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

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

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

¿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]

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.

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
Corresponding author: 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

INTRODUCTION

In Brazil, prostate cancer is the second most common cancer type among men[1]. Countries like the United States estimate that one in eight men will be diagnosed with prostate cancer during his life[2]. This type of cancer can have multiple causes, with main risk factors including age, skin color/ethnicity, nationality, family history and genetic alterations. There are also other life habits-associated factors related to prostate cancer that have been studied to better identify a causal relationship, among them: diet, obesity, smoking, occupational exposure, prostate inflammation, and sexually transmitted diseases[3]. Additionally, social and economic factors significantly affect the population's life habits and their conditions of access to health services, which can, theoretically, influence the occurrence and diagnosis of prostate cancer (DPC) in the male population.

Studies on prostate cancer in Brazil in the aspects of public health and epidemiology have already been published. For instance, there are literature review studies on the theme in the context of public health in Brazil[4,5], others that characterize individuals with the disease or the staging of the disease through clinical and sociodemographic variables[6,7], discussions on population screening of prostate cancer[8], and studies about the temporal tendency of prostate cancer mortality in Brazil or specific regions of the country[9-11].

In recent years, there has been a growing interest in machine learning methods or artificial intelligence (AI) for cancer research and prevention[12,13], for instance, through analysis of the several variables that may influence the occurrence rate of different types of cancer, and medical imaging analysis. In this context, the present study aims at analyzing the factors associated with DPC in Brazil through data from the 2019 National Health Survey (NHS) and testing predictive machine learning models on prostate cancer using those data.

METHOD

Cross-sectional study using secondary data from the NHS conducted in 2019. The database comprises 42,799 individuals, with 339 (0.79%) of them claiming to "having received" DPC at some point in their lives and 42,460 (99.21%) claiming to "not having received".

Among the 58 initially selected variables, there was the dependent variable "prostate cancer": having received or not having received (yes or no) the diagnosis; and 57 independent variables.

After applying the elimination criteria of variables (detailed in the next paragraphs of this section), 13 independent variables remained, being applied to the predictive models. The variables are: age; years since last doctor appointment; how the individual assesses their own health (very good, good, bad etc.); their weekly intake of fruits; their weekly intake of vegetables and/or legumes; their weekly intake of artificial juice; if they have already received a high cholesterol diagnosis (yes or no); if they manage or used to manage chemical substances at work that may be prejudicial to health (yes or no); if they have health insurance (yes or no); skin color/ethnicity (white, black, brown etc.); education level (complete elementary school, complete high school, incomplete higher education etc.); if they smoke any tobacco product, and how frequently (doesn't smoke, daily, less than daily); and if they had received a depression diagnosis (yes or no).

The original base was obtained from the *Instituto Brasileiro de Geografia e Estatística* (IBGE), in the NHS section, microdata section[14]. This survey, a home questionnaire about health, was conducted in 2019, with a representative sample of the Brazilian population. The original base, contained in the NHS_2019.txt, has 346 Mb, a total of 279,382 lines (interviewed individuals) and 817 columns (characteristics).

The NHS was approved by the National Research Ethics Committee (CONEP) in August 2019, approval report number 3.529.376 (CAAE (submission for ethical review): 11713319.7.0000.0008) for the 2019 edition. As the present study uses publicly available data, no further approval is needed from the Research Ethics Committee according to Resolution number 510/2016[15] of the National Health Council (NHC).

In the first filtering process, conducted through the R[16] NHS-IBGE library[14], it was possible to extract a data frame of the original base with the variables of interest. A .csv file with 279,382 lines (interviewed individuals) and 68 columns (characteristics) was extracted. In the second filtering process, the database was reduced to contain only male individuals that responded to two questions that compose the dependent variable: a) if they had received any cancer diagnosis (1 = yes, 2 = no); b) if they had received a prostate cancer diagnosis (1 =yes; 2 = no). After the second filtering, a data frame of 42,799 lines (individuals) and 58 columns (characteristics) was obtained. The database was then properly prepared to be used in the models, including treatment of missing data and organization of categorical variables, resulting in a final database of 42,799 lines (individuals) and 58 variables.

