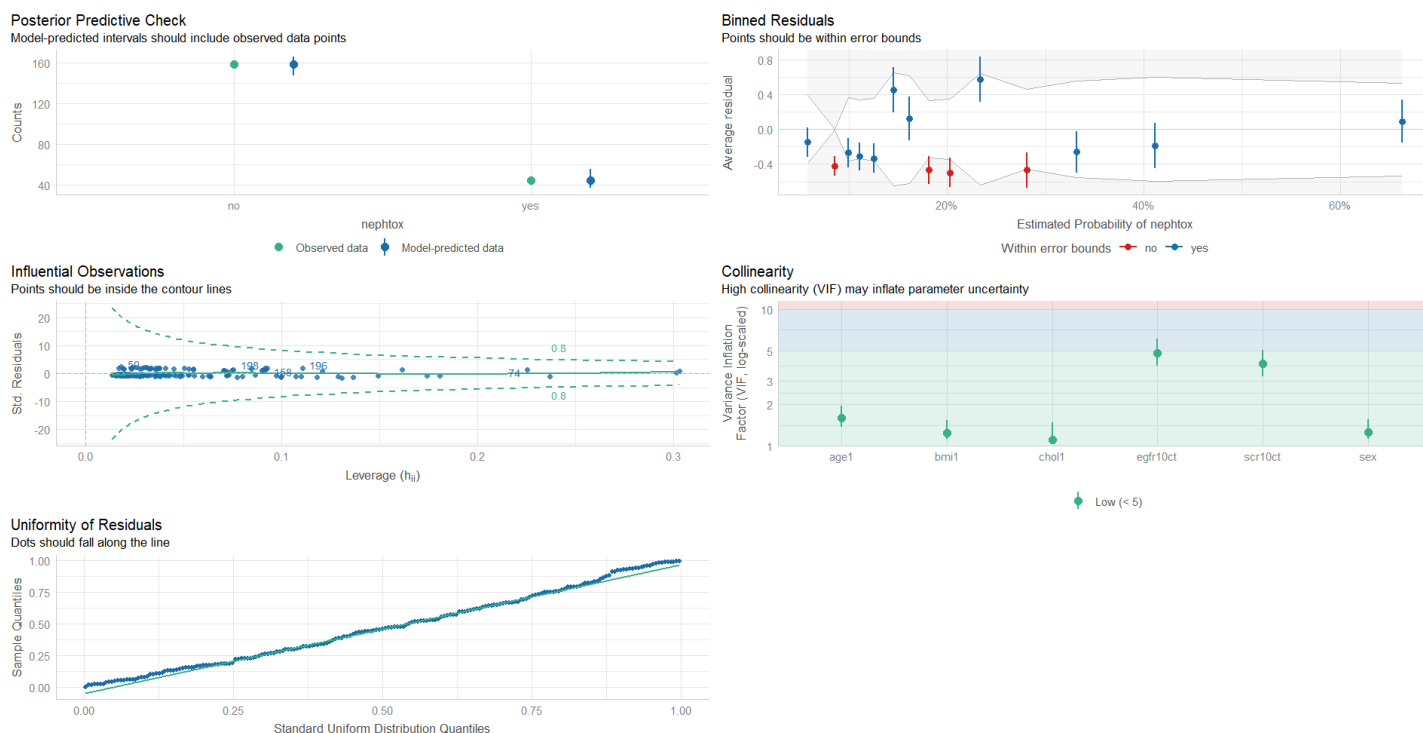


Na Figura 9, observa-se a relação entre as categorias de IMC e a probabilidade predita de nefrotoxicidade. Indivíduos classificados como “obesos” apresentam a maior probabilidade estimada de nefrotoxicidade, em torno de 50%, com um intervalo de confiança que se estende de aproximadamente 25% a 75%. Em seguida, estão os indivíduos com baixo peso (“underwgt”), com uma probabilidade predita de cerca de 55%, também com grande incerteza (IC amplo). Os indivíduos com IMC normal apresentam uma probabilidade intermediária, enquanto aqueles com sobrepeso (“overwgt”) têm a menor probabilidade predita. Esses resultados sugerem que tanto o excesso de peso extremo quanto o baixo peso pode estar associados a um risco aumentado de toxicidade, enquanto o sobrepeso.

O segundo gráfico mostra que, para valores negativos de idade padronizada (indivíduos mais jovens), a probabilidade estimada é inferior a 20%. À medida que a idade aumenta, essa probabilidade sobe progressivamente, alcançando aproximadamente 55% para os indivíduos mais velhos.

