UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA EM SAÚDE
MESTRADO EM CIÊNCIA E TECNOLOGIA EM SAÚDE

GUSTAVO DIAS DA SILVA

UTILIZAÇÃO DO ALGORITMO APRIORI PARA TRAÇAR O PERFIL SOCIODEMOGRÁFICO DO HOMEM BRASILEIRO COM CÂNCER DE PRÓSTATA

CAMPINA GRANDE
2022

GUSTAVO DIAS DA SILVA

UTILIZAÇÃO DO ALGORITMO APRIORI PARA TRAÇAR O PERFIL SOCIODEMOGRÁFICO DO HOMEM BRASILEIRO COM CÂNCER DE PRÓSTATA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como parte dos requisitos para obtenção do Título de Mestre em Tecnologias em Saúde.

Área de concentração: Saúde Coletiva: Saúde Pública.

Orientador: Prof. Dr. Wellington Candeia de Araújo

CAMPINA GRANDE
2022

S586u Silva, Gustavo Dias da.

Utilização do algoritmo Apriori para traçar o perfil sociodemográfico do homem brasileiro com câncer de próstata [manuscrito] / Gustavo Dias da Silva. - 2022.

85 p. : il. colorido.

Digitado.

Dissertação (Mestrado em Profissional em Ciência e Tecnologia em Saúde) - Universidade Estadual da Paraíba, Pró-Reitoria de Pós-Graduação e Pesquisa , 2022.

"Orientação : Prof. Dr. Wellington Candeia de Araújo , Coordenação do Curso de Computação - CCT."

1. Algoritmo Apriori. 2. Câncer de próstata. 3. Mineração de dados. 4. Regras de associação. I. Título

21. ed. CDD 006.31

Elaborada por Deivid A. M. Neves - CRB - 15/899 BC/UEPB

GUSTAVO DIAS DA SILVA

UTILIZAÇÃO DO ALGORITMO APRIORI PARA TRAÇAR O PERFIL SOCIODEMOGRÁFICO DO HOMEM BRASILEIRO COM CÂNCER DE PRÓSTATA

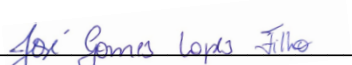
Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como requisito para obtenção do título de Mestre em Ciência e Tecnologia em Saúde.

Dissertação aprovada em: 09/03/2022

BANCA EXAMINADORA:

Prof. Dr. Wellington Candeia de Araújo
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Frederico Moreira Bublitz
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. José Gomes Lopes Filho
Parque Tecnológico Itaipu (PTI-BR)

Dedico este trabalho primeiramente a Deus; sem ele eu não teria capacidade para desenvolver este trabalho. Ademais, a todos aqueles que neste trabalho puderam ajudar**.**

AGRADECIMENTOS

Agradeço ao meu orientador: seu apoio, orientação e ideias que fizeram desta uma experiência inspiradora para mim. Além disso, por realizar este sonho que, com todas as dificuldades que surgiram, desde o dia que resolvi deixar tudo e ingressar nessa jornada – que é o mestrado –, consegui concluir.

A minha esposa, por me apoiar durante todo o período de estudo até a finalização desta dissertação.

A todos os amigos que me apoiaram e me incentivaram a realizar este sonho.

“Se você torturar os dados por tempo suficiente, eles irão confessar. ” (Ronald H. Coase)