With this database ready, the following procedures were conducted: variable filtering for greater relevance, statistical analysis, and machine learning modeling. For this study, the chosen classification models for describing

the dependent variable "DPC" were logistic regression and decision tree. Such procedures are described as follows.

Before applying the database to the logistic regression and decision tree models, an initial filtering of the descriptive variables was conducted via information value (IV), calculated by RStudio[17]. All the variables with very weak IV (≤ 0.02) were previously excluded from the models, leaving 42 descriptive variables at this stage.

After IV filtering, the logistic regression model was applied using RStudio[17], considering DPC as a dependent variable (binary variable, with **1** for "yes", and **0** for "no"). Thus, the database was divided into a training base (70% of individuals) and a test base (30%). The database separation in training and test is important to perform the cross-validation of the model. Therefore, the model is completely developed (or parameterized) with the training base and then validated in the test base through metrics such as accuracy, sensibility, and specificity. The logistic regression model enables predicting if the dependent variable will have a positive or negative result for a certain individual in the base.

Using the training base, the backward method, and a significance level of 0.10 (that is, $p \leq 0.10$) were applied for the selection of variables. After this process, the final model was obtained with a total of 13 descriptive variables, all categorical, which are mentioned in the 3rd paragraph of this section.

The decision tree model was executed in RStudio[17] with the same 13 variables, considering trees of 2 and 3 levels. The decision tree is also used as a predictive model for the dependent DPC variable and, additionally, enables the identification of specific groups with greater frequency of positive cases for the dependent variable, according to the level of association with descriptive variables.

RESULTS

Among the selected descriptive variables, the one that presented the highest IV value was "age" (IV = 2.86), which represents a strong predictive capacity regarding prostate cancer. The verified mean age of the interviewed men was 45.9 years old, with a median of 45; this shows that the 50% group of individuals above the median contain the age group more prone to receive a DPC (Figure 1). Graph 1(c) shows that the median age of men that responded "yes" to DPC (72 years old) is well above the median age of men who responded "no" (44 years old).

Table 1 shows the distribution of individuals in the descriptive variables categories and the crossing with dependent variable DPC; the results that can be more clearly contextualized are:

Age: DPC frequency increase starting at 50 years old, with a higher frequency at the "≥ 80 years old" category;
Medical appointments: greater DPC frequency in individuals whose last medical appointment was closer to the research's date (up to two years prior);
Health self-assessment: greater DPC frequency in individuals who claimed to have a "bad or very bad" health;
High cholesterol diagnosis: greater DPC frequency in individuals who received a high cholesterol diagnosis;
Health insurance: greater DPC frequency in individuals who claimed to have an insurance;
Depression diagnosis: greater DPC frequency in individuals who received a depression diagnosis.

The selection of descriptive variables such as "health insurance" and "education level" indicates an influence of the socioeconomic situation in the predictive models. In fact, such an indication may be verified by crossing the DPC dependent variable with the individuals' *per capita* family income. Considering the 50 years and older age group, the following DPC rates of 0.84%, 1.82%, 2.49% and 3.07% were observed for the respective *per capita*

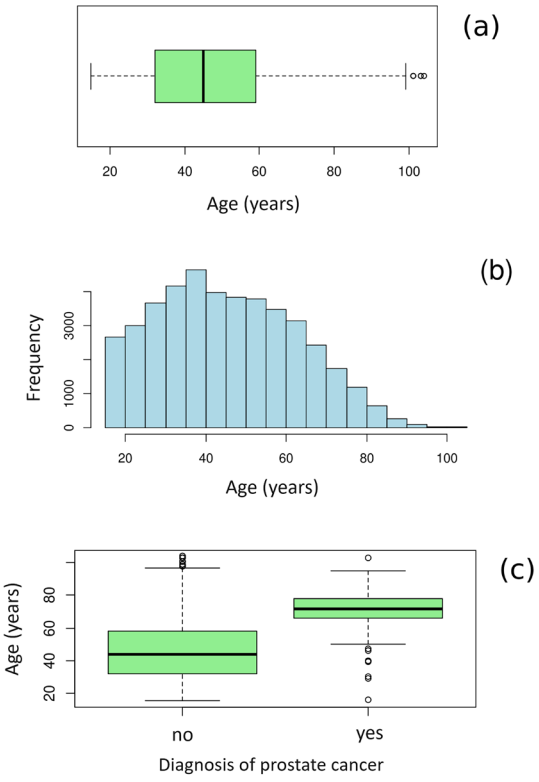

Figure 1. Description of "age" variable of male individuals selected from the National Health Survey base. **(a)** Box plot showing age distribution. **(b)** Histogram showing age distribution. **(c)** Box plot comparison of age distribution in the following groups: with prostate cancer diagnosis at some point in life (yes); with no prostate cancer diagnosis (no)

Table 1. Descriptive variables used on the logistic regression and decision tree final models. The table shows the categories of each variable, the frequency of individuals in each category, and the frequency of positive (yes) and negative (no) cases of prostate cancer in each category. N.A.: non-applied question

Descriptive Variable			Prostate cancer (%)	
Age	**Age group (years)**	**Frequency (%)**	**No**	**Yes**
	< 35	29.46	99.98	0.02
	≥ 35 and < 50	29.45	99.96	0.04
	≥ 50 and < 65	24.76	99.35	0.65
	≥ 65 and < 80	13.50	96.63	3.37
	≥ 80	2.83	94.48	5.52
Medical appointments	**Time since last medical appointment (years)**	**Frequency (%)**	**No**	**Yes**
	Up to 2 years	83.02	99.05	0.95
	Over 2 years	15.77	99.96	0.04
	Never been	1.21	100.00	0.00
Health self-assessment	**Self-assessment**	**Frequency (%)**	**No**	**Yes**
	Very good or good	66.77	99.54	0.46
	Reasonable	27.82	98.58	1.42
	Bad or very bad	5.41	98.36	1.64
Fruit intake	**Weekly intake**	**Frequency (%)**	**No**	**Yes**
	1 to 3 days	40.45	99.53	0.47
	4 to 6 days	20.61	99.29	0.71
	Never or too little	13.10	99.48	0.52
	Every day	25.84	98.50	1.50
Artificial juice intake[a]	**Weekly intake**	**Frequency (%)**	**No**	**Yes**
	1 to 3 days	21.37	99.68	0.32
	4 to 6 days	7.89	99.62	0.38
	Never or too little	64.08	99.00	1.00
	Every day	6.66	99.16	0.84
High cholesterol diagnosis	**Answer**	**Frequency (%)**	**No**	**Yes**
	No or N.A.	89.12	99.35	0.65
	Yes	10.88	98.02	1.98
Exposure to chemical products at work[b]	**Answer**	**Frequency (%)**	**No**	**Yes**
	No or N.A.	87.21	99.12	0.88
	Yes	12.79	99.82	0.18
Has health insurance	**Answer**	**Frequency (%)**	**No**	**Yes**
	No	78.82	99.38	0.62
	Yes	21.18	98.58	1.42
Skin color/ethnicity	**Skin color/ethnicity**	**Frequency (%)**	**No**	**Yes**
	Yellow	0.73	98.72	1.28
	White	36.14	98.89	1.11
	Unknown	0.02	100.00	0.00
	Indigenous	0.78	100.00	0.00
	Brown	50.51	99.45	0.55
	Black	11.82	99.11	0.89

