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

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

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

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.

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.

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.

[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
Dirección para correspondencia: 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

INTRODUCCIÓN

En el Brasil, el cáncer de próstata es el segundo más común entre los hombres[1]. En los Estados Unidos, se estimó que uno de cada ocho hombres padecerá cáncer de próstata en el transcurso de su vida[2]. Este tipo de cáncer es multicausal y los principales factores de riesgo identificados son edad, raza/etnia, nacionalidad, antecedentes familiares y alteraciones genéticas. Hay también otros factores asociados a los hábitos de vida relacionados con el cáncer de próstata que han sido estudiados para un mejor conocimiento de posible relación causal, entre ellos: dieta, obesidad, tabaquismo, exposición ocupacional, inflamación de la próstata y enfermedades sexualmente transmisibles[3]. Además, factores sociales y económicos tienen una fuerte influencia sobre los hábitos de vida de la población y sus condiciones de acceso a los servicios de salud, pudiendo, en teoría, influir en la ocurrencia y el diagnóstico de cáncer de próstata (DCP) en la población masculina.

Estudios anteriores fueron ya publicados sobre el cáncer de próstata en el Brasil en los aspectos de salud pública y epidemiología. Por ejemplo, hay estudios de revisión de la literatura sobre el tema con una contextualización para la salud pública en el Brasil[4,5], otros que caracterizan a los individuos con la enfermedad o la estadificación de esta mediante variables clínicas y sociodemográficas[6,7], discusiones sobre la forma de rastreo poblacional del cáncer de próstata[8], y estudios sobre la tendencia temporal de la mortalidad por cáncer de próstata en el Brasil o regiones específicas del país[9-11].

En los últimos años, existe un creciente interés por el uso de métodos de *machine learning* o inteligencia artificial (IA) para la investigación y prevención del cáncer[12,13], por ejemplo, por el análisis de diversas variables que pueden influir en la tasa de ocurrencia de diferentes tipos de cáncer, y por análisis de imágenes médicas. En este contexto, el presente estudio tiene como objetivos: analizar los factores asociados al DCP en el Brasil mediante los datos de la Encuesta Nacional de Salud (ENS) 2019; y probar modelos predictivos de *machine learning* sobre el cáncer de próstata con el uso de estos datos.

MÉTODO

Se trata de un estudio transversal con datos secundarios procedentes de la ENS realizada en 2019. La base de datos contiene 42 799 individuos; de estos, 339 (0,79%) dijeron "haber recibido" DCP en algún momento de su vida y 42 460 (99,21%) "no haber recibido".

Entre las 58 variables seleccionadas inicialmente, estaba la variable dependiente "cáncer de próstata": haber recibido el diagnóstico o no haber recibido (sí o no); y 57 eran variables independientes.

Tras la aplicación de criterios de eliminación de variables que se explicarán en esta sección, quedaron 13 variables independientes, las cuales fueron aplicadas en los modelos predictivos. Ellas son: edad; hace cuántos años la persona consultó a un médico por última vez; cómo la persona evalúa su propia salud (muy buena, buena, mala, etc.); cuántos días por semana consume frutas; cuántos días por semana consume verduras y/o legumbres; cuántos días por semana consume jugos artificiales; si ha recibido diagnóstico de nivel alto de colesterol (sí o no); si manipula o manipulaba sustancias químicas en el trabajo que son potencialmente perjudiciales para la salud (sí o no); si tiene seguro médico privado (sí o no); raza/etnia (blanca, negra, parda etc.); nivel de educación (primaria completa, secundaria completa, superior incompleta, etc.); si fuma algún producto del tabaco, y con cuál frecuencia (no fuma, diariamente, menos que diariamente); y si ha recibido diagnóstico de depresión (sí o no).

La base original se obtuvo del sitio web del Instituto Brasileño de Geografía y Estadística (IBGE), en la sección de la ENS, subsección de microdatos[14]. Esta encuesta domiciliaria de salud fue realizada en 2019, con una muestra representativa de la población brasileña. La base original, contenida en el archivo PNS_2019.txt, tiene 346 MB, un total de 279 382 líneas (individuos entrevistados) y 817 columnas (características).

La ENS obtuvo la aprobación de la Comisión Nacional de Ética en Pesquisa (Conep) en agosto de 2019 con el número 3.529.376 (CAAE: 11713319.7.0000.0008) para la edición de 2019. Como esta investigación utilizó datos puestos a disposición pública, se justifica que, para el presente estudio, no hay necesidad de análisis por parte del Comité de Ética en Pesquisa (CEP), de acuerdo con la Resolución n.º 510/2016[15] del Consejo Nacional de Salud (CNS).