RESUMO

Entre as doenças que acometem a população masculina, o câncer de próstata é a que tem aumentado a taxa de mortalidade, pois, no mundo, é a sexta neoplasia maligna e, no Brasil, a primeira. Ele passa a ser um caso de saúde pública que preocupa, porém, se descoberto no início, o índice de cura é alto. Apesar das inciativas de ajuda à população masculina contra a neoplasia prostática, ainda falta um direcionamento quanto ao diagnóstico e tratamento. Mas as iniciativas seriam mais bem direcionadas se tivessem os perfis dos pacientes assistidos por elas, no entanto, esse ainda é um campo de pesquisa com lacunas. Além disso, dados que possam ajudar se encontram armazenados em grandes bases de dados com muitas informações, principalmente devido ao processo de informatização do setor de saúde, que dificulta uma análise manual desses dados. Nesse sentido, este trabalho tem o propósito de determinar o perfil do homem que é propenso ao câncer de próstata através do algoritmo Apriori para a formação de regras de associação no contexto brasileiro. Com isso, aplicamos na base de dados do INCA o algoritmo Apriori com a finalidade de termos as regras de associação. Ao final, percebemos que os fatores de tabagismo, alcoolismo, raça e estado conjugal são os fatores que mais destacaram por aparecerem nas regras com os maiores índices de confiança. Entretanto, depreendemos que a raça parda é de maior incidência do câncer de próstata no Brasil. Apesar da incompletude dos dados opcionais na base do INCA, é importante destacar que a análise foi feita a nível nacional e pode ser utilizada para nortear campanhas no contexto da saúde do homem.

Palavras-Chave: Câncer de Próstata. Mineração de Dados. Regras de Associação. Algoritmo Apriori.

ABSTRACT

Among the diseases that affect the male population, prostate cancer has increased the mortality rate among them, where it is the sixth malignant neoplasm in the world and in Brazil the first. It becomes a matter of public health that concerns, but if discovered early, the cure rate is high. Despite the initiatives to help the male population against prostate cancer, there is still a lack of guidance regarding diagnosis and treatment. But the initiatives would be better targeted if they had the profiles of patients assisted by them, but it is still a field of research with gaps. In addition, data that can help are stored in large databases with a lot of information, mainly due to the computerization process of the health sector, which makes manual analysis of this data difficult. This work aims to determine the profile of men who are prone to prostate cancer through the Apriori algorithm for the formation of association rules in the Brazilian context. With that, we applied the Apriori algorithm to the INCA database in order to have the association rules. In the end, we realized that the factors of smoking, alcoholism, race and marital status are the factors that stood out the most for appearing in the rules with the highest levels of confidence. However, we infer that the brown race has a higher incidence of prostate cancer in Brazil. Despite the incompleteness of the optional data in the INCA database, it is important to highlight that the analysis was carried out at the national level and can be used to guide campaigns in the context of men's health.

Keywords: Prostate Cancer. Data Mining. Association Rules. Apriori Algorithm.

LISTA DE ILUSTRAÇÕES

Figura 1 - Representação do processo de KDD 19
Figura 2 - Software Weka 26
Figura 3 - Portal integrador RHC 35
Figura 4 - Download de dados do portal integrador RHC 35
Figura 5 - *TabWin* 36
Figura 6 - *Spyder* 37
Figura 7 - Código de conversão do atributo numérico raça para categórico 38

LISTA DE GRÁFICOS

Gráfico 1 – Taxas de mortalidade das 5 localizações primarias mais frequentes, ajustadas por idade, pela população mundial, por 100 mil homens, Brasil, entre 2009 e 2019. 29
Gráfico 2 – Quantitativo de respostas para o quesito sobre o consumo de álcool. 43
Gráfico 3 – Quantitativo de respostas para o quesito sobre o consumo de tabaco. 45
Gráfico 4 – Quantitativo de respostas para o quesito sobre caso de câncer na família. 46
Gráfico 5 – Quantitativo de respostas para o quesito sobre estado conjugal. 47
Gráfico 6 – Histograma das idades dos homens propensos ao câncer de próstata no período de 2010 e 2019 48
Gráfico 7 – Mapa de calor das correlações entre os atributos idade, raça, histórico familiar, alcoolismo e tabagismo. 49
Gráfico 8 – Comparativo entre as raças dos homens com câncer de próstata. 50
Gráfico 9 – População residente, por cor ou raça 51
Gráfico 10 – Distribuição das pessoas de 25 anos ou mais de idade, segundo o nível de instrução 52
Gráfico 11 – Comparativo entre os graus de instrução dos homens com câncer de próstata ... 53
Gráfico 12 – Distribuição das ocorrências de casos de próstata no Brasil no período de 2010 a 2019 54