to be continued

Table 1. continuation

Descriptive variable			Prostate cancer – frequency in category (%)	
Education level	**Education level**	**Frequency (%)**	**No**	**Yes**
	No education or incomplete elementary school	42.67	99.03	0.97
	Complete elementary school or incomplete high school	15.48	99.41	0.59
	Complete high school or incomplete higher education	28.93	99.43	0.57
	Complete higher education	12.92	99.06	0.94
Vegetable and legume intake	**Weekly intake**	**Frequency (%)**	**No**	**Yes**
	1 to 3 days	35.67	99.34	0.66
	4 to 6 days	21.24	99.21	0.79
	Never or too little	9.54	99.46	0.54
	Every day	33.55	98.99	1.01
Currently smokes	**Weekly use**	**Frequency (%)**	**No**	**Yes**
	Daily	14.32	99.45	0.55
	Less than daily	1.75	99.73	0.27
	Doesn't smoke	83.93	99.16	0.84
Diagnosed with depression	**Answer**	**Frequency (%)**	**No**	**Yes**
	No	95.43	99.24	0.76
	Yes	4.57	98.46	1.54

[a] According to the National Health Survey (NHS), this variable refers to the intake of the so-called "boxed", canned or powdered juices.
[b] According to the PNS, this variable refers to the handling of chemical products such as: pesticides, gasoline, diesel, formalin, lead, mercury, chrome, chemotherapeutics etc.

family income groups (in minimum wages (MW)): up to ½ MW; from ½ MW up to 2 MW; from 2 MW to 5 MW; over 5 MW. Thus, the higher income group (over 5 MW) has a DPC rate ≈ 3.7 × greater than the lower income group (up to ½ MW).

Based on the premise above, the possible socioeconomic effect on the descriptive variables was investigated; so, an association level was determined between each variable described and the *per capita* family income through the Cramer's V and Cohen's ω measures. A high level of association between the *per capita* family income and the "health insurance" (Cramer's V = 0.487 and Cohen's ω = 0.487) and "education level" (Cramer's V = 0.314 and Cohen's ω = 0.544) variables was verified. A medium level of association with the *per capita* family income was observed in the variable "skin color/ethnicity" (Cramer's V = 0.150 and Cohen's ω = 0.260), and a reasonable level was observed in "fruit intake" and "vegetable and legume intake" (Cohen's 7 = 0.218 and 0.299, respectively). Therefore, the influence of family income in the described variables should be considered in the interpretation of results from Table 1.

The results of accuracy, sensibility, specificity, and ROC-AUC for logistic regression are shown in Table 2, referring to the training and test bases. Accuracy, sensibility, and specificity were observed to be very satisfactory for these bases, being all above 80%. In the test base, there was a slight increase in accuracy and specificity, and a slight decrease in sensibility. The obtained result for ROC-AUC in the training base (0.822) can also be considered very satisfactory, while the base result (0.780) was found to be very near the same condition.

Table 2. Accuracy, sensibility, specificity, and ROC-AUC obtained in the models of logistic regression, 2-level decision tree and 3-level decision tree, considering the training (70% of the total) and test (30% of the total) bases

Model	Base	Accuracy	Sensibility	Specificity	ROC-AUC
Logistic regression	Training	0.828	0.828	0.828	0.822
	Test	0.835	0.802	0.835	0.780
Decision tree (2 levels)	Training	0.801	0.850	0.800	0.823
	Test	0.805	0.821	0.804	0.812
Decision tree (3 levels)	Training	0.767	0.906	0.766	0.832
	Test	0.773	0.887	0.772	0.813

The results obtained for the 2 and 3-level decision trees are graphically shown in Figures 2 and 3, respectively, referring to the training bases. From the observation of frequency of prostate cancer cases in the **2-level** tree's final nodes, it can be verified that:

- **Nodes 3 and 4** suggest that there is a greater probability of prostate cancer occurring in black men in comparison to other ethnicities, considering the below 50 age group;
- **Nodes 6 and 7** indicate that the group of men with high cholesterol levels have a greater DPC frequency (≈ 3× greater) in comparison to the group of normal or unknown cholesterol levels, considering the 50 to 65 age group;
- **Nodes 9 and 10** indicate that the group of men with education levels between complete elementary school and complete higher education have a greater DPC frequency (≈ 2× greater) in comparison to the group of lower education levels, considering the 65 to 80 age group;
- **Nodes 12 and 13** indicate that the group of men with health insurance have a greater DPC frequency (≈ 2× greater) in comparison to the group of men who don't have insurance, considering the 80 and over age group.