En el primer proceso de filtrado hecho por la biblioteca PNS-IBGE[14] del R[16], fue posible extraer un *dataframe* de la base original con las variables de interés. Entonces se extrajo un archivo CSV de 279 382 líneas (individuos entrevistados) y 68 columnas (características). En el segundo proceso de filtrado, la base de datos se redujo para que solo contenga a los individuos de sexo masculino que respondieron dos preguntas que conforman la variable dependiente: a) si el individuo recibió diagnóstico de algún tipo de cáncer en su vida (1 = sí, 2 = no); b) si el individuo recibió DCP a lo largo de su vida (1 = sí, 2 = no). Luego del 2º proceso de filtrado, se obtuvo un *dataframe* de 42 799 líneas (individuos) y 58 columnas (características). La base de datos fue entonces debidamente preparada para ser usada en los modelados, incluyendo el tratamiento

de datos faltantes (*missing*) y la organización de variables categóricas, resultando en una base de datos final con 42 799 líneas (individuos) y 58 variables.

Con la base de datos debidamente preparada, se realizaron los procedimientos de filtrado de las variables de mayor relevancia, análisis estadístico y modelado usando métodos de *machine learning*. En el caso de este trabajo, los modelos de clasificación escogidos para la descripción de la variable dependiente "DCP" fueron: regresión logística y árbol de decisión. Tales procedimientos se detallan a continuación.

Antes de la aplicación de la base de datos en los modelos de regresión logística y árbol de decisión, se hizo un filtrado inicial de las variables descriptivas usando *information value* (IV), calculado por el RStudio[17]. Todas las variables con IV considerado muy débil (≤ 0,02) fueron excluidas de los modelos previamente, quedando 42 variables descriptivas en esta etapa.

Tras el filtrado usando el IV, se aplicó el modelo de regresión logística con el uso del RStudio[17], considerando como variable dependiente el DCP (variable binaria, con **1** para "sí", y **0** para "no"). Para esto, la base de datos fue dividida en base de entrenamiento (70% de los individuos) y base para pruebas (30%). La división de la base de datos en entrenamiento y pruebas es importante para realizar la validación cruzada del modelo. Por lo tanto, todo el modelo es desarrollado (o parametrizado) con la base de entrenamiento, y, enseguida, validado en la base para pruebas mediante métricas como la exactitud, sensibilidad y especificidad. El modelo de regresión logística permite prever si la variable dependiente tendrá resultado positivo o negativo para un determinado individuo de la base.

Usando la base de entrenamiento, se aplicó el método *backward* y un nivel de significación de 0,10 (o sea, $p \leq 0,10$) para la selección de las variables. Después de este proceso, se llegó a un modelo final con un total de 13 variables descriptivas, todas categóricas, las cuales fueron citadas anticipadamente en el método.

El modelo de árbol de decisión fue ejecutado en el RStudio[17], con las mismas 13 variables descriptivas, considerando árboles de 2 y 3 niveles. El árbol de decisión también se utiliza como un modelo predictivo para la variable dependiente DCP, y, además, permite identificar a grupos específicos con mayor frecuencia de casos positivos para la variable dependiente, de acuerdo con el nivel de asociación con las variables descriptivas.

RESULTADOS

Entre las variables descriptivas seleccionadas, la que presentó el valor más alto de IV fue la variable "edad" (IV = 2,86), lo que representa una capacidad predictiva muy fuerte en relación con el cáncer de próstata. Se verifica que la edad promedio de los hombres entrevistados es de 45,9 años y la mediana es de 45 años; esto muestra que el grupo del 50% de los individuos por encima de la mediana contiene al grupo etario de mayor propensión a recibir DCP (Figura 1). Observando el Gráfico 1(c), se nota que la mediana de edad de los hombres que respondieron "sí" para DCP (72 años) está muy por encima de la mediana de los hombres que respondieron "no" (44 años).

La Tabla 1 muestra la distribución de los individuos en las categorías de las variables descriptivas y el cruce con la variable dependiente DCP; los resultados que pueden ser contextualizados más claramente son:

Edad: aumento de la frecuencia de DCP a partir de los 50 años, con frecuencia más alta en la categoría "≥ 80 años".

Consultas médicas: mayor frecuencia de DCP en los individuos que tuvieron su última consulta médica más próxima del momento de la encuesta (hasta dos años antes).

Autoevaluación de salud: mayor frecuencia de DCP en los individuos que afirmaron tener una calidad de salud "mala o muy mala".

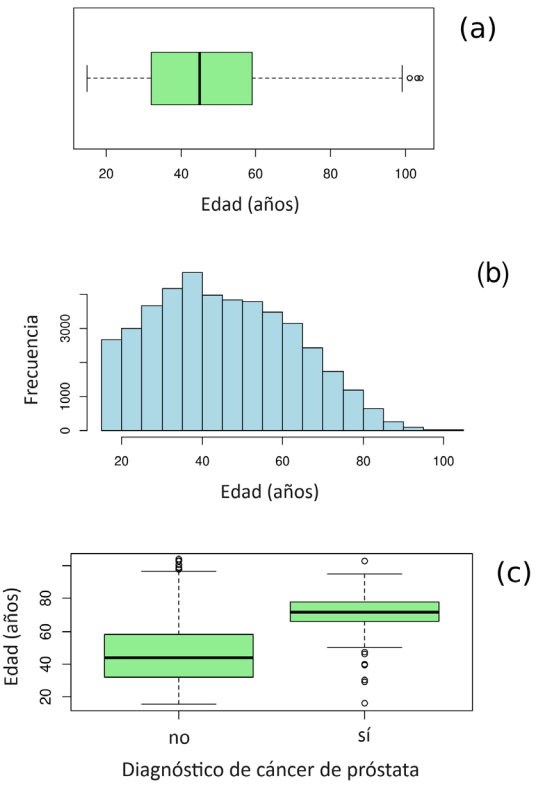

Figura 1. Descripción de la variable "edad" de los individuos de sexo masculino seleccionados en la base de datos de la Encuesta Nacional de Salud (ENS). (a) Diagrama de caja mostrando la distribución de edad. (b) Histograma mostrando la distribución de edad. (c) Comparación en diagrama de caja de las distribuciones de edad en los grupos: con diagnóstico de cáncer de próstata en algún momento de la vida (sí); sin diagnóstico de cáncer de próstata (no)

Tabla 1. Variables descriptivas usadas en los modelos finales de regresión logística y árbol de decisión. Se muestran las categorías de cada variable, la frecuencia de individuos en cada categoría, y la frecuencia de casos positivos (sí) y negativos (no) de cáncer de próstata en cada categoría. *N.A.*: pregunta no aplicada

Variable descriptiva			Cáncer de próstata (%)	
Edad	**Grupo de edad (años)**	**Frecuencia (%)**	**No**	**Sí**
	< 35	29,46	99,98	0,02
	≥ 35 y < 50	29,45	99,96	0,04
	≥ 50 y < 65	24,76	99,35	0,65
	≥ 65 y < 80	13,50	96,63	3,37
	≥ 80	2,83	94,48	5,52
Consultas médicas	**Tiempo desde la última consulta médica (años)**	**Frecuencia (%)**	**No**	**Sí**
	hasta 2 años	83,02	99,05	0,95
	más de 2 años	15,77	99,96	0,04
	nunca fue	1,21	100,00	0,00
Autoevaluación de salud	**Autoevaluación**	**Frecuencia (%)**	**No**	**Sí**
	muy buena o buena	66,77	99,54	0,46
	regular	27,82	98,58	1,42
	mala o muy mala	5,41	98,36	1,64
Consumo de frutas	**Consumo semanal**	**Frecuencia (%)**	**No**	**Sí**
	1 a 3 días	40,45	99,53	0,47
	4 a 6 días	20,61	99,29	0,71
	nunca o muy poco	13,10	99,48	0,52
	todos los días	25,84	98,50	1,50
Consumo de jugo artificial[a]	**Consumo semanal**	**Frecuencia (%)**	**No**	**Sí**
	1 a 3 días	21,37	99,68	0,32
	4 a 6 días	7,89	99,62	0,38
	nunca o muy poco	64,08	99,00	1,00
	todos los días	6,66	99,16	0,84
Diagnóstico de colesterol alto	**Respuesta**	**Frecuencia (%)**	**No**	**Sí**
	no o *N.A.*	89,12	99,35	0,65
	sí	10,88	98,02	1,98
Exposición a productos químicos en el trabajo[b]	**Respuesta**	**Frecuencia (%)**	**No**	**Sí**
	no o *N.A.*	87,21	99,12	0,88
	sí	12,79	99,82	0,18
Tiene seguro médico privado	**Respuesta**	**Frecuencia (%)**	**No**	**Sí**
	no	78,82	99,38	0,62
	sí	21,18	98,58	1,42
Raza/etnia	**Raza/etnia**	**Frecuencia (%)**	**No**	**Sí**
	amarilla	0,73	98,72	1,28
	blanca	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
	negra	11,82	99,11	0,89

continua

Tabla 1. continuación

Variable descriptiva			Cáncer de próstata (%)	
Nivel de educación	**Nivel de educación**	**Frecuencia (%)**	**No**	**Sí**
	sin educación o primaria incompleta	42,67	99,03	0,97
	primaria completa o secundaria incompleta	15,48	99,41	0,59
	secundaria completa o superior incompleta	28,93	99,43	0,57
	superior completa	12,92	99,06	0,94
Consumo de verduras y legumbres	**Consumo semanal**	**Frecuencia (%)**	**No**	**Sí**
	1 a 3 días	35,67	99,34	0,66
	4 a 6 días	21,24	99,21	0,79
	nunca o muy poco	9,54	99,46	0,54
	todos los días	33,55	98,99	1,01
Fuma actualmente	**Uso semanal**	**Frecuencia (%)**	**No**	**Sí**
	diariamente	14,32	99,45	0,55
	menos que diario	1,75	99,73	0,27
	no fuma	83,93	99,16	0,84
Tuvo diagnóstico de depresión	**Respuesta**	**Frecuencia (%)**	**No**	**Sí**
	no	95,43	99,24	0,76
	sí	4,57	98,46	1,54

a De acuerdo con la Encuesta Nacional de Salud (ENS), esta variable se refiere al consumo de los llamados jugos de "cajita", en lata o refresco en polvo.

b De acuerdo con la ENS, esta variable se refiere a la manipulación de productos químicos como: pesticidas agrícolas, gasolina, diésel, formol, plomo, mercurio, cromo, quimioterápicos, etc.

Diagnóstico de colesterol alto: mayor frecuencia de DCP en los individuos que recibieron diagnóstico de nivel alto de colesterol.

Seguro médico privado: mayor frecuencia de DCP en los individuos que afirman tener seguro médico privado.

Diagnóstico de depresión: mayor frecuencia de DCP en los individuos que recibieron diagnóstico de depresión.

La selección de variables descriptivas como "seguro médico privado" y "nivel de educación" indica una influencia de la situación socioeconómica en los modelos predictivos. De hecho, tal indicación puede verificarse por el cruce de la variable dependiente DCP con los ingresos familiares per cápita de los individuos. Considerando el grupo con edades desde los 50 años, se observan tasas de DCP del 0,84%, 1,82%, 2,49% y 3,07% para los respectivos segmentos de ingresos familiares per cápita (en salarios mínimos – SM): hasta ½ SM; más de ½ SM hasta 2 SM; más de 2 SM hasta 5 SM; más de 5 SM. Esto es, el segmento con ingresos más altos (más de 5 SM) tiene una tasa de DCP unas 3,7 veces mayor que en el segmento con ingresos más bajos (hasta ½ SM).

Bajo la premisa antes descrita, el posible efecto socioeconómico sobre las variables descriptivas fue investigado; para esto, se determinó el nivel de asociación entre cada variable descriptiva y los ingresos familiares per cápita usando las medidas V de Cramer y ω de Cohen. Se constató un nivel alto de asociación de los ingresos familiares per cápita con las variables "seguro médico privado" (V de Cramer = 0,487 y ω de Cohen = 0,487) y "nivel de educación" (V de Cramer = 0,314 y ω de Cohen = 0,544). Un nivel medio de asociación con los ingresos familiares per cápita fue observado en la variable "raza/etnia" (V de Cramer = 0,150 y ω de Cohen = 0,260), y un nivel razonable fue observado en "consumo de frutas" y "consumo de verduras y legumbres" (ω de Cohen = 0,218 y 0,229, respectivamente). Por lo tanto, la influencia de los ingresos familiares en las variables descriptivas debe tenerse en consideración en la interpretación de los resultados de la Tabla 1.

Los resultados de exactitud, sensibilidad, especificidad y ROC-AUC para la regresión logística son mostrados en la Tabla 2**,** referentes a las bases de entrenamiento y para pruebas. Se observa que la exactitud, sensibilidad y especificidad para estas bases de datos son muy satisfactorias, pues todas superan el 80%. En la base de datos para pruebas, hubo un pequeño aumento en la exactitud y especificidad, y una pequeña disminución en la sensibilidad. El resultado obtenido para el ROC-AUC en la base de datos de entrenamiento (0,822) también puede ser considerado muy satisfactorio, mientras que el resultado en la base de datos de pruebas (0,780) se encuentra muy próximo de la misma condición.

Los resultados obtenidos para los árboles de decisión de 2 y 3 niveles son mostrados gráficamente en las Figuras 2 y 3, respectivamente, referentes a la base de entrenamiento. Observando las frecuencias de casos de cáncer de próstata en los nodos finales del árbol de **2 niveles** se constata que:

- Los **nodos 3 y 4** sugieren que existe una probabilidad de ocurrencia de cáncer de próstata en hombres negros comparativamente mayor al conjunto de las otras etnias, considerando el grupo etario menor de 50 años;
- Los **nodos 6 y 7** indican que el grupo de hombres con nivel de colesterol alto tiene una mayor frecuencia de DCP (unas tres veces mayor) en comparación con el grupo de nivel de colesterol normal o desconocido, considerando el grupo etario de 50 a 65 años;
- Los **nodos 9 y 10** indican que el grupo de hombres con nivel de educación entre la educación primaria completa y superior completo tiene una mayor frecuencia de DCP (una dos veces mayor) en comparación con el grupo de menor nivel educativo, considerando al grupo etario entre 65 y 80 años;
- Los **nodos 12 y 13** indican que el grupo de hombres que tienen seguro médico privado tiene una mayor frecuencia de DCP (unas dos veces mayor) en comparación con el

Tabla 2. Exactitud, sensibilidad, especificidad y ROC-AUC obtenidos en los modelos de regresión logística, árbol de decisión de 2 niveles y árbol de decisión de 3 niveles, considerando las bases de entrenamiento (70% del total) y pruebas (30% del total)

Modelo	Base	Exactitud	Sensibilidad	Especificidad	ROC-AUC
Regresión logística	entrenamiento	0,828	0,828	0,828	0,822
	pruebas	0,835	0,802	0,835	0,780
Árbol de decisión (2 niveles)	entrenamiento	0,801	0,850	0,800	0,823
	pruebas	0,805	0,821	0,804	0,812
Árbol de decisión (3 niveles)	entrenamiento	0,767	0,906	0,766	0,832
	pruebas	0,773	0,887	0,772	0,813

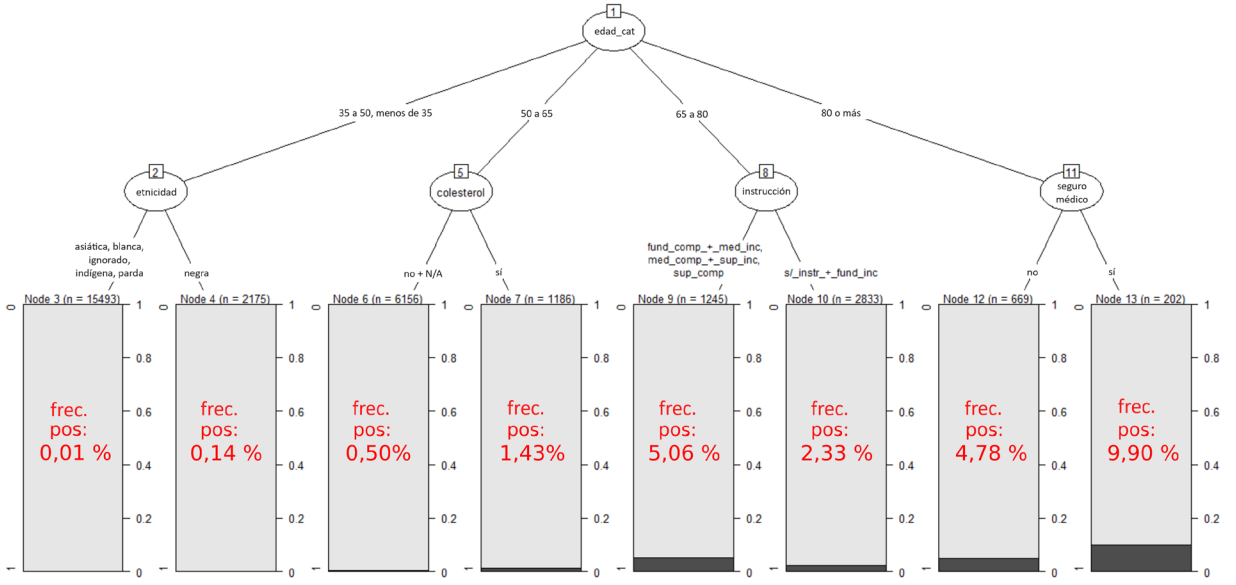

Figura 2. Árbol de decisión de 2 niveles, considerando la base de entrenamiento con el 70% de los individuos seleccionados. Los porcentajes en rojo indican la frecuencia de casos positivos de cáncer de próstata en los nodos finales del árbol

Leyendas: *edad_cat* = edad en la forma categórica (años); *etnicidad* = raza/etnia; *colesterol* = nivel de colesterol alto; *instrucción* = nivel de educación; *seguro_médico* = seguro médico privado; *n_a* = pregunta no aplicada; *fund_comp* = nivel primario completo; *med_inc* = nivel secundario incompleto; *med_comp* = nivel secundario completo; *sup_inc* = nivel superior incompleto; *sup_comp* = nivel superior completo; *s/_instr* = sin educación; y *fund_inc* = nivel primario incompleto.