Avaliando o diagnóstico do modelo:

Figura 10 – Diagnóstico do modelo completo padronizado StepWise



Na Figura 10, o gráfico de "Posterior Predictive Check" mostra uma verificação da previsão posterior, onde os intervalos previstos pelo modelo devem conter os dados observados. Como os pontos previstos estão próximos dos observados para

ambas as categorias (sim e não para nefrotoxicidade), isso indica que o modelo está reproduzindo bem os dados reais.

O gráfico de "Binned Residuals" avalia os resíduos médios em diferentes faixas de probabilidade estimada para nefrotoxicidade. Idealmente, os resíduos (diferença entre observado e previsto) devem estar dentro dos limites de erro (área sombreada). No entanto, observa-se que alguns pontos vermelhos (casos com nefrotoxicidade) estão fora dos limites, o que sugere que o modelo pode estar subestimando a probabilidade de nefrotoxicidade em certas faixas.

O gráfico de "Influential Observations" mostra os resíduos padronizados em função da alavancagem (leverage). Os pontos devem estar dentro das linhas de contorno verdes (limites aceitáveis). A maioria dos pontos está dentro dessas linhas, o que indica que não há observações altamente influentes que estejam distorcendo o modelo de forma significativa.

O gráfico "Collinearity", avalia a multicolinearidade entre as variáveis explicativas usando o Fator de Inflação da Variância (VIF). Valores acima de 5 (área azul) indicam colinearidade elevada, o que pode inflacionar a incerteza dos parâmetros. Aqui, todos os VIFs estão abaixo de 5 (área verde), indicando baixa colinearidade e, portanto, maior confiabilidade nas estimativas dos coeficientes.

Enfim, o gráfico "Uniformity of Residuals" compara os quantis dos resíduos com uma distribuição uniforme padrão. Os pontos devem alinhar-se com a linha diagonal se os resíduos estiverem bem distribuídos. A boa aderência dos pontos à linha sugere que os resíduos seguem a distribuição esperada, indicando ausência de padrões sistemáticos não capturados pelo modelo.

3.3 Modelo de Regressão Logística (Aplicando Lasso)

Com o modelo aplicando Lasso, foi identificado algumas possíveis variáveis que explicaria a nefrotoxicidade. Por meio da OR, excluímos as possíveis variáveis que não nos agregariam tanto (sex, akd35, akd50, scr120_1, egfr60_1, viremia, suppressed, phcr1, bmi1, age1, egfr1, egfr3 e scr1). Segue os resultados:

Tabela 7 – Estatísticas das variáveis (Modelo Lasso)

Variável	Estimativa (β)	Erro Padrão	p-valor
Intercepto	-1.35147	0.18662	<0.001
map	0.21566	0.18280	0.2381
cd4	0.12348	0.17756	0.4868
egfr10ct	0.56459	0.26535	0.0334
scr10ct	0.51656	0.33474	0.1228
vl1	-0.38850	0.38534	0.3134
chol1	-0.22854	0.18499	0.2167
uacr1	-0.02425	0.19963	0.9033
time	-0.21638	0.17316	0.2114
lnuacr1	0.15359	0.21803	0.4811
lnvl1	-0.01174	0.21512	0.9565
glu1	-0.02616	0.17587	0.8818

A variável “egfr10ct” na tabela 7, apresenta coeficiente positivo e significativo, sugerindo que maiores valores dessa variável estão associados a um aumento na chance de nefrotoxicidade, enquanto as outras, não existe tanta significância nelas.

Avaliando as OR's e seus intervalos:

Tabela 8 – Estatísticas das variáveis (OR do Modelo Lasso)