Regarding node 5 (which results in final nodes 6 and 7), it is possible to assess the association level between the "high cholesterol" risk factor and the DPC outcome for the subgroup of men aged between 50 and 65, using the odds ratio (OR). In this case, an OR = 2.87 was obtained (confidence interval – CI 95%: 1.58 – 5.21), representing a significant level of association.

When interpreting the results from nodes 9 and 10, the previously discussed strong association between education level and *per capita* family income should be considered[18]. Regarding nodes 12 and 13, the different characteristics of the public and private health systems in Brazil should be

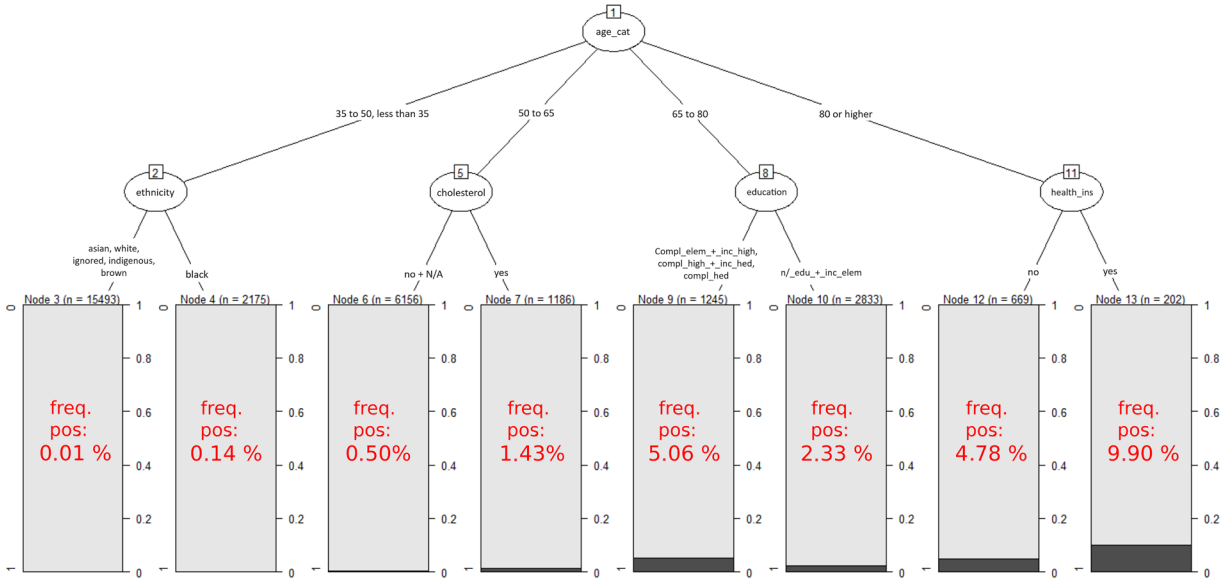

Figure 2. Two-level decision tree, considering the training base with 70% of the selected individuals. Percentages in red indicate the frequency of positive prostate cancer cases at the end of the tree

Captions: *age_cat* = age in categorical form (years); *ethnicity* = skin color/ethnicity; *cholesterol* = high cholesterol level; *education* = education level; *health_ins* = health insurance; *n_a* = not applied question; *compl_elem* = complete elementary school; *inc_high* = incomplete high school; *comp_high* = complete high school; *inc_hed* = incomplete higher education; *comp_hed* = complete higher education; *n/_edu* = no education; and *inc_elem* = incomplete elementary school.

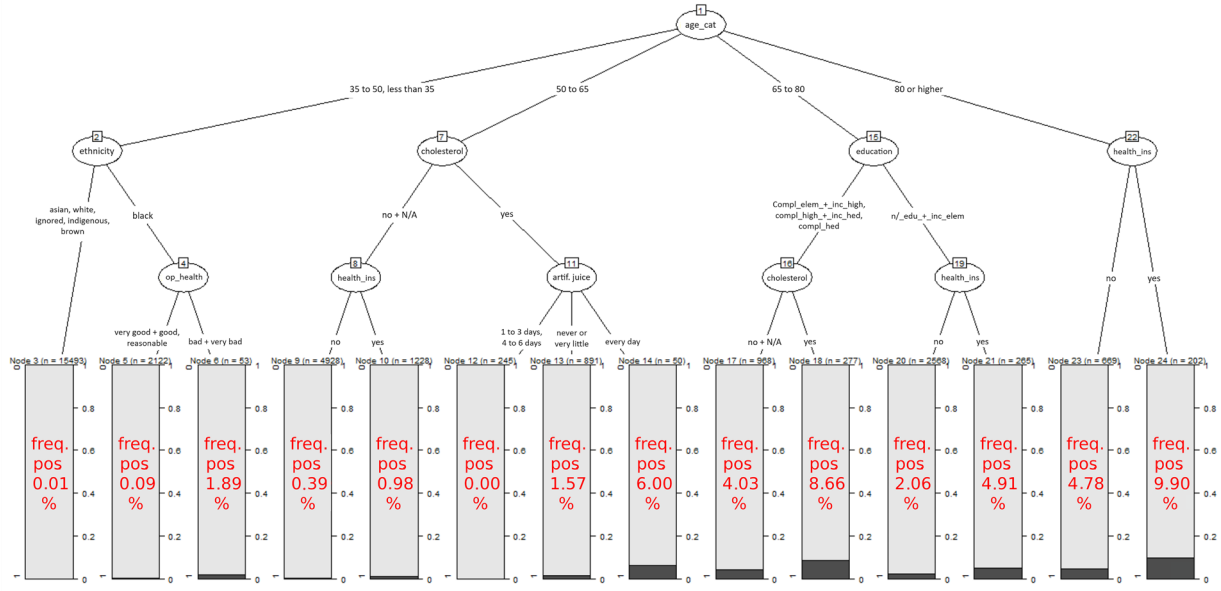

Figure 3. Three-level decision tree, considering the training base with 70% of the selected individuals. Percentages in red indicate the frequency of positive prostate cancer cases at the end of the tree

Captions: *age_cat* = age in categorical form (years); *ethnicity* = skin color/ethnicity; *cholesterol* = high cholesterol level; *education* = education level; *health_ins* = health insurance; *op_health* = opinion on their own health; *artf_juice* = weekly intake of artificial juice; *n_a* = not applied question; *compl_elem* = complete elementary school; *inc_high* = incomplete high school; *comp_high* = complete high school; *inc_hed* = incomplete higher education; *comp_hed* = complete higher education; *n/_edu* = no education; *inc_elem* = incomplete elementary school; and t*oo_ltl* = too little.

considered, in addition to the previously discussed strong association between the "health insurance" variable and the *per capita* family income[19].

The main results from the **3-level** decision tree are described as follows:

- Among men aged 50 to 65 years old, who present normal or unknown cholesterol levels, the DPC frequency was ≈ 2,5× greater in the subgroup who owns health insurance in comparison to the subgroup who doesn't own health insurance.
- Among men aged 50 to 65 years old, who present high cholesterol levels, the DPC frequency was ≈ 4× greater in the subgroup of men who drink artificial juice every day in comparison to the subgroup who claimed not to drink or to drink too little artificial juice.
- Among men aged 65 to 80 years old, who have at least completed elementary school, the DPC frequency was ≈ 2× greater in the subgroup of men who present high cholesterol levels in comparison to the subgroup who present normal or unknown levels of cholesterol.
- Among men aged 65 to 80 years old, who have an education level below complete elementary school, the DPC frequency was ≈ 2.5× greater in the subgroup of men who have health insurance in comparison to the subgroup of men who don't have health insurance.

The results of the 3-level decision tree reinforce aspects discussed on the 2-level decision tree, though in more specific groups. For instance, the {9,10}, {20,21} and {23,24} node pairs reinforce the probability of DPC for individuals who claim to have health insurance. The {17,18} node pair reinforces the greater probability of DPC for individuals with high cholesterol levels, yet in the age group of 65 to 80 years old who have at least completed elementary school. Regarding node 16 (which results in final nodes 17 and 18), it is possible to assess the association level between the "high cholesterol" risk factor and the DPC outcome for this subgroup. In this case, an OR = 2.26 (CI 95%: 1.33 – 3.83) was obtained, representing a significant level of association.

A relevant information is shown in the {12,13,14} node trio, which indicates a significantly greater probability of DPC for individuals who claim to drink artificial/industrialized juices every day, within the age group of 50 to 65 years old who show high cholesterol levels. However, such a result must be carefully analyzed, as will be discussed in the following section.

DISCUSSION

The median age of men who answered ("yes") to having received DPC was 72 years old, a value well above the median of men who answered "no" (44 years old), result that is compatible with the general tendency of men developing prostate cancer approximately from the age of 50[3].

Nodes 3 and 4 in the 2-level tree suggest that black people have a greater tendency for developing prostate cancer before 50 years old; such a fact seems to corroborate prior studies about prostate cancer[3], in which this disease

was verified to be more frequent in people with African and Caribbean ancestry. However, prostate cancer cases in ages younger than 50 years old are rare at the studied base (2 cases in 15,493 individuals in node 3, 3 cases in 2,175 individuals in node 4), and thus, it is suggested that conclusions regarding this age group should be corroborated by a larger database of individuals younger than 50 years old.

Regarding nodes 6 and 7 in the 2-level tree, it is important to consider studies such as Pelton et al.[20], which claims that high cholesterol levels in the blood is related to more aggressive prostate cancer cases, and Jamnagerwalla et al.[21], which shows that high levels of total serum cholesterol and HDL are associated to an increased risk of a high degree prostate cancer. A communication from Johns Hopkins Medicine[22] describes more recent studies that point to similar conclusions. As the present study is cross-sectional, it cannot be said with certainty that it corroborates the results of previous works but indicates the importance of additional studies on the cholesterol/DPC relationship including individuals from Brazil.

The cross-sectional design of the present study limits the possibility of conclusions that involve the causal relationship due to the lack of temporality control. Another limitation is the use of secondary data, which hinders the precision of the interviewees' answers regarding some variables; a clear example in this aspect is the weekly intake of artificial juices. However, previous studies suggest the relationship between high intake of sugary drinks and a greater incidence of prostate cancer, such as Miles et al.[23] and Llaha et al.[24]. Moreover, the Makarem et al.[25] study about sugary food intake suggests an increase in the risk of developing prostate cancer in men who drink fruit juice more frequently. The results of this research indicate the importance of complementary studies on the influence of sugary drinks, artificial, and industrialized juices intake in the DPC rate in the Brazilian context. Therefore, longitudinal studies will be useful to analyzing the possible effect of factors such as high cholesterol level and artificial and industrialized juice intake in the rate of prostate cancer occurrence.

The decision trees were used as prediction models for prostate cancer cases. The training base was used in the model's parameterization, which was later applied to the test base. The results of accuracy, sensibility, specificity, and ROC-AUC for the 2 and 3 level decision trees are shown in Table 2, referring to the training and test bases. The 2-level model performed well in both the training and test bases. There is a slight tendency for a better reproduction of positive prostate cancer events (due to sensibility) in relation to negative events (due to specificity). Regarding the 3-level tree, there was a considerable increase in sensibility in relation to the 2-level tree, a very satisfactory result in the aspect of reproducing positive prostate cancer cases in the training and test bases; however, the 3-level tree is slightly less efficient in reproducing negative cases, which can be verified by the slight decrease in specificity.