LISTA DE TABELAS

Tabela 1- Taxas de mortalidade por câncer de próstata, brutas e ajustadas por idade, pelas populações mundial e brasileira de 2010, por 100 mil homens, no Brasil entre 2009 e 2019 . 30
Tabela 2 – Computador disponível..... 37
Tabela 3 – Proporcionalidade dos casos de câncer de e a população masculina por faixa etária 48
Tabela 4 – Proporcionalidade dos casos de câncer de próstata e a população masculina por Estado 55
Tabela 5 – Regras de associação com dados do período de 2010 a 2019 58

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Problema 14
1.2 Limitação das abordagens atuais 15
1.3 Solução proposta 16
1.4 Objetivos 18
1.4.1 Objetivo geral 18
1.4.2 Objetivos específicos 18
2 REVISÃO DE LITERATURA 19
2.2 Mineração de dados 19
2.3 Mineração de dados na saúde 25
2.4 Câncer de próstata 27
2.5 Trabalhos relacionados 31
3 METODOLOGIA 33
3.1 Fonte de dados 34
3.2 Recursos de softwares 36
3.3 Transformação dos dados 38
3.4 Regras de associação 38
3.4.1 Definição de parâmetros do algoritmo apriori 39
3.5 Seleção dos atributos 40
3.6 Pós-processamento dos dados 41
4 RESULTADOS E DISCUSSÕES 42
5 CONCLUSÕES 61
5.1 Dificuldades encontradas 62
5.2 Trabalhos futuros 62
REFERÊNCIAS 63
APÊNDICE A – PROTOCOLO DE REVISÃO SISTEMÁTICA APLICADA AO ESTUDO 68
ANEXO A – DICIONÁRIO DA BASE DE DADOS DO INCA 82

1 INTRODUÇÃO

A saúde é um bem a que todos têm direito e a Carta Magna Brasileira garante isso. Porém, muitas vezes, as políticas assistencialistas que objetivam levar a uma melhor qualidade de vida não chegam às classes menos favorecidas de conhecimento e recursos financeiros, precisando, inúmeras vezes, de que o governo procure meios alternativos para isso.

O pouco conhecimento da população pode dificultar a difusão da assistência, deixando esse público carente de algumas políticas sociais. Além disso, os tratamentos e diagnósticos de doenças ficam prejudicados quando o cenário do sistema de saúde é desigual em se tratando do acesso (SACRAMENTO et al., 2019).

Dentre tantas doenças que acometem a população, destaca-se o câncer, já que esse atinge o indivíduo tanto em termos físicos quanto em termos psicológicos. E os números sobre essa doença tendem a aumentos consideráveis nos anos seguintes (AGUIAR; PEREIRA; THOMAS, 2017; FERRÃO; BETTINELLI; PORTELA, 2017; INCA, 2018).

Se considerar a população masculina, o câncer de próstata tem se apresentado como destaque tanto a nível mundial quanto brasileiro, haja vista que essa doença tem um índice de mortalidade alto, sendo considerado o segundo câncer que mais mata entre os homens, ficando atrás apenas do câncer de pulmão (BARROS, 2019; REGO et al., 2020).

A neoplasia prostática acomete homens com histórico familiar de câncer de próstata, hábitos alimentares, atividades sexuais, consumo de álcool, idade entre outros (HUSSAIN et al., 2019; REGO et al., 2020). Com o destaque para idade, pois é quase unânime que o câncer de próstata ocorre em homens com idades acima de 50 anos, e sua ocorrência em idades inferiores é considerado caso raro (REGO et al., 2020).

Além disso, estudos demonstram que há um índice considerável de diagnóstico de câncer de próstata em homens afrodescendentes, por exemplo, chegando a valores de 127 para cada 100 mil homens diagnosticados na Nigéria, enquanto que, na América do Sul, esse índice não passa de 61 para cada 100 mil (CASSELL et al., 2019; MOTA; BARROS, 2019).