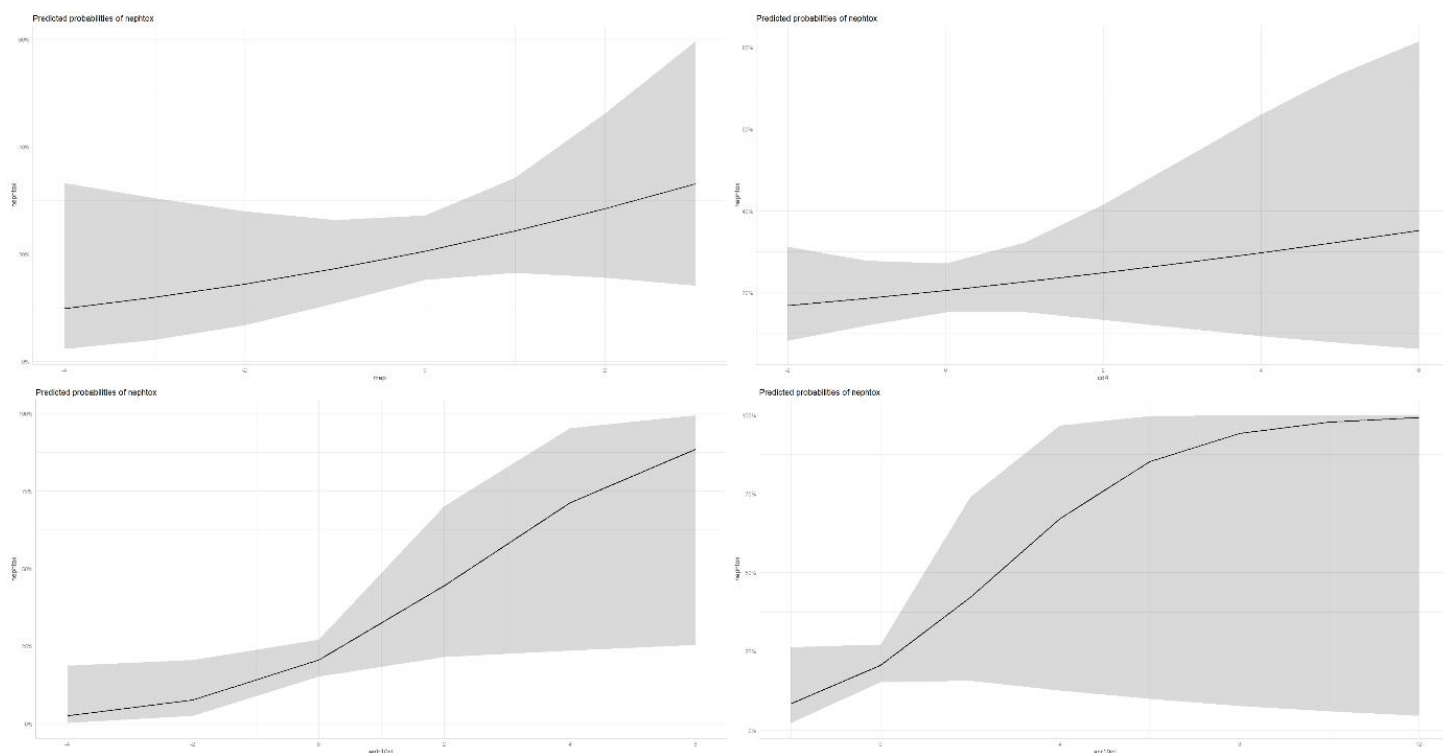
Variável	OR	IC 95% (2.5% – 97.5%)	Interpretação
Intercepto	0.26	0.18 – 0.37	-
map	1.24	0.87 – 1.78	Não significativo.
cd4	1.13	0.77 – 1.59	Não significativo.
egfr10ct	1.76	1.10 – 3.30	Significativo; aumenta em ~76% as chances do evento.
scr10ct	1.68	1.05 – 3.98	Significativo; aumenta em ~68% as chances do evento.
vl1	0.68	0.23 – 1.14	Não significativo.
chol1	0.80	0.55 – 1.14	Não significativo.
uacr1	0.98	0.61 – 1.43	Não significativo.
time	0.81	0.57 – 1.14	Não significativo.

Variável	OR	IC 95% (2.5% – 97.5%)	Interpretação
Inuacr1	1.17	0.75 – 1.78	Não significativo.
Invl1	0.99	0.66 – 1.54	Não significativo.
glu1	0.97	0.69 – 1.38	Não significativo.

Nas variáveis da Tabela 8, “egfr10ct” e “scr10ct” foram estatisticamente significativas. Isso significa que, para cada aumento unitário nessas variáveis, a chance de desenvolver nefrotoxicidade cresce em cerca de 76% e 68%, respectivamente.

Avaliando graficamente:

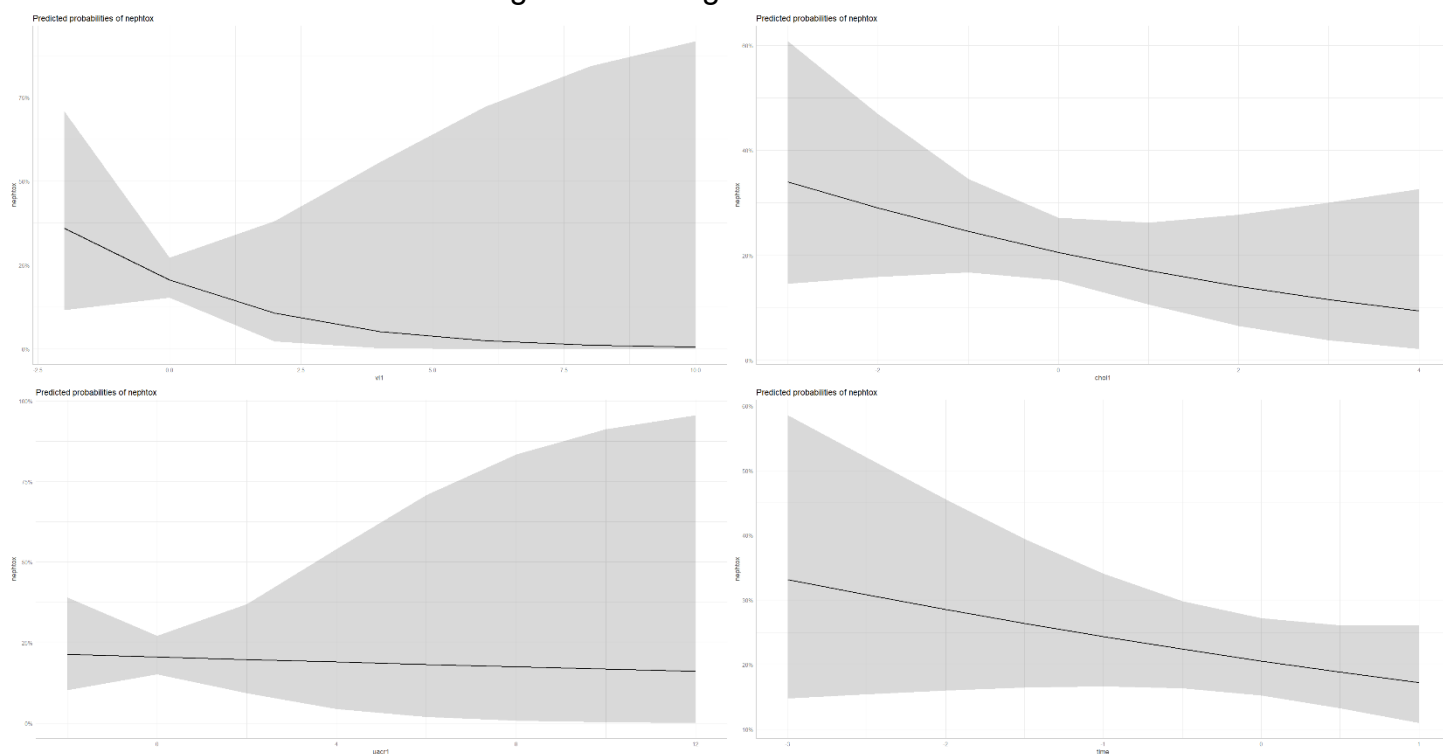
Figura 11 – Diagnóstico do Modelo Lasso



Na Figura 11, observa-se que tanto “map” quanto “cd4” apresentam tendência de aumento da probabilidade do desfecho à medida que seus valores crescem, porém esse efeito é sutil e acompanhado de intervalos de confiança amplos, o que sugere ausência de evidência estatística robusta.