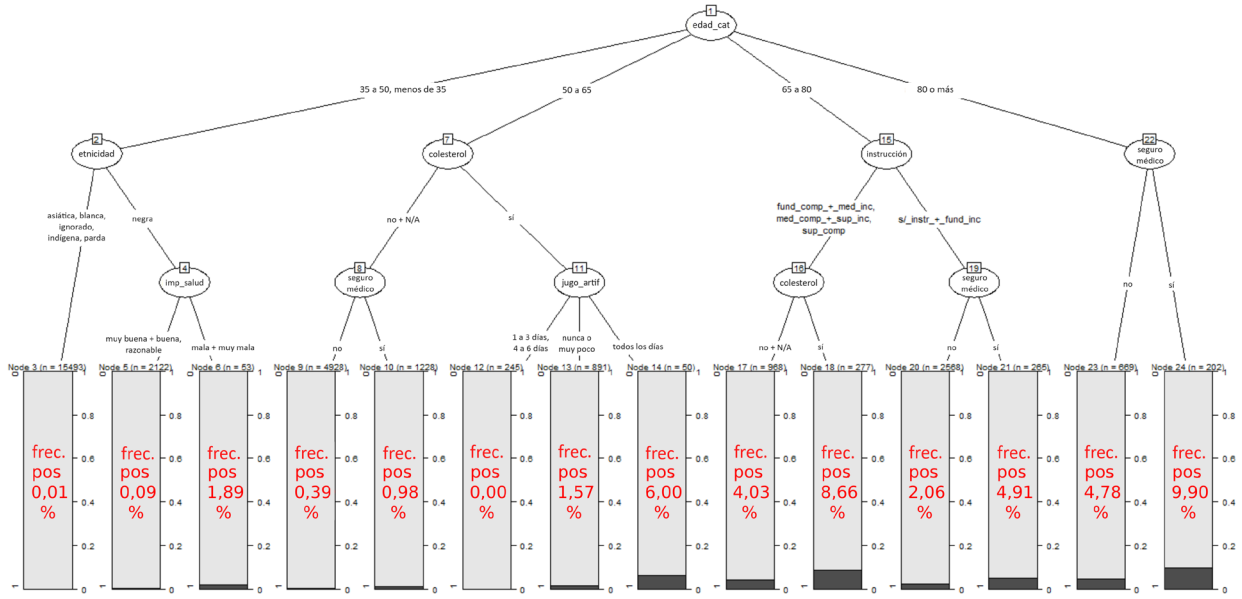

Figura 3. Arbol de decisión de 3 niveles, considerando la base de datos de entrenamiento con el 70% de los individuos seleccionados. Los porcentajes en rojo indican la frecuencia de casos positivos de cáncer de próstata en los nodos finales del árbol

Leyendas: *edad_cat* = edad en la forma categórica (años); *etnicidad* = raza/etnia; *colesterol* = nivel de colesterol alto; *instrucción* = nivel de educación; *seguro_médico* = seguro médico privado; *imp_salud* = impresión sobre la propia salud; *jugo_artif* = consumo semanal de jugo artificial; *n_a* = pregunta no aplicada; *fund_comp* = nivel primario completo; *med_inc* = nivel secundario incompleto; *med_comp* = nivel secundario completo; *sup_inc* = nivel superior incompleto; *sup_comp* = nivel superior completo; *s/_instr* = sin educación; *fund_inc* = nivel fundamental incompleto; y *muy_poco* = muy poco

grupo que no tiene seguro médico privado, considerando al grupo etario a partir de 80 años.

Con relación al nodo 5 (el cual resulta en los nodos finales 6 y 7), se puede evaluar el nivel de asociación entre el factor de riesgo "colesterol alto" y el resultado DCP para el subgrupo de hombres con edad entre 50 y 65 años, utilizando el *odds ratio* (OR). En este caso, se obtuvo OR = 2,87 (intervalo de confianza – IC 95%: 1,58 – 5,21), representando un nivel de asociación significativo.

Al interpretar los resultados de los nodos 9 y 10, se debe considerar la fuerte asociación entre nivel de educación e ingresos familiares per cápita discutida anteriormente[18]. Respecto a los nodos 12 y 13, se debe tener en cuenta las diferentes características de los sistemas público y privado de salud en el Brasil, además de la fuerte asociación entre las variables "seguro médico privado" e ingresos familiares per cápita discutida anteriormente[19].

Los principales resultados del árbol de decisión de **3 niveles** se describen a continuación:

- Entre los hombres con edad entre 50 y 65 años, y que tienen nivel de colesterol normal o desconocido, se observa que la frecuencia de DCP es unas 2,5 veces mayor en el subgrupo que tiene seguro médico privado en comparación con el subgrupo que no tiene seguro médico privado;
- Entre los hombres con edad a partir de 50 y < 65 años, y que tienen nivel de colesterol alto, se observa que el subgrupo de hombres que toma jugos artificiales todos los días tiene una frecuencia de DCP unas cuatro veces mayor que el subgrupo de hombres que dice no tomar o tomar muy poco jugos artificiales.
- Entre los hombres con edad entre 65 y 80 años, y que tienen al menos el nivel primario completo de educación, se observa que el subgrupo de hombres con nivel de colesterol alto tiene una frecuencia de DCP unas dos veces mayor que el subgrupo de hombres que tienen nivel de colesterol normal o desconocido.
- Entre los hombres con edad entre 65 y 80 años, y que tienen educación inferior a la primaria completa, se observa que el subgrupo de hombres que tiene seguro médico privado tiene una frecuencia de DCP unas 2,5 veces mayor que el subgrupo de hombres que no tienen seguro médico privado.

El árbol de decisión de 3 niveles muestra resultados que refuerzan aspectos discutidos sobre el árbol de 2 niveles, aunque en grupos más específicos. Por ejemplo, los pares de nodos {9,10}, {20,21} y {23,24} refuerzan la mayor probabilidad de DCP para los individuos que afirman tener seguro médico privado. El par de nodos {17,18} refuerza la mayor probabilidad de DCP para los individuos con alto nivel de colesterol, aunque en el grupo etario entre 65 y 80 años y que tienen al menos el nivel primario completo de educación. Con relación al nodo 16 (el cual resulta en los nodos finales 17 y 18), se puede evaluar el nivel de asociación entre el factor de riesgo "colesterol alto" y el resultado DCP para este subgrupo. En este caso,

se obtuvo OR = 2,26 (IC 95%: 1,33 – 3,83), mostrando nuevamente un nivel de asociación significativo.

Una información relevante se da en el trío de nodos {12,13,14}, el cual indica una probabilidad significativamente mayor de DCP para los individuos que afirman consumir jugos artificiales/industrializados todos los días, dentro del grupo etario entre 50 y 65 años y que tienen alto nivel de colesterol. Sin embargo, este resultado debe ser analizado con precaución, como se discute en la próxima sección.

DISCUSIÓN

La mediana de edad de los hombres que respondieron haber recibido DCP ("sí") fue 72 años, valor muy superior a la mediana de los hombres que respondieron "no" (44 años), resultado que es compatible con la tendencia general en la cual los hombres desarrollan el cáncer de próstata a partir de los 50 años aproximadamente[3].

Con relación a los nodos 3 y 4 del árbol de 2 niveles, se sugiere que las personas negras tienen una mayor tendencia para el desarrollo de cáncer de próstata en edades menores de 50 años; tal constatación parece corroborar estudios anteriores sobre el cáncer de próstata[3], en los cuales se comprobó que esta enfermedad es más frecuente en hombres con ascendencia africana y caribeña. Aun así, los casos de cáncer de próstata debajo de los 50 años son raros en la base (dos casos entre 15 493 individuos en el nodo 3, y tres casos entre 2175 individuos en el nodo 4) y, por esto, se sugiere que conclusiones respecto de este grupo etario deban ser corroboradas por una base de datos más numerosa debajo de los 50 años.

Respecto a los nodos 6 y 7 del árbol de 2 niveles, se debe considerar estudios como el de Pelton *et al.*[20], el cual afirma que altos niveles de colesterol en la sangre están relacionados a casos de cáncer de próstata más agresivos, y de Jamnagerwalla *et al.*[21], el cual señala que altos niveles de colesterol sérico total y HDL están asociados a un riesgo aumentado de cáncer de próstata de alto grado. Una comunicación del *Johns Hopkins Medicine*[22] describe investigaciones más recientes que apuntan hacia conclusiones semejantes. Como el presente estudio es transversal, no se puede afirmar con seguridad que corrobora los resultados de los trabajos anteriores, pero se indica la importancia de estudios adicionales sobre la relación colesterol/DCP con la inclusión de individuos del Brasil.

El hecho de que el presente estudio sea transversal limita la posibilidad de conclusiones que implican la relación causal por la falta de control de la temporalidad. Otra limitación se refiere al uso de datos secundarios, lo que dificulta la precisión de las respuestas de los entrevistados respecto a algunas variables; un ejemplo claro en ese aspecto es la frecuencia semanal de consumo de jugos artificiales. Aun así, hay estudios anteriores que sugieren la relación entre el alto consumo de bebidas azucaradas y una mayor incidencia de cáncer de próstata, como el de Miles *et al.*[23] y el de Llaha *et al.*[24]. Además, el estudio de Makarem *et al.*[25] sobre el consumo de alimentos azucarados sugiere un aumento del riesgo de cáncer de próstata en los hombres que consumen jugos de frutas con mayor frecuencia. Los resultados de este trabajo indican la importancia de estudios complementarios sobre la influencia del consumo de bebidas azucaradas, jugos artificiales e industrializados en la tasa de DCP en el contexto brasileño. Por lo tanto, estudios longitudinales serán útiles para analizar el posible efecto de factores como el alto nivel de colesterol y el consumo de jugos artificiales e industrializados en la tasa de ocurrencia de cáncer de próstata.

Los árboles de decisión también fueron usados como modelos para la previsión de casos de cáncer de próstata. La base de datos de entrenamiento fue usada en la parametrización del modelo el cual fue aplicado a continuación en la base de datos para pruebas. Los resultados de exactitud, sensibilidad, especificidad y ROC-AUC para los árboles de decisión de 2 y 3 niveles son mostrados en la Tabla 2, referentes a las bases de entrenamiento y de pruebas. El modelo de 2 niveles presenta buen desempeño, tanto en la base de entrenamiento como en la base de pruebas. Hay una ligera tendencia hacia una mejor reproducción de los eventos positivos de cáncer de próstata (dado por la sensibilidad) en relación con los eventos negativos (dado por la especificidad). En el caso del árbol de 3 niveles, hubo un aumento considerable en la sensibilidad con relación al árbol de 2 niveles, siendo un resultado muy satisfactorio en el aspecto de la reproducción de casos positivos de cáncer de próstata en las bases de entrenamiento y de pruebas; sin embargo, el árbol de 3 niveles es un poco menos eficiente en la reproducción de los casos negativos, lo que se comprueba por la ligera disminución de la especificidad.

Complementando los resultados de la Tabla 2, se calculó el valor predictivo positivo (VPP) para los modelos en las bases de entrenamiento y de pruebas, obteniéndose valores entre el 3% y el 4%. Resultados de este orden de magnitud son esperados, pues la base de datos de 42 799 hombres analizada tiene más del 99% de individuos que dijeron "no haber recibido" DCP. Debido al gran predominio de casos negativos en la base, es natural que los modelos acaben generando, en términos absolutos, un número alto de casos falsos positivos (del orden de algunos millares), que tiene un efecto relativamente suave para la especificidad y exactitud, pero reduce fuertemente el

VPP. Por lo tanto, tal limitación predictiva no desmerece los exitosos resultados presentados en las otras métricas.

En principio, los modelos presentados podrían usarse para identificar a hombres con características físicas, socioeconómicas y de hábitos de vida que los vuelven más propensos a recibir DCP. Aunque, hay que considerar que existe una diferencia importante entre desarrollar una enfermedad y recibir el diagnóstico de la enfermedad. Existen variables socioeconómicas como ingresos familiares, tener o no seguro médico privado, nivel de educación, etc. que pueden influir en la obtención del diagnóstico temprano del cáncer de próstata. Esto es, hombres en situación socioeconómica deficiente (bajos ingresos, bajo nivel de educación etc.) son más susceptibles a un subdiagnóstico. Por eso, si los modelos aquí presentados se aplicaren con el objetivo restringido de identificar a los hombres con tendencia a desarrollar el cáncer de próstata, entonces se debe considerar que el modelo tendrá naturalmente limitaciones, dependiendo del grupo socioeconómico analizado.

Los resultados de este estudio pueden ser trabajados más profundamente en investigaciones futuras, incluyendo la influencia del consumo de ciertos alimentos en la ocurrencia del cáncer de próstata en estudios longitudinales, las diferencias estadísticas entre grupos sociales en el diagnóstico de la enfermedad, la relación entre el cáncer de próstata y otras enfermedades, entre otros aspectos. El cruce de la variable dependiente DCP con las variables descriptivas ni siempre permite una interpretación clara de los resultados, pues algunas variables pueden sufrir una influencia significativa de factores socioeconómicos, como se describe en el análisis exploratorio. Además, hay algunas variables descriptivas cuyas preguntas correspondientes de la ENS no fueron respondidas por la totalidad de los individuos (usada la sigla N.A. en el caso de preguntas no aplicadas).

Se debe tener en cuenta que el cuestionario de la ENS no fue planeado específicamente para un análisis profundizado del cáncer de próstata. Una pregunta importante no incluida en el cuestionario es referente a los casos de cáncer de próstata en miembros de la familia. Investigaciones en el área[3] muestran que tener un pariente de primer grado con DCP aumenta significativamente el riesgo de que un hombre desarrolle la enfermedad. En general, el cuestionario de la ENS hace preguntas que no aclaran cómo fueron los hábitos de vida y el estado de salud de la persona a lo largo de su vida, esto es, la mayoría de las preguntas hace solo un "retrato" del entrevistado en el momento de aquella encuesta. Por esto, los modelos aquí presentados pueden ser perfeccionados futuramente si hubiere una base de datos histórica de la persona entrevistada y de su familia, con preguntas que se refieren a la evolución de su salud y hábitos de vida.

El análisis exploratorio de esa base de datos y los modelos desarrollados pueden ser usados para estimaciones de las demandas del sistema público o privado de salud, como recursos humanos e infraestructura, para la prevención y tratamiento del cáncer de próstata en grupos o regiones específicos del Brasil, así como para identificar grupos de individuos que sean blancos preferenciales de campañas de prevención del cáncer de próstata.

CONCLUSIÓN

Los modelos de *machine learning* aplicados a los datos de la ENS señalan una asociación significativa de factores socioeconómicos, físicos y de hábitos de vida con el DCP en el Brasil. Los modelos de árbol de decisión muestran que las variables "edad", "diagnóstico de alto nivel de colesterol", "si tiene seguro médico privado" y "nivel de educación" tienen fuerte asociación con la tasa de DCP, en diferentes grupos de individuos. El alto nivel de exactitud (próximo o mayor al 80%) y sensibilidad (entre el 80% y el 90%) de los modelos muestra el potencial de los métodos de *machine learning* para el estudio y prevención del cáncer de próstata en el contexto brasileño, especialmente si hubiere disponibilidad de bases de datos longitudinales sobre la enfermedad en el futuro.

Esta investigación presenta resultados útiles para la planificación en el uso de recursos públicos o privados en el tratamiento o prevención del cáncer de próstata en el contexto brasileño, así como para el direccionamiento de investigaciones futuras sobre la enfermedad.

APORTES

Marco Antonio de Souza contribuyó substancialmente en la concepción y en el planeamiento del estudio; en la obtención, análisis e interpretación de los datos; en la redacción y revisión crítica. Camila Nascimento Monteiro y Cláudia Renata dos Santos Barros contribuyeron en el análisis e interpretación de los datos; en la redacción y revisión crítica. Todos los autores aprobaron la versión final a publicarse.

DECLARACIÓN DE CONFLICTO DE INTERESES

Nada a declarar.

FUENTES DE FINANCIAMIENTO

No hay.

REFERENCIAS

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 nº 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