Mundialmente, a Organização Mundial de Saúde (OMS) considera o câncer de próstata como um problema de saúde pública (GONÇALVES et al., 2008), haja vista que é um problema que assola uma grande parcela da população masculina, deste modo existe a necessidade de políticas de ajuda tanto para o diagnóstico, quanto para o tratamento (INCA, 2018).

No Brasil, os números são preocupantes, como são demonstrados pelo boletim da vigilância e Análise de situação, que lançou, em sua sétima edição, os números de diagnósticos de câncer do período de 2012 a 2016, em que o câncer de próstata figurou em primeiro lugar sendo responsável por 25% dos casos (INCA, 2020).

O aumento nos números é justificado, em muitos casos, pela evolução do diagnóstico, pelas as políticas de informação, e pelo aumento da expectativa de vida. Segundo Menezes et al. (2019), há uma política de atenção à saúde do homem promovido pelo Ministério da Saúde brasileiro, que objetiva promover a saúde e prevenir doenças, isso leva a um aumento na expectativa de vida do homem.

Esse câncer é uma doença que, de imediato, não apresenta sintomas perceptíveis para o paciente, pois, incialmente, ele evolui de forma silenciosa, contudo, na sua forma agressiva, chega a provocar infecção generalizada ou insuficiência renal (REGO et al., 2020).

Para o diagnóstico do câncer de próstata, há duas formas, uma que passa pelo exame de toque retal e outra pelo exame de Antígeno Específico Prostático (PSA). No primeiro, o homem é submetido ao toque para verificar e avaliar a presença de nódulos na próstata. Contudo, segundo Rego et al. (2020), é um exame que deixa muito a desejar, pois não consegue tocar toda a próstata. E o segundo método é um dos mais usados para diagnosticar o câncer nos seus estágios iniciais, ele tem o propósito de verificar o nível de uma proteína originária na próstata.

No mundo, o PSA é um exame bastante utilizado para detecção do câncer de próstata, por exemplo, nos Estados Unidos, o teste é utilizado principalmente para detectar a presença de tumor nos estágios iniciais do câncer. Além disso, ele é considerado como importante meio de diagnóstico de doenças prostáticas, inclusive o câncer (HUSSAIN et al., 2019; REGO et al., 2020).

Independentemente de o câncer de próstata ser uma doença que afeta uma grande parcela da população masculina, e como estudos demonstram os negros, principalmente, há um grande preconceito no Brasil na realização de exames para diagnóstico, devido às questões culturais. Segundo Menezes et al. (2019), a questão cultural barra o homem na procura da sua saúde, dificultando assim a identificação de pacientes propensos a este tipo de câncer.

É importante, ainda, frisar que o diagnóstico tem que se demonstrar eficaz e seguro, pois o tratamento pode causar disfunção erétil, incontinência urinária e sintomas intestinais, assim pode causar desconforto ao paciente que poderia estar com uma neoplasia de baixa agressividade (FERRÃO; BETTINELLI; PORTELA, 2017; SBU; SBPC/ML, 2018). E a falta

de consideração de fatores como idade, raça e histórico familiar pode vir a acarretar erros (REGO et al., 2020).

1.1 Problema

O Ministério da Saúde (doravante MS) brasileiro não recomenda o rastreamento populacional para o câncer de próstata, pois há riscos e limitações dos exames (REGO et al., 2020). Mas, aconselha que os homens procurem as unidades de saúde e submetam-se aos exames, uma vez que o MS recomenda que a procura seja por livre demanda ou por rastreamento oportunista. Contudo, o homem deve ser alertado sobre as possíveis complicações caso seja necessário tratamento (REGO et al., 2020).

A Sociedade Brasileira de Urologia recomenda que homens acima de 50 anos devem procurar um profissional especializado para avaliação individual, e, nos casos de serem negros ou com parentes de primeiro grau com casos de câncer de próