Qual a Relação de Hábitos de Vida e Fatores Socioeconômicos com o Diagnóstico de Câncer de Próstata no Brasil?

https://doi.org/10.32635/2176-9745.RBC.2024v70n2.4633

What is the Relation of Lifestyle Habits and Socioeconomic Factors with Prostate Cancer Diagnosis in Brazil?

¿Cuál es la Relación de los Hábitos de Vida y los Factores Socioeconómicos con el Diagnóstico de Cáncer de Próstata en el Brasil?

Marco Antonio de Souza[1]; Camila Nascimento Monteiro[2]; Cláudia Renata dos Santos Barros[3]

RESUMO

Introdução: O câncer de próstata é o segundo mais comum entre os homens no Brasil, atrás apenas do câncer de pele não melanoma. Atualmente, há interesse em analisar dados referentes ao câncer com métodos do tipo *machine learning*. **Objetivo:** Investigar as características físicas, socioeconômicas e de hábitos de vida que podem estar associadas ao diagnóstico de câncer de próstata no Brasil. **Método:** Uma base de microdados referente à Pesquisa Nacional de Saúde 2019 foi utilizada, com a seleção de 42.799 indivíduos do sexo masculino; esse grupo foi analisado por meio de métodos estatísticos e modelagem por *machine learning* (regressão logística e árvore de decisão). **Resultados:** Os modelos aplicados permitiram identificar com bom nível de acurácia (próximo ou acima de 80%) os indivíduos que receberam o diagnóstico de câncer de próstata (DCP), além de grupos com características específicas mais fortemente associadas a essa doença. Entre as variáveis mais significativamente ligadas à taxa de DCP, destacam-se: idade, diagnóstico de alto nível de colesterol, se possui plano de saúde e nível de instrução. **Conclusão:** Os modelos indicam um nível de associação significativo de fatores socioeconômicos, físicos e alimentares com a frequência de DCP no grupo analisado. O alto nível de acurácia e a sensibilidade dos modelos demonstram o potencial dos métodos de *machine learning* para a previsão de DCP.

Palavras-chave: Neoplasias da Próstata; Estilo de vida/etnologia; Estudos Transversais; Aprendizado de Máquina.

ABSTRACT

Introduction: Prostate cancer is the second most common cancer among men in Brazil, behind only non-melanoma skin cancer. Currently, there is interest in analyzing data related to cancer with machine learning type methods. **Objective:** The investigation of physical, lifestyle and socioeconomic features that may be associated with prostate cancer diagnosis in Brazil. **Method:** A microdata base referring to the 2019 National Health Survey in Brazil was utilized, and 42,799 male individuals were selected; this group was analyzed through statistical methods and machine learning modeling (logistic regression and decision tree). **Results:** The models applied allowed to identify with a good level of accuracy (near or above 80%) individuals with diagnosis of prostate cancer (DPC), in addition to groups with specific features more strongly associated with this disease. Among the variables more significantly associated with DPC rate, the following stand out: age, diagnosis of high level of cholesterol, health insurance, and level of education. **Conclusion:** The models indicate a significant level of association of socioeconomic, physical, and dietary factors with the frequency of DPC in the group analyzed. The high level of accuracy and sensitivity of the models demonstrates the potential of machine learning methods for predicting DPC.

Key words: Prostatic Neoplasms; Life Style/ethnology; Cross-Sectional Studies; Machine Learning.

RESUMEN

Introducción: El cáncer de próstata es el segundo cáncer más común entre los hombres en el Brasil, sólo detrás del cáncer de piel no melanoma. Actualmente existe interés en analizar datos relacionados con el cáncer con métodos de tipo *machine learning*. **Objetivo:** Investigar características físicas, de estilo de vida y socioeconómicas que pueden estar asociadas con el diagnóstico de cáncer de próstata en el Brasil. **Método:** Se utilizó una base de microdatos referente a la Encuesta Nacional de Salud de 2019, con la selección de 42 799 individuos de sexo masculino; este grupo fue analizado mediante métodos estadísticos y modelado de *machine learning* (regresión logística y árbol de decisión). **Resultados:** Los modelos aplicados permitieron identificar con buen nivel de exactitud (cerca o por encima del 80%) a los individuos con diagnóstico de cáncer de próstata (DCP), además de grupos con características específicas más fuertemente asociadas a esta enfermedad. Entre las variables más significativamente asociadas a la tasa de DCP destacan las siguientes: la edad, el diagnóstico de nivel alto de colesterol, si se tiene seguro médico y el nivel de educación. **Conclusión:** Los modelos indican un nivel significativo de asociación de factores socioeconómicos, físicos y dietéticos con la frecuencia de DCP en el grupo analizado. El alto nivel de exactitud y sensibilidad de los modelos demuestra el potencial de los métodos de *machine learning* para predecir la DCP.

Palabras clave: Neoplasias de la Próstata; Estilo de Vida/etnología; Estudios Transversales; Aprendizaje Automático.

[1]Universidade de São Paulo, Instituto de Física. São Paulo (SP), Brasil. E-mail: marsouza@if.usp.br. Orcid iD: https://orcid.org/0000-0003-3340-5912
[2]Hospital Sírio-Libanês, Saúde Populacional. São Paulo (SP), Brasil. E-mail: c.nascimentomonteiro@gmail.com. Orcid iD: https://orcid.org/0000-0002-0121-0398
[3]Instituto Butantan. São Paulo (SP), Brasil. E-mail: barros.crs3@gmail.com. Orcid iD: https://orcid.org/0000-0002-1582-2010
Endereço para correspondência: Marco Antonio de Souza. Rua Manuel Jacinto, 667 – bloco 9, apto. 73 – Vila Morse. São Paulo (SP), Brasil. CEP 05624-001
E-mail: marsouza@if.usp.br

INTRODUÇÃO

No Brasil, o câncer de próstata é o segundo mais comum entre os homens[1]. Em países como Estados Unidos, há uma estimativa de um em cada oito homens com câncer de próstata no decorrer da vida[2]. Esse tipo de câncer é multicausal e os principais fatores de risco identificados são idade, cor/etnia, nacionalidade, histórico familiar e alterações genéticas. Há também outros fatores associados ao hábito de vida relacionados ao câncer de próstata que têm sido estudados para melhor conhecimento de possível relação causal, entre eles: dieta, obesidade, tabagismo, exposição ocupacional, inflamação da próstata e doenças sexualmente transmissíveis[3]. Além disso, fatores sociais e econômicos têm uma forte influência sobre os hábitos de vida da população e as suas condições de acesso aos serviços de saúde, podendo, em tese, influenciar a ocorrência e o diagnóstico de câncer de próstata (DCP) na população masculina.