To complement the results from Table 2, the positive predictive value (PPV) was calculated for the training and test base models, and values between 3% and 4% were obtained. Such results are expected, as the analyzed base of 42,799 men contains over 99% of individuals who claimed to "not having received" a DPC. Due to the great predominance of negative cases in the base, the models are expected to generate, in absolute terms, a high number of false-positive cases (about a few thousand), which has a relatively mild effect on specificity and accuracy, but strongly reduces PPV. Such a predictive limitation, however, does not belittle the successful results presented in the other metrics.

In principle, the models presented could be used to identify men with physical, socioeconomic, and lifestyle characteristics that make them more likely to receive DPC. However, it must be noted that there is an important difference between developing an illness and being diagnosed with it. There are socioeconomic variables such as family income, having or not having health insurance, education level etc. that may influence obtaining a precocious prostate cancer diagnosis. That is, men in deficient socioeconomic situations (low income, low education level etc.) are more prone to being underdiagnosed. If the models presented here are applied with the restricted goal of identifying men with a tendency to develop prostate cancer, then it should be noted that the model will naturally have limitations, depending on the analyzed socioeconomic group.

This study' results can be more deeply investigated in future research by including the influence of certain food intake in the occurrence of prostate cancer in longitudinal studies, the statistical differences among social groups in the disease diagnosis, the relationship between prostate cancer and other diseases, among other aspects. The crossing of the dependent variable DPC with descriptive variables does not always allow for a clear interpretation of the results, since some variables may suffer a significant influence of socioeconomic factors, as is described in the exploratory analysis. Moreover, NHS questions that correspond to some descriptive variables were not answered by the totality of individuals (the N.A. acronym was used for questions that were not applied).

It should also be considered that the NHS questionnaire was not specifically planned for a profound analysis of prostate cancer. An important question not included in the questionnaire refers to the prostate cancer cases in

family members. Studies in the field[3] show that having a first degree relative with DPC significantly increases the risk of a man developing the disease. In general, the NHS questionnaire asked questions that did not clarify the life habits and health state of the interviewee throughout their life, that is, most questions only frame the interviewee at the time of that survey. Therefore, the models presented here may be improved in the future by using a historical database of the interviewee and their family, with questions related to the evolution of their health and life habits.

The exploratory analysis of this database and the developed models can be used for estimating the public or private health system demands, such as human resources and infrastructure, prevention, and treatment of prostate cancer in specific groups or regions of Brazil, as well as identifying groups of people as the preferential targets of prostate cancer prevention campaigns.

CONCLUSION

The machine learning models applied to the NHS data indicate a significant association between socioeconomic, physical, and behavioral factors and diagnosis of prostate cancer (DPC) in Brazil. The decision tree models show that the variables "age", "diagnosis of high cholesterol level", "having health insurance" and "education level" are strongly associated to the DPC rate in different groups of people. The models' high level of accuracy (near or above 80%) and sensibility (between 80% and 90%) show the potential of machine learning methods for study and prevention of prostate cancer in the Brazilian context, particularly if longitudinal databases on the disease are available in the future.

This research shows useful results for planning the use of public or private resources in the treatment or prevention of prostate cancer in the Brazilian context, as well as for directing future studies on the disease.

CONTRIBUTIONS

Marco Antonio de Souza has substantially contributed to the study design, acquisition, analysis and interpretation of the data, wording, and critical review. Camila Nascimento Monteiro and Cláudia Renata dos Santos Barros have contributed to the analysis and interpretation of the data, wording, and critical review. All the authors approved the final version for publication.

DECLARATION OF CONFLICT OF INTERESTS

There is no conflict of interest to declare.

FUNDING SOURCES

None.

REFERENCES

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

Associate-editor: Jeane Tomazelli. Orcid iD: https://orcid.org/0000-0002-2472-3444
Scientific-editor: Anke Bergmann. Orcid iD: https://orcid.org/0000-0002-1972-8777