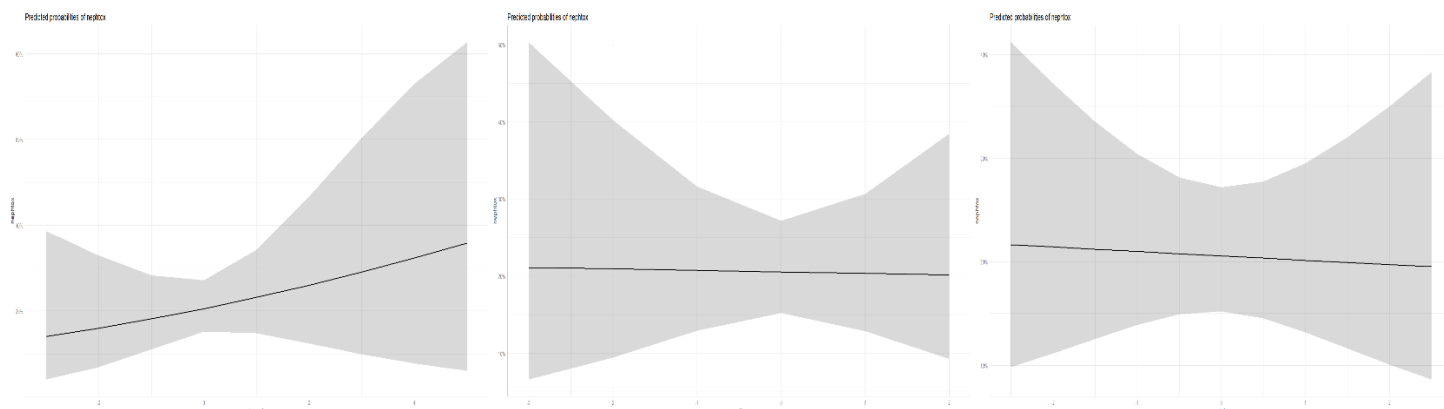
Por outro lado, as variáveis “egfr10ct” e “scr10ct” mostram um padrão mais consistente: há um crescimento acentuado e progressivo da probabilidade de ocorrência de nefrotoxicidade conforme seus valores aumentam, chegando a probabilidades próximas de 100% em níveis mais altos. Esse comportamento visual reforça os resultados da regressão logística, que identificaram essas duas variáveis como preditores significativos.

Figura 12 – Diagnóstico do Modelo Lasso



Na Figura 12, para “v1” e “chol”, há uma relação inversa, em que o aumento desses valores tende a reduzir a probabilidade de nefrotoxicidade, ainda que com intervalos de confiança largos em determinadas faixas. No caso de “uacr”, a probabilidade se mantém relativamente estável, indicando pouca influência sobre o desfecho. Já para “time”, nota-se uma tendência decrescente, sugerindo que, com o tempo, a chance de ocorrência de nefrotoxicidade tende a diminuir.

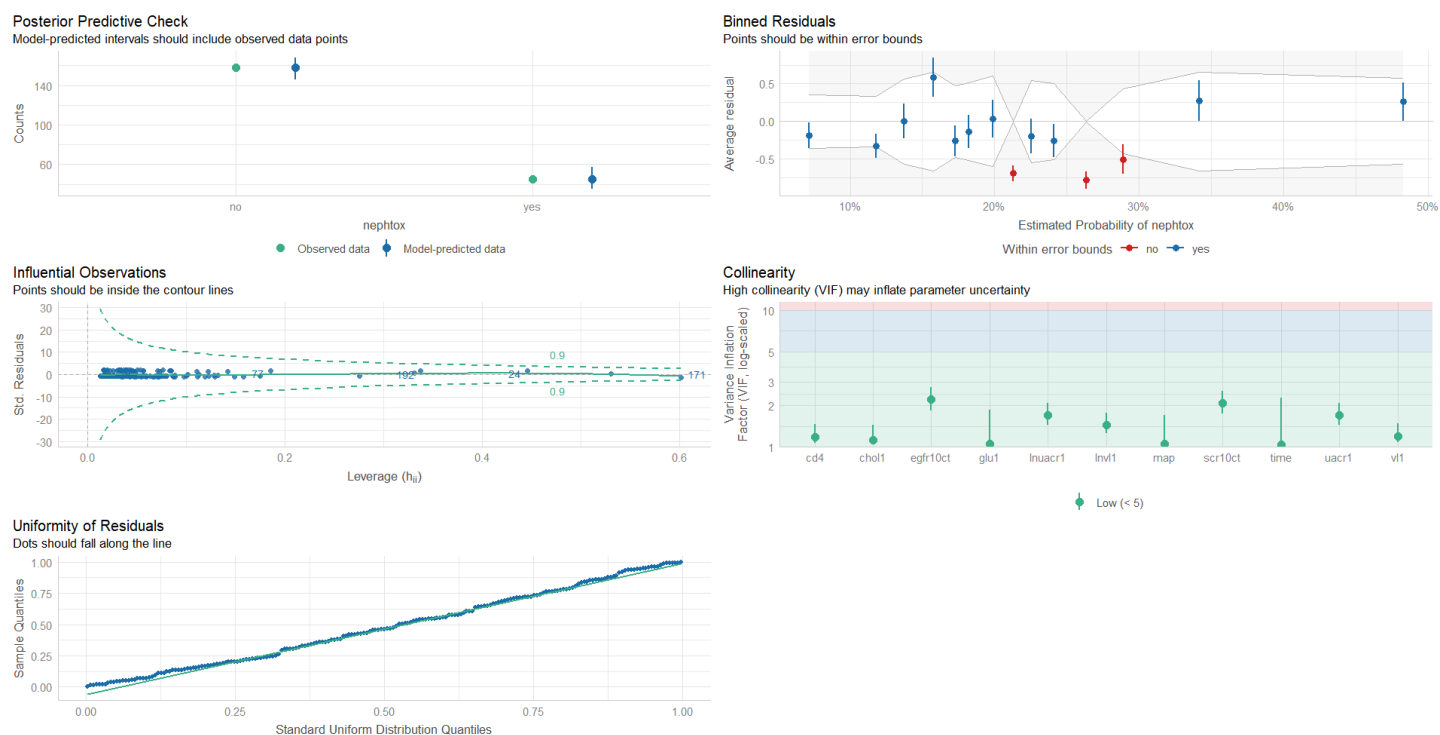
Figura 13 – Diagnóstico do Modelo Lasso



Na Figura 13, o “Inuacr1” apresenta maior influência, indicando aumento consistente da probabilidade de toxicidade renal conforme seus valores crescem, ainda que com intervalos de confiança relativamente amplos. Já para as variáveis “Invl1” e “glu1”, o efeito estimado é muito discreto, com curvas praticamente estáveis e intervalos de confiança largos, sugerindo pouca ou nenhuma associação significativa com a ocorrência de nefrotoxicidade.

Avaliando o diagnóstico do modelo:

Figura 14 – Diagnóstico do Modelo Lasso



Na Figura 14, em “Posterior Predictive Check” indica que os intervalos preditos pelo modelo incluem adequadamente os dados observados, sugerindo bom ajuste geral.

Os resíduos agrupados por faixa de probabilidade estimada (“Binned Residuals”) a maioria dos pontos permanece dentro dos limites de erro, embora se observe certa variabilidade em probabilidades mais altas, o que pode indicar leve tendência de falta de ajuste em regiões específicas.

No gráfico de observações influentes, quase todos os pontos encontram-se dentro das linhas de referência, com poucas observações de maior alavancagem, mas que não chegam a comprometer de forma crítica o modelo.

A análise de colinearidade, expressa pelos fatores de inflação da variância (VIF), mostra que todos os preditores apresentam valores abaixo de 5, o que indica baixa multicolinearidade e menor risco de inflação da variância dos parâmetros.

Por fim, o gráfico de uniformidade dos resíduos demonstra que os resíduos seguem de forma adequada a linha de referência. Em conjunto, esses resultados sugerem que o modelo apresenta bom desempenho, sem problemas importantes de ajuste, influência excessiva ou multicolinearidade entre as variáveis.

Vamos agora comparar o modelo Lasso com o modelo Step que fizemos anteriormente:

Figura 15 – Diagnóstico do Modelo Lasso

		Previsão das Classes		
		Positivo	Negativo	
Classes Atuais	Positivo	<i>True Positive</i> (TP) 24	<i>False Negative</i> (FN) 21	Sensibilidade TP $\frac{TP}{TP + FN}$ 0.5333
	Negativo	<i>False Positive</i> (FP) 77	<i>True Negative</i> (TN) 81	Especificidade TN $\frac{TN}{TN + FP}$ 0.5126
		Precisão TP $\frac{TP}{TP + FP}$ 0.2376	Previsão Negativa TN $\frac{TN}{TN + FN}$ 0.7941	Acurácia TP + TN $\frac{TP + TN}{TP + TN + FP + FN}$ 0.5172

Figura 16 – Diagnóstico do Modelo Step

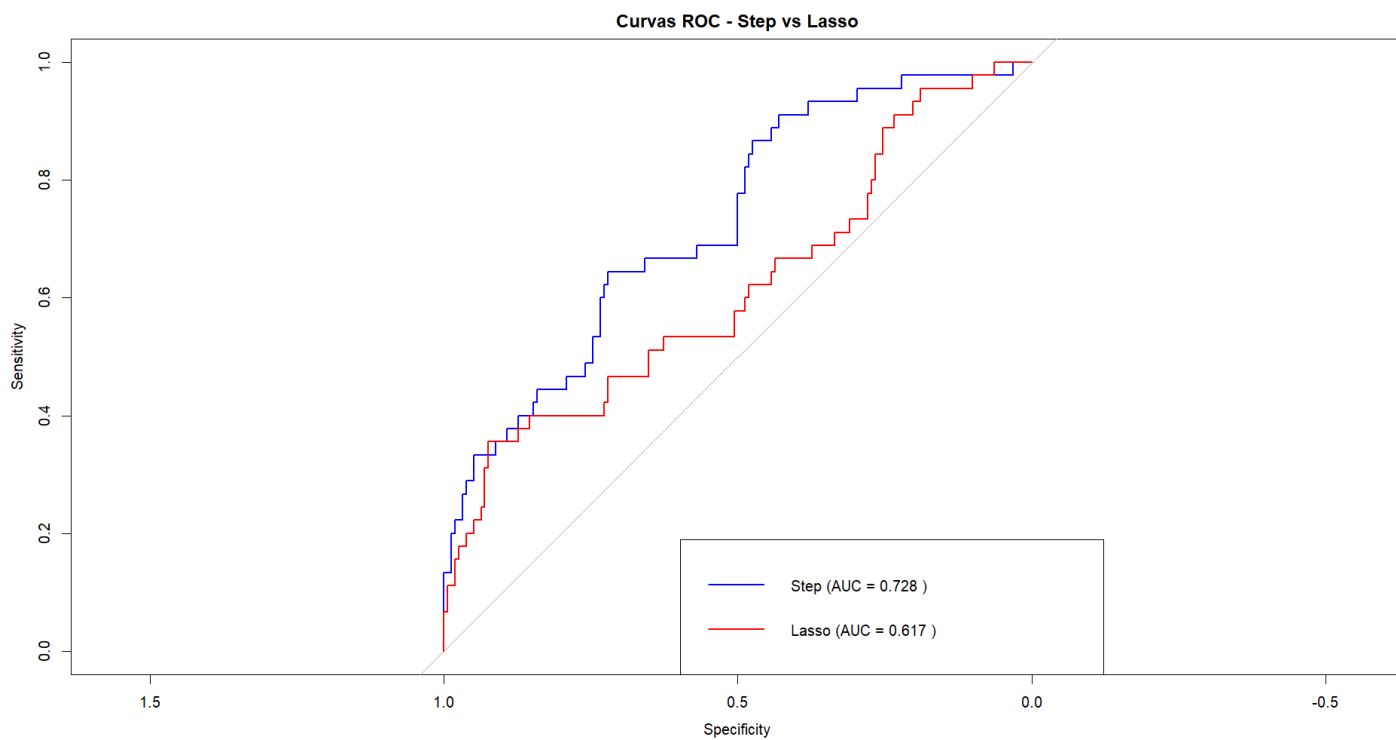
		Previsão das Classes		
		Positivo	Negativo	
Classes Atuais	Positivo	<i>True Positive</i> (TP) 31	<i>False Negative</i> (FN) 14	Sensibilidade TP $\frac{TP}{TP + FN}$ 0.6888
	Negativo	<i>False Positive</i> (FP) 74	<i>True Negative</i> (TN) 84	Especificidade TN $\frac{TN}{TN + FP}$ 0.5316
		Precisão TP $\frac{TP}{TP + FP}$ 0.2952	Previsão Negativa TN $\frac{TN}{TN + FN}$ 0.8571	Acurácia TP + TN $\frac{TP + TN}{TP + TN + FP + FN}$ 0.5665

O modelo Lasso apresentou acurácia de 51,7%, sensibilidade de 53,3% e especificidade de 51,2%. Esses valores indicam um desempenho apenas ligeiramente melhor que o acaso, com destaque para uma baixa precisão (23,7%), o que significa que grande parte das previsões positivas foram incorretas. Apesar disso, o valor da previsão negativa (79,4%) foi razoável, mostrando que o modelo se saiu melhor em identificar corretamente os casos negativos.

Já o modelo Step mostrou métricas mais equilibradas e superiores em quase todos os aspectos: acurácia de 56,6%, sensibilidade de 68,8% e especificidade de 53,1%. Isso indica que ele conseguiu identificar uma proporção maior de casos positivos corretamente sem perder muito em termos de especificidade. Além disso, sua precisão, 29,5%, também foi superior à do modelo Lasso, e a previsão negativa atingiu 85,7%, evidenciando uma boa capacidade de descartar corretamente os casos negativos.

Pela curva ROC na Figura 17, também vemos esta diferença nos modelos, sendo que o Step desempenha melhor que o do Lasso:

Figura 17 – Curva ROC dos dois modelos



4 CONSIDERAÇÕES PARCIAIS

A análise desenvolvida ao longo deste relatório permitiu compreender de forma abrangente os fatores associados à ocorrência de nefrotoxicidade, bem como avaliar diferentes estratégias de modelagem estatística aplicadas ao conjunto de dados. Inicialmente, a caracterização descritiva dos dados forneceu uma visão clara do perfil da amostra, evidenciando variáveis clínicas e laboratoriais que poderiam desempenhar papel relevante na predição do desfecho. Em seguida, foram aplicados modelos de regressão logística tradicionais e penalizados (Lasso), além de abordagens de seleção de variáveis via Stepwise, com o objetivo de comparar o desempenho preditivo entre eles.

Os diagnósticos dos modelos revelaram pontos importantes. As análises de resíduos e de colinearidade indicaram adequação geral, ainda que alguns preditores tenham mostrado baixa contribuição explicativa. O modelo Lasso, embora útil para reduzir dimensionalidade e selecionar variáveis, apresentou desempenho limitado, com acurácia em torno de 51% e baixa precisão, caracterizando elevado número de falsos positivos. Já o modelo Step mostrou resultados superiores, com acurácia de aproximadamente 61%, maior sensibilidade e melhor equilíbrio entre especificidade e valor preditivo negativo. Essa diferença sugere que a estratégia de seleção de variáveis empregada no Step foi mais eficaz para identificar padrões relevantes no conjunto de dados em comparação ao Lasso.

De forma geral, a avaliação comparativa aponta que, embora nenhum dos modelos tenha alcançado desempenho preditivo considerado elevado, o modelo Step se mostrou mais robusto e adequado à realidade do estudo, sobretudo pela maior capacidade de identificar corretamente os casos positivos de nefrotoxicidade.

REFERÊNCIAS

AGÊNCIA BRASIL. HIV: Brasil cumpre meta de pessoas em tratamento antirretroviral. *Agência Brasil*, 13 nov. 2023. Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia/2023-11/hiv-brasil-cumpr-met-de-pessoas-em-tratamento-antirretroviral>. Acesso em: 1 jun. 2025.

BRASIL. Ministério da Saúde. Saúde divulga dados epidemiológicos sobre HIV e aids no Brasil. *gov.br*, dez. 2024. Disponível em: <https://www.gov.br/aids/pt-br/assuntos/noticias/2024/dezembro/saude-divulga-dados-epidemiologicos-sobre-hiv-e-aids-no-brasil>. Acesso em: 1 jun. 2025.

BRASIL. Ministério da Saúde. Aids: etiologia, clínica, diagnóstico e tratamento. Brasília: Ministério da Saúde, 2008. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/Aids_etiologia_clinica_diagnostico_tratamento.pdf. Acesso em: 1 jun. 2025.

GIRO, Mario. O desafio da Aids não está vencido na África: a resistência está aumentando. *Instituto Humanitas Unisinos (IHU)*, 2023. Disponível em: <https://www.ihu.unisinos.br/categorias/649953-o-desafio-da-aids-nao-esta-vencido-na-africa-a-resistencia-esta-aumentando-artigo-de-mario-giro>. Acesso em: 1 jun. 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009. Disponível em: <https://hastie.su.domains/ElemStatLearn/>. Acesso em: 5 jun. 2025.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. *Applied Logistic Regression*. 3. ed. Hoboken: John Wiley & Sons, 2013. Disponível em: https://books.google.com/books/about/Applied_Logistic_Regression.html?id=Po0RLQ7USIMC. Acesso em: 5 jun. 2025.

MELLO, Luciana Magalhães de; RODRIGUES, Larissa Soares; LIMA, Yuri dos Santos; et al. Tendência temporal da taxa de mortalidade por HIV/aids no Brasil, 2000–2019: estudo ecológico. *Epidemiologia e Serviços de Saúde*, Brasília, v. 30, n. 5, e2020767, 2021. Disponível em: http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742021000500013. Acesso em: 1 jun. 2025.

SANTOS, Marcos André dos; et al. Trends in HIV/AIDS in Brazil: a nationwide analysis of the 2007–2021 period. *The Lancet Regional Health – Americas*, [s. l.], v. 29, 2023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S221339842300249X>. Acesso em: 1 jun. 2025.