Estudos anteriores já foram publicados sobre o câncer de próstata no Brasil nos aspectos de saúde pública e epidemiologia. Por exemplo, há estudos de revisão da literatura sobre o tema com uma contextualização para a saúde pública no Brasil[4,5], outros que caracterizam os indivíduos com a doença ou o estadiamento da doença por meio de variáveis clínicas e sociodemográficas[6,7], discussões sobre a forma de rastreamento populacional do câncer de próstata[8], e estudos sobre a tendência temporal da mortalidade por câncer de próstata no Brasil ou regiões específicas do país[9-11].

Nos últimos anos, há um interesse crescente pelo uso de métodos de *machine learning* ou inteligência artificial (IA) para a pesquisa e prevenção do câncer[12,13], por exemplo, pela análise de diversas variáveis que podem influenciar a taxa de ocorrência de diferentes tipos de câncer, e por análise de imagens médicas. Nesse contexto, o presente estudo tem como objetivos: analisar os fatores associados ao DCP no Brasil por meio dos dados da Pesquisa Nacional de Saúde (PNS) 2019; e testar modelos preditivos de *machine learning* sobre o câncer de próstata com o uso desses dados.

MÉTODO

Trata-se de um estudo transversal com dados secundários provenientes da PNS realizada em 2019. A base de dados contém 42.799 indivíduos; destes, 339 (0,79%) disseram "ter recebido" DCP em algum momento da sua vida e 42.460 (99,21%) "não ter recebido".

Entre as 58 variáveis selecionadas inicialmente, estava a variável dependente "câncer de próstata": ter recebido o diagnóstico ou não ter recebido (sim ou não); e 57 eram variáveis independentes.

Após a aplicação de critérios de eliminação de variáveis que serão explicados nessa seção, restaram 13 variáveis independentes, as quais foram aplicadas nos modelos preditivos. São elas: idade; há quantos anos a pessoa consultou um médico pela última vez; como a pessoa avalia a sua própria saúde (muito boa, boa, ruim etc.); quantos dias da semana consome frutas; quantos dias da semana consome verduras e/ou legumes; quantos dias da semana consome sucos artificiais; se já recebeu diagnóstico de nível alto de colesterol (sim ou não); se manuseia ou manuseava substâncias químicas no trabalho que são potencialmente prejudiciais à saúde (sim ou não); se possui plano de saúde (sim ou não); cor/etnia (branca, preta, parda etc.); nível de instrução (fundamental completo, médio completo, superior incompleto etc.); se fuma algum produto do tabaco, e com qual frequência (não fuma, diariamente, menos que diariamente); e se já recebeu diagnóstico de depressão (sim ou não).

A base original foi obtida no site do Instituto Brasileiro de Geografia e Estatística (IBGE), na seção da PNS, subseção de microdados[14]. Essa pesquisa, inquérito domiciliar de saúde, foi realizada em 2019, com amostra representativa da população brasileira. A base original, contida no arquivo PNS_2019.txt, possui 346 Mb, um total de 279.382 linhas (indivíduos entrevistados) e 817 colunas (características).

A PNS teve aprovação da Comissão Nacional de Ética em Pesquisa (Conep) em agosto de 2019 sob o número n.º 3.529.376 (CAAE: 11713319.7.0000.0008) para a edição de 2019. Como esta pesquisa utilizou dados disponibilizados publicamente, justifica-se que, para o presente estudo, não há necessidade de análise pelo Comitê de Ética em Pesquisa (CEP), de acordo com a Resolução n.º 510/2016[15] do Conselho Nacional de Saúde (CNS).

No primeiro processo de filtragem, feito pela biblioteca PNS-IBGE[14] do R[16], foi possível extrair um *dataframe* da base original com as variáveis de interesse. Foi, então, extraído um arquivo csv de 279.382 linhas (indivíduos entrevistados) e 68 colunas (características). No segundo processo de filtragem, a base de dados foi reduzida para conter apenas os indivíduos do sexo masculino que responderam duas perguntas que compõem a variável dependente: a) se o indivíduo recebeu diagnóstico de algum tipo de câncer na vida (1 = sim, 2 = não); b) se o indivíduo recebeu DCP ao longo da vida (1 = sim, 2 = não). Após o 2º processo de filtragem, foi obtido um *dataframe* de 42.799 linhas (indivíduos) e 58 colunas (características). A base de dados então foi devidamente preparada para ser usada nas modelagens, incluindo o tratamento de dados faltantes (*missings*) e a organização de variáveis categóricas, resultando em uma base de dados final com 42.799 linhas (indivíduos) e 58 variáveis.

Com a base de dados devidamente preparada, foram realizados os procedimentos de filtragem das variáveis de maior relevância, análise estatística e modelagem por métodos de *machine learning*. No caso deste trabalho, os modelos de classificação escolhidos para a descrição da variável dependente "DCP" foram: regressão logística e árvore de decisão. Tais procedimentos são detalhados a seguir.

Antes da aplicação da base de dados nos modelos de regressão logística e árvore de decisão, foi feita uma filtragem inicial das variáveis descritivas via *information value* (IV), calculado pelo RStudio[17]. Todas as variáveis com IV considerado fraquíssimo (≤ 0,02) foram excluídas dos modelos previamente, restando 42 variáveis descritivas nessa etapa.

Após filtragem pelo IV, foi aplicado o modelo de regressão logística com o uso do RStudio[17], considerando como variável dependente o DCP (variável binária, com **1** para "sim", e **0** para "não"). Para isso, a base de dados foi dividida em base de treino (70% dos indivíduos) e base de teste (30%). A divisão da base de dados em treino e teste é importante para realizar a validação cruzada do modelo. Portanto, o modelo é todo desenvolvido (ou parametrizado) com a base de treino, e, em seguida, validado na base de teste mediante métricas como a acurácia, sensibilidade e especificidade. O modelo de regressão logística permite prever se a variável dependente terá resultado positivo ou negativo para um certo indivíduo da base.

Usando a base de treino, foi aplicado o método *backward* e um nível de significância de 0,10 (ou seja, $p \leq 0,10$) para a seleção das variáveis. Após esse processo, chegou-se a um modelo final com um total de 13 variáveis descritivas, todas categóricas, as quais foram citadas antecipadamente no método.

O modelo de árvore de decisão foi executado no RStudio[17], com as mesmas 13 variáveis descritivas, considerando árvores de 2 e 3 níveis. A árvore de decisão também é utilizada como um modelo preditivo para a variável dependente DCP, e, além disso, permite identificar grupos específicos com maior frequência de casos positivos para a variável dependente, de acordo com o nível de associação com as variáveis descritivas.

RESULTADOS

Entre as variáveis descritivas selecionadas, a que apresentou o valor mais alto de IV foi a variável "idade" (IV = 2,86), o que representa uma capacidade preditiva muito forte em relação ao câncer de próstata. Verifica-se que a idade média dos homens entrevistados é de 45,9 anos e a mediana é de 45 anos; isso mostra que o grupo de 50% dos indivíduos acima da mediana contém a faixa etária de maior propensão a receber DCP (Figura 1). Observando o Gráfico 1(c), nota-se que a mediana de idade dos homens que responderam "sim" para DCP (72 anos) está bem acima da mediana dos homens que responderam "não" (44 anos).

A Tabela 1 mostra a distribuição dos indivíduos nas categorias das variáveis descritivas e o cruzamento com a variável dependente DCP; os resultados que podem ser contextualizados mais claramente são:

Idade: aumento da frequência de DCP a partir de 50 anos de idade, com frequência mais alta na categoria "≥ 80 anos".

Consultas médicas: maior frequência de DCP nos indivíduos que tiveram sua última consulta médica mais próxima do momento da pesquisa (até dois anos antes).

Autoavaliação de saúde: maior frequência de DCP nos indivíduos que afirmaram ter uma qualidade de saúde "ruim ou muito ruim".

Diagnóstico de colesterol alto: maior frequência de DCP nos indivíduos que receberam diagnóstico de nível alto de colesterol.

Plano de saúde: maior frequência de DCP nos indivíduos que afirmam ter plano de saúde.

Diagnóstico de depressão: maior frequência de DCP nos indivíduos que receberam diagnóstico de depressão.

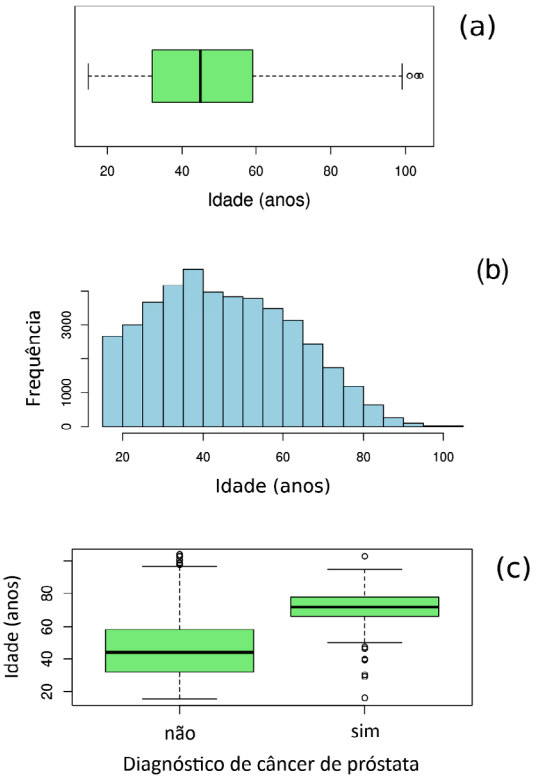

Figura 1. Descrição da variável "idade" dos indivíduos do sexo masculino selecionados na base da Pesquisa Nacional de Saúde. **(a)** *Boxplot* mostrando a distribuição de idade. **(b)** Histograma mostrando a distribuição de idade. **(c)** Comparação em *boxplot* das distribuições de idade nos grupos: com diagnóstico de câncer de próstata em algum momento da vida (sim); sem diagnóstico de câncer de próstata (não)

Tabela 1. Variáveis descritivas usadas nos modelos finais de regressão logística e árvore de decisão. São mostradas as categorias de cada variável, a frequência de indivíduos em cada categoria, e a frequência de casos positivos (sim) e negativos (não) de câncer de próstata em cada categoria. N.A.: pergunta não aplicada

Variável descritiva			Câncer de próstata (%)	
Idade	**Faixa de idade (anos)**	**Frequência (%)**	**Não**	**Sim**
	< 35	29,46	99,98	0,02
	≥ 35 e < 50	29,45	99,96	0,04
	≥ 50 e < 65	24,76	99,35	0,65
	≥ 65 e < 80	13,50	96,63	3,37
	≥ 80	2,83	94,48	5,52
Consultas médicas	**Tempo desde a última consulta médica (anos)**	**Frequência (%)**	**Não**	**Sim**
	até 2 anos	83,02	99,05	0,95
	mais de 2 anos	15,77	99,96	0,04
	nunca foi	1,21	100,00	0,00
Autoavaliação de saúde	**Autoavaliação**	**Frequência (%)**	**Não**	**Sim**
	muito boa ou boa	66,77	99,54	0,46
	regular	27,82	98,58	1,42
	ruim ou muito ruim	5,41	98,36	1,64
Consumo de frutas	**Consumo semanal**	**Frequência (%)**	**Não**	**Sim**
	1 a 3 dias	40,45	99,53	0,47
	4 a 6 dias	20,61	99,29	0,71
	nunca ou muito pouco	13,10	99,48	0,52
	todos os dias	25,84	98,50	1,50
Consumo de suco artificial [a]	**Consumo semanal**	**Frequência (%)**	**Não**	**Sim**
	1 a 3 dias	21,37	99,68	0,32
	4 a 6 dias	7,89	99,62	0,38
	nunca ou muito pouco	64,08	99,00	1,00
	todos os dias	6,66	99,16	0,84
Diagnóstico de colesterol alto	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não ou N.A.	89,12	99,35	0,65
	sim	10,88	98,02	1,98
Exposição a produtos químicos no trabalho [b]	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não ou N.A.	87,21	99,12	0,88
	sim	12,79	99,82	0,18
Possui plano de saúde	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não	78,82	99,38	0,62
	sim	21,18	98,58	1,42
Cor/etnia	**Cor/etnia**	**Frequência (%)**	**Não**	**Sim**
	amarela	0,73	98,72	1,28
	branca	36,14	98,89	1,11
	ignorado	0,02	100,00	0,00
	indígena	0,78	100,00	0,00
	parda	50,51	99,45	0,55
	preta	11,82	99,11	0,89

continua

Tabela 1. continuação

Variável descritiva			Câncer de próstata (%)	
Nível de instrução	**Nível de instrução**	**Frequência (%)**	**Não**	**Sim**
	sem instrução ou fundamental incompleto	42,67	99,03	0,97
	fundamental completo ou médio incompleto	15,48	99,41	0,59
	médio completo ou superior incompleto	28,93	99,43	0,57
	superior completo	12,92	99,06	0,94
Consumo de verduras e legumes	**Consumo semanal**	**Frequência (%)**	**Não**	**Sim**
	1 a 3 dias	35,67	99,34	0,66
	4 a 6 dias	21,24	99,21	0,79
	nunca ou muito pouco	9,54	99,46	0,54
	todos os dias	33,55	98,99	1,01
Fuma atualmente	**Uso semanal**	**Frequência (%)**	**Não**	**Sim**
	diariamente	14,32	99,45	0,55
	menos que diário	1,75	99,73	0,27
	não fuma	83,93	99,16	0,84
Teve diagnóstico de depressão	**Resposta**	**Frequência (%)**	**Não**	**Sim**
	não	95,43	99,24	0,76
	sim	4,57	98,46	1,54

[a] De acordo com a Pesquisa Nacional de Saúde (PNS), essa variável refere-se ao consumo dos chamados sucos de "caixinha", em lata ou refresco em pó.
[b] De acordo com a PNS, essa variável refere-se ao manuseio de produtos químicos como: agrotóxicos, gasolina, diesel, formol, chumbo, mercúrio, cromo, quimioterápicos etc.

A seleção de variáveis descritivas como "plano de saúde" e "nível de instrução" indica uma influência da situação socioeconômica nos modelos preditivos. De fato, tal indicação pode ser verificada pelo cruzamento da variável dependente DCP com a renda familiar *per capita* dos indivíduos. Considerando a faixa de idade a partir de 50 anos, observam-se taxas de DCP de 0,84%, 1,82%, 2,49% e 3,07% para as respectivas faixas de renda familiar *per capita* (em salários-mínimos – SM): até ½ SM; mais de ½ SM até 2 SM; mais de 2 SM até 5 SM; mais de 5 SM. Isto é, a faixa de renda mais alta (mais de 5 SM) tem uma taxa de DCP ≈ 3,7 × maior do que na faixa de renda mais baixa (até ½ SM).

Com a premissa descrita acima, o possível efeito socioeconômico sobre as variáveis descritivas foi investigado; para isso, foi determinado o nível de associação entre cada variável descritiva e a renda familiar *per capita* pelas medidas V de Cramer e ω de Cohen. Verificou-se um nível alto de associação da renda familiar *per capita* com as variáveis "plano de saúde" (V de Cramer = 0,487 e ω de Cohen = 0,487) e "nível de instrução" (V de Cramer = 0,314 e ω de Cohen = 0,544). Um nível médio de associação com a renda familiar per capita foi observado na variável "cor/etnia" (V de Cramer = 0,150 e ω de Cohen = 0,260), e um nível razoável foi observado em "consumo de frutas" e "consumo de verduras e legumes" (ω de Cohen = 0,218 e 0,229, respectivamente). Portanto, a influência da renda familiar nas variáveis descritivas deve ser levada em conta na interpretação dos resultados da Tabela 1.

Os resultados de acurácia, sensibilidade, especificidade e ROC-AUC para a regressão logística são mostrados na Tabela 2**,** referentes às bases de treino e teste. Observa-se que a acurácia, sensibilidade e especificidade para essas bases são muito satisfatórias, pois todas estão acima de 80%. Na base de teste, houve um pequeno aumento na acurácia e especificidade, e uma pequena diminuição na

Tabela 2. Acurácia, sensibilidade, especificidade e ROC-AUC obtidos nos modelos de regressão logística, árvore de decisão de 2 níveis e árvore de decisão de 3 níveis, considerando as bases de treino (70% do total) e teste (30% do total)

Modelo	Base	Acurácia	Sensibilidade	Especificidade	ROC-AUC
Regressão logística	treino	0,828	0,828	0,828	0,822
	teste	0,835	0,802	0,835	0,780
Árvore de decisão (2 níveis)	treino	0,801	0,850	0,800	0,823
	teste	0,805	0,821	0,804	0,812
Árvore de decisão (3 níveis)	treino	0,767	0,906	0,766	0,832
	teste	0,773	0,887	0,772	0,813

sensibilidade. O resultado obtido para o ROC-AUC na base de treino (0,822) também pode ser considerado muito satisfatório, enquanto o resultado na base de teste (0,780) encontra-se muito próximo da mesma condição.

Os resultados obtidos para as árvores de decisão de 2 e 3 níveis são mostrados graficamente nas Figuras 2 e 3, respectivamente, referentes à base de treino. Observando as frequências de casos de câncer de próstata nos nós finais da árvore de **2 níveis**, verifica-se:

- Os **nós 3 e 4** sugerem que há uma probabilidade maior de ocorrência de câncer de próstata em homens negros comparativamente ao conjunto das outras etnias, considerando a faixa etária abaixo de 50 anos;
- Os **nós 6 e 7** indicam que o grupo de homens com nível de colesterol alto tem uma maior frequência de DCP (≈ 3× maior) em comparação com o grupo de nível de colesterol normal ou desconhecido, considerando a faixa etária de 50 a 65 anos;
- Os **nós 9 e 10** indicam que o grupo de homens com nível de instrução entre o ensino fundamental completo e superior completo tem uma maior frequência de DCP (≈ 2× maior) em comparação com o grupo de menor nível de instrução, considerando a faixa etária a partir de 65 anos e abaixo de 80 anos;
- Os **nós 12 e 13** indicam que o grupo de homens que possuem plano de saúde tem uma maior frequência de DCP (≈ 2× maior) em comparação com o grupo que não possui plano de saúde, considerando a faixa etária a partir de 80 anos.

Com relação ao nó 5 (o qual resulta nos nós finais 6 e 7), pode-se avaliar o nível de associação entre o fator de risco "colesterol alto" e o desfecho DCP para o subgrupo de homens com idade a partir de 50 anos e menor que 65 anos, utilizando o *odds ratio* (OR). Nesse caso, foi obtido OR = 2,87 (intervalo de confiança – IC 95%: 1,58 – 5,21), representando um nível de associação significativo.

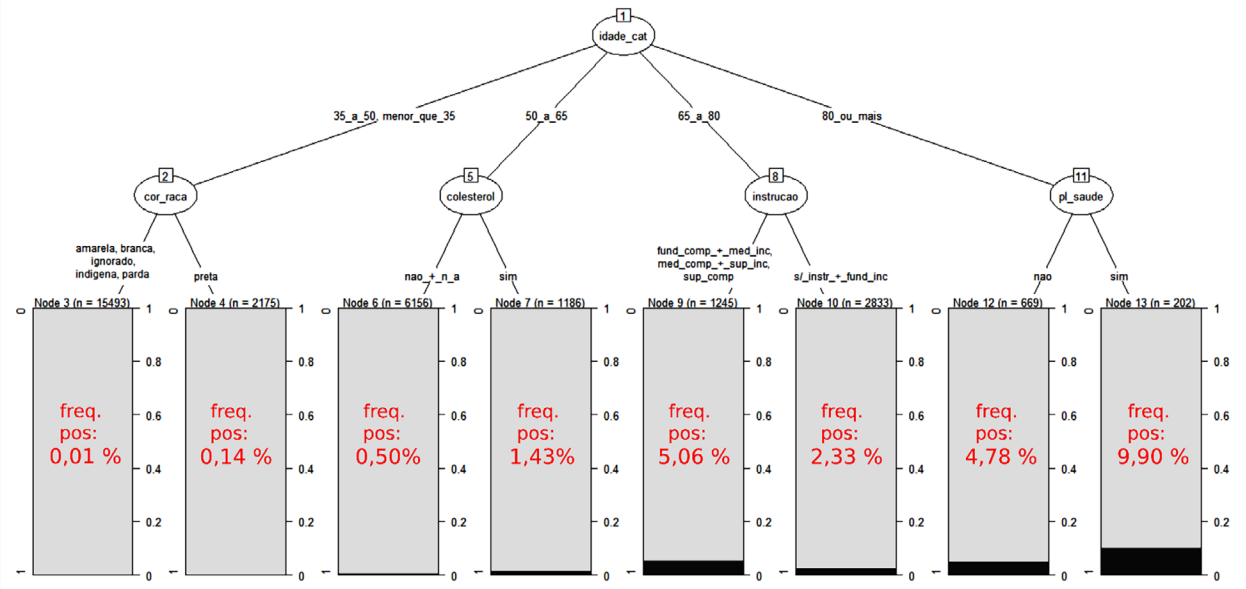

Figura 2. Árvore de decisão de 2 níveis, considerando a base de treino com 70% dos indivíduos selecionados. As porcentagens em vermelho indicam a frequência de casos positivos de câncer de próstata nos nós finais da árvore

Legendas: *idade_cat* = idade na forma categórica (anos); *cor_raca* = cor/etnia; *colesterol* = nível de colesterol alto; *instrucao* = nível de instrução; *pl_saude* = plano de saúde; *n_a* = pergunta não aplicada; *fund_comp* = nível fundamental completo; *med_inc* = nível médio incompleto; *med_comp* = nível médio completo; *sup_inc* = nível superior incompleto; *sup_comp* = nível superior completo; *s/_instr* = sem instrução; e *fund_inc* = nível fundamental incompleto.

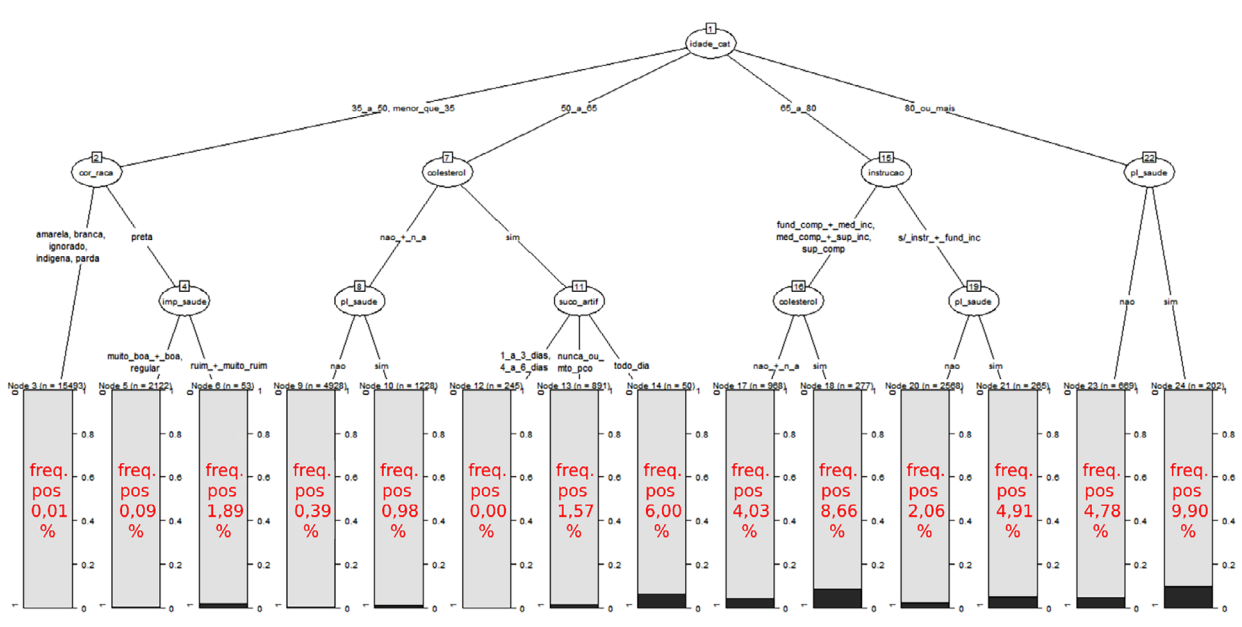

Figura 3. Arvore de decisão de 3 níveis, considerando a base de treino com 70% dos indivíduos selecionados. As porcentagens em vermelho indicam a frequência de casos positivos de câncer de próstata nos nós finais da árvore

Legendas: *idade_cat* = idade na forma categórica (anos); *cor_raca* = cor/etnia; *colesterol* = nível de colesterol alto; *instrucao* = nível de instrução; *pl_saude* = plano de saúde; *imp_saude* = impressão sobre a própria saúde; *suco_artif* = consumo semanal de suco artificial; *n_a* = pergunta não aplicada; *fund_comp* = nível fundamental completo; *med_inc* = nível médio incompleto; *med_comp* = nível médio completo; *sup_inc* = nível superior incompleto; *sup_comp* = nível superior completo; *s/_instr* = sem instrução; *fund_inc* = nível fundamental incompleto; e *mto_pco* = muito pouco

Ao interpretar os resultados dos nós 9 e 10, deve-se levar em conta a forte associação entre nível de instrução e renda familiar *per capita* discutida anteriormente[18]. A respeito dos nós 12 e 13, deve-se levar em conta as diferentes características dos sistemas público e privado de saúde no Brasil, além da forte associação entre a variável "plano de saúde" e a renda familiar *per capita* discutida anteriormente[19].

Os principais resultados da árvore de decisão de **3 níveis** são descritos a seguir:

- Entre os homens com idade a partir de 50 e < 65 anos, e que possuem nível de colesterol normal ou desconhecido, observa-se que a frequência de DCP é ≈ 2,5× maior no subgrupo que possui plano de saúde em comparação com o subgrupo que não possui plano de saúde;
- Entre os homens com idade a partir de 50 e < 65 anos, e que possuem nível de colesterol alto, observa-se que o subgrupo de homens que toma sucos artificiais todos os dias tem uma frequência de DCP ≈ 4× maior do que o subgrupo de homens que diz não tomar ou tomar muito pouco sucos artificiais.
- Entre os homens com idade a partir de 65 e < 80 anos, e que possuem pelo menos o nível fundamental completo de escolaridade, observa-se que o subgrupo de homens com nível de colesterol alto tem uma frequência de DCP ≈ 2× maior do que o subgrupo de homens que têm nível de colesterol normal ou desconhecido.
- Entre os homens com idade a partir de 65 e < 80 anos, e que possuem nível de escolaridade abaixo do fundamental completo, observa-se que o subgrupo de homens que possui plano de saúde tem uma frequência de DCP ≈ 2,5× maior do que o subgrupo de homens que não possui plano de saúde.

A árvore de decisão de 3 níveis mostra resultados que reforçam aspectos discutidos sobre a árvore de 2 níveis, contudo, em grupos mais específicos. Por exemplo, os pares de nós {9,10}, {20,21} e {23,24} reforçam a maior probabilidade de DCP para os indivíduos que afirmam ter plano de saúde. O par de nós {17,18} reforça a maior probabilidade de DCP para os indivíduos com alto nível de colesterol, contudo, na faixa etária a partir de 65 e < 80 anos e que possuem pelo menos o nível fundamental completo de escolaridade. Com relação ao nó 16 (o qual resulta nos nós finais 17 e 18), pode-se avaliar o nível de associação entre o fator de risco "colesterol alto" e o desfecho DCP para esse subgrupo. Nesse caso, foi obtido OR = 2,26 (IC 95%: 1,33 – 3,83), mostrando novamente um nível de associação significativo.

Uma informação relevante ocorre no trio de nós {12,13,14}, o qual indica uma probabilidade significativamente maior de DCP para os indivíduos que afirmam consumir sucos artificiais/industrializados todos os dias, dentro da faixa etária a partir de 50 e < 65 anos e que possuem alto nível de colesterol. Contudo, tal resultado deve ser analisado de forma cautelosa, como se discute na próxima seção.

DISCUSSÃO

A mediana de idade dos homens que responderam ter recebido DCP ("sim") foi 72 anos, valor bem acima da mediana dos homens que responderam "não" (44 anos), resultado que é compatível com a tendência geral na qual os homens desenvolvem o câncer de próstata a partir de 50 anos aproximadamente[3].

Com relação aos nós 3 e 4 da árvore de 2 níveis, sugere-se que as pessoas negras têm uma maior tendência para o desenvolvimento de câncer de próstata em idades abaixo de 50 anos; tal constatação parece corroborar estudos anteriores sobre o câncer de próstata[3], nos quais se verificou que essa doença é mais frequente em homens com ascendência africana e caribenha. Contudo, os casos de câncer de próstata abaixo de 50 anos são raros na base (2 casos entre 15.493 indivíduos no nó 3, e 3 casos entre 2.175 indivíduos no nó 4) e, por isso, sugere-se que conclusões a respeito dessa faixa etária devam ser corroboradas por uma base de dados mais numerosa abaixo de 50 anos.

A respeito dos nós 6 e 7 da árvore de 2 níveis, deve-se considerar estudos como o de Pelton et al.[20], o qual afirma que altos níveis de colesterol no sangue estão relacionados a casos de câncer de próstata mais agressivos, e de Jamnagerwalla et al.[21], o qual aponta que altos níveis de colesterol sérico total e HDL estão associados a um risco aumentado de câncer de próstata de alto grau. Uma comunicação do *Johns Hopkins Medicine*[22] descreve pesquisas mais recentes que apontam para conclusões semelhantes. Como o presente estudo é transversal, não se pode afirmar com segurança que ele corrobora os resultados dos trabalhos anteriores, mas indica-se a importância de estudos adicionais sobre a relação colesterol/DCP com a inclusão de indivíduos do Brasil.

O fato do presente estudo ser transversal limita a possibilidade de conclusões que implicam a relação causal pela falta de controle da temporalidade. Outra limitação refere-se ao uso de dados secundários, o que dificulta a precisão das respostas dos entrevistados quanto a algumas variáveis; um exemplo claro nesse aspecto é a frequência semanal de consumo de sucos artificiais. Contudo, há estudos anteriores que sugerem a relação entre o alto consumo de bebidas açucaradas e uma maior incidência de câncer de próstata, como o de Miles et al.[23] e o de Llaha et al.[24]. Além disso, o estudo de Makarem et al.[25] sobre o consumo de alimentos açucarados sugere um aumento do risco de câncer de próstata nos homens que consomem sucos de frutas com maior frequência. Os resultados deste trabalho indicam a importância de estudos complementares sobre a influência do consumo de bebidas açucaradas, sucos artificiais e industrializados na taxa de DCP no contexto brasileiro. Portanto, estudos longitudinais serão úteis para analisar o possível efeito de fatores como o alto nível de colesterol e o consumo de sucos artificiais e industrializados na taxa de ocorrência de câncer de próstata.

As árvores de decisão também foram usadas como modelos para a previsão de casos de câncer de próstata. A base de treino foi usada na parametrização do modelo o qual foi aplicado em seguida na base de teste. Os resultados de acurácia, sensibilidade, especificidade e ROC-AUC para as árvores de decisão de 2 e 3 níveis são mostrados na Tabela 2, referentes às bases de treino e de teste. O modelo de 2 níveis apresenta bom desempenho, tanto na base de treino quanto na base de teste. Há uma leve tendência para uma melhor reprodução dos eventos positivos de câncer de próstata (dado pela sensibilidade) em relação aos eventos negativos (dado pela especificidade). No caso da árvore de 3 níveis, houve um aumento considerável na sensibilidade em relação à árvore de 2 níveis, sendo um resultado muito satisfatório no aspecto da reprodução de casos positivos de câncer de próstata nas bases de treino e de teste; contudo, a árvore de 3 níveis é um pouco menos eficiente na reprodução dos casos negativos, o que se verifica pela leve diminuição da especificidade.

Em complemento aos resultados da Tabela 2, calculou-se o valor preditivo positivo (VPP) para os modelos nas bases de treino e de teste, obtendo-se valores entre 3% e 4%. Resultados dessa ordem de grandeza são esperados, pois a base de 42.799 homens analisada tem mais de 99% de indivíduos que disseram "não ter recebido" DCP. Em razão da grande predominância de casos negativos na base, é natural que os modelos acabem gerando, em termos absolutos, um número alto de casos falso-positivos (da ordem de alguns milhares), que tem um efeito relativamente brando para a especificidade e acurácia, mas reduz fortemente o VPP. Portanto, tal limitação preditiva não desmerece os resultados bem-sucedidos apresentados nas outras métricas.

Em princípio, os modelos apresentados poderiam ser usados para identificar homens com características físicas, socioeconômicas e de hábitos de vida que os tornam mais propensos a receber DCP. Porém, deve-se levar em conta que há uma diferença importante entre desenvolver uma doença e receber o diagnóstico da doença. Existem variáveis socioeconômicas como renda familiar, possuir ou não plano de saúde, nível de instrução etc. que podem influenciar a obtenção do diagnóstico precoce do câncer de próstata. Isto é, homens em situação socioeconômica deficiente (baixa renda, baixo nível de instrução etc.) são mais suscetíveis a um subdiagnóstico. Por isso, se os modelos aqui apresentados forem aplicados com o objetivo restrito de identificar homens com tendência a desenvolver o câncer de próstata, então deve-se atentar que o modelo terá naturalmente limitações, dependendo do grupo socioeconômico analisado.

Os resultados deste estudo podem ser investigados mais profundamente em pesquisas futuras, incluindo a

influência do consumo de certos alimentos na ocorrência do câncer de próstata em estudos longitudinais, as diferenças estatísticas entre grupos sociais no diagnóstico da doença, a relação entre o câncer de próstata e outras doenças, entre outros aspectos. O cruzamento da variável dependente DCP com as variáveis descritivas nem sempre permite uma interpretação clara dos resultados, pois algumas variáveis podem sofrer uma influência significativa de fatores socioeconômicos, como é descrito na análise exploratória. Além disso, há algumas variáveis descritivas cujas perguntas correspondentes da PNS não foram respondidas pela totalidade dos indivíduos (usada a sigla N.A. no caso de perguntas não aplicadas).

Deve-se levar em conta que o questionário da PNS não foi planejado especificamente para a análise aprofundada do câncer de próstata. Uma pergunta importante não incluída no questionário é referente aos casos de câncer de próstata em membros da família. Pesquisas na área[3] mostram que ter um parente de primeiro grau com DCP aumenta significativamente o risco de um homem desenvolver a doença. Em geral, o questionário da PNS faz perguntas que não esclarecem como foram os hábitos de vida e o estado de saúde da pessoa ao longo da sua vida, isto é, a maioria das perguntas faz somente um "retrato" do entrevistado no momento daquela pesquisa. Por isso, os modelos aqui apresentados podem ser aperfeiçoados futuramente se houver uma base de dados histórica da pessoa entrevistada e da sua família, com perguntas que se referem à evolução da sua saúde e hábitos de vida.

A análise exploratória dessa base de dados e os modelos desenvolvidos podem ser usados para estimativas das demandas do sistema público ou privado de saúde, como recursos humanos e infraestrutura, para a prevenção e tratamento do câncer de próstata em grupos específicos ou regiões específicas do Brasil, bem como para identificar grupos de indivíduos que sejam alvos preferenciais de campanhas de prevenção ao câncer de próstata.

CONCLUSÃO

Os modelos de *machine learning* aplicados aos dados da PNS indicam uma associação significativa de fatores socioeconômicos, físicos e de hábitos de vida com o DCP no Brasil. Os modelos de árvore de decisão mostram que as variáveis "idade", "diagnóstico de alto nível de colesterol", "se possui plano de saúde" e "nível de instrução" têm forte associação com a taxa de DCP, em diferentes grupos de indivíduos. O alto nível de acurácia (próximo ou acima de 80%) e sensibilidade (entre 80% e 90%) dos modelos mostra o potencial dos métodos de *machine learning* para o estudo e prevenção do câncer de próstata no contexto brasileiro, especialmente se houver a disponibilidade de bases de dados longitudinais sobre a doença no futuro.

Esta pesquisa apresenta resultados úteis para o planejamento no uso de recursos públicos ou privados no tratamento ou prevenção do câncer de próstata no contexto brasileiro, bem como para o direcionamento de pesquisas futuras sobre a doença.

CONTRIBUIÇÕES

Marco Antonio de Souza contribuiu substancialmente na concepção e no planejamento do estudo; na obtenção, análise e interpretação dos dados; na redação e revisão crítica. Camila Nascimento Monteiro e Cláudia Renata dos Santos Barros contribuíram na análise e interpretação dos dados; na redação e revisão crítica. Todos os autores aprovaram a versão final a ser publicada.

DECLARAÇÃO DE CONFLITO DE INTERESSES

Nada a declarar.

FONTES DE FINANCIAMENTO

Não há.

REFERÊNCIAS

1. Santos MO, Lima FCS, Martins LFL, et al. Estimativa de incidência de câncer no Brasil, 2023-2025. Rev Bras Cancerol. 2023;69(1):e-213700. doi: https://doi.org/10.32635/2176-9745.RBC.2023v69n1.3700
2. Prostate Cancer Foundation [Internet]. Santa Monica: PCF; [2023]. Prostate Cancer Survival Rates. [acesso 2024 mar 1]. Disponível em: https://www.pcf.org/about-prostate-cancer/what-is-prostate-cancer/prostate-cancer-survival-rates/
3. Instituto Oncoguia [Internet]. São Paulo: Oncoguia; 2015. Fatores de Risco para Câncer de Próstata. 2023 nov 22. [acesso 2024 mar 1 atualizado em 2024 abr 16]. Disponível em: http://www.oncoguia.org.br/conteudo/fatores-de-risco-para-cancer-de-prostata/5850/1130/
4. Krüger FPG, Cavalcanti G. Conhecimento e atitudes sobre o câncer de próstata no Brasil: revisão integrativa. Rev Bras Cancerol. 2018;64(4):561-67. doi: https://doi.org/10.32635/2176-9745.RBC.2018v64n4.206
5. Gomes R, Rebello LEFS, Araújo FC, et al. A prevenção do câncer de próstata: uma revisão da literatura. Ciênc saúde coletiva. 2008;13(1):235-46. doi: https://doi.org/10.1590/S1413-81232008000100027
6. Zacchi SR, Amorim MHC, Souza MAC, et al. Associação de variáveis sociodemográficas e clínicas com o estadiamento inicial em homens com câncer de próstata. Cad saúde colet. 2014;22(1):93-100. doi: https://doi.org/10.1590/1414-462X201400010014

7. Moraes-Araújo MS, Sardinha AHL, Figueiredo Neto JA, et al. Caracterização sociodemográfica e clínica de homens com câncer de próstata. Rev Salud Pública. 2019;21(3):362-67. doi: https://doi.org/10.15446/rsap.V21n3.70678
8. Steffen RE, Trajman A, Santos M, et al. Rastreamento populacional para o câncer de próstata: mais riscos que benefícios. Physis. 2018;28(2):e280209. doi: https://doi.org/10.1590/S0103-73312018280209
9. Conceição MBM, Boing AF, Peres KG. Time trends in prostate cancer mortality according to major geographic regions of Brazil: an analysis of three decades. Cad Saúde Pública. 2014;30(3):559-66. doi: https://doi.org/10.1590/0102-311X00005813
10. Jerez-Roig J, Souza DLB, Medeiros PFM, et al. Future burden of prostate cancer mortality in Brazil: a population-based study. Cad Saúde Pública. 2014;30(11):2451-58. doi: https://doi.org/10.1590/0102-311X00007314
11. Evangelista FM, Melanda FN, Modesto VC, et al. Incidência, mortalidade e sobrevida do câncer de próstata em dois municípios com alto índice de desenvolvimento humano de Mato Grosso, Brasil. Rev Bras Epidemiol. 2022;25:25(Supl 1):e220016. doi: https://doi.org/10.1590/1980-549720220016.supl.1.1
12. Fundação Oswaldo Cruz [Internet]. Rio de Janeiro: Fiocruz; [2000]. Pesquisa mostra expansão de aplicações de inteligência artificial contra o câncer. 2024 jan 22. [acesso 2024 mar 1]. Disponível em: https://portal.fiocruz.br/noticia/pesquisa-mostra-expansao-de-aplicacoes-de-inteligencia-artificial-contra-o-cancer
13. Braga L, Lopes R, Alves L, et al. The global patent landscape of artificial intelligence applications for cancer. Nat Biotechnol. 2023;41:1679-87. doi: https://doi.org/10.1038/s41587-023-02051-9
14. Instituto Brasileiro de Geografia e Estatística [Internet]. Rio de Janeiro: IBGE; 2014. PNS: Pesquisa Nacional de Saúde. Microdados. [acesso 2024 mar 1]. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=&t=microdados
15. Conselho Nacional de Saúde (BR). Resolução n° 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais cujos procedimentos metodológicos envolvam a utilização de dados diretamente obtidos com os participantes ou de informações identificáveis ou que possam acarretar riscos maiores do que os existentes na vida cotidiana, na forma definida nesta Resolução [Internet]. Diário Oficial da União, Brasília, DF. 2016 maio 24 [acesso 2024 mar 1]; Seção I:44. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/cns/2016/res0510_07_04_2016.html
16. R: The R Project for Statistical Computing [Internet]. Version 4.4.0 [place unknown]: The R foundation. 2024 abr 24. [acesso 2024 mar 1]. Disponível em: https://www.r-project.org/
17. RStudio [Internet]. Version 2024.04.1+748. Boston: Posit Software, PBC. 2024 abr 1. [acesso 2024 mar 1]. Disponível em: http://www.rstudio.com/ide
18. Victora CG, Horta BL, Mola CL, et al. Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. Lancet Glob Health. 2015;3(4):E199-E205. doi: https://doi.org/10.1016/S2214-109X(15)70002-1
19. Paim J, Travassos C, Almeida C, et al. The Brazilian health system: history, advances, and challenges. Lancet. 2011;377(9779):1778-97. doi: https://doi.org/10.1016/S0140-6736(11)60054-8
20. Pelton K, Freeman MR, Solomon KR. Cholesterol and prostate cancer. Curr Opin Pharmacol. 2012;12(6):751-9. doi: https://doi.org/10.1016/j.coph.2012.07.006
21. Jamnagerwalla J, Howard LE, Allott EH, et al. Serum cholesterol and risk of high-grade prostate cancer: results from the REDUCE study. Prostate Cancer Prostatic Dis. 2018;21(2):252-59. doi: https://doi.org/10.1038/s41391-017-0030-9
22. Johns Hopkins Medicine [Internet]. Cholesterol, prostate cancer, and race. Baltimore: Johns Hopkins Medicine. 2021 dez 11. [acesso 2024 mar 1]. Disponível em: https://www.hopkinsmedicine.org/news/articles/cholesterol-prostate-cancer-and-race
23. Miles FL, Neuhouser ML, Zhang Z-F. Concentrated sugars and incidence of prostate cancer in a prospective cohort. Br J Nutr. 2018;120(6):703-10. doi: https://doi.org/10.1017/S0007114518001812
24. Llaha F, Gil-Lespinard M, Unal P, et al. consumption of sweet beverages and cancer risk. a systematic review and meta-analysis of observational studies. Nutrients. 2021;13(2):516. doi: https://doi.org/10.3390/nu13020516
25. Makarem N, Bandera EV, Lin Y, et al. Consumption of sugars, sugary foods, and sugary beverages in relation to adiposity-related cancer risk in the framingham offspring cohort (1991–2013). Cancer Prev Res 2018;11(6):347-58. doi: https://doi.org/10.1158/1940-6207.CAPR-17-0218

Recebido em 18/3/2024
Aprovado em 3/5/2024

Editora-associada: Jeane Tomazelli. Orcid iD: https://orcid.org/0000-0002-2472-3444
Editora-científica: Anke Bergmann. Orcid iD: https://orcid.org/0000-0002-1972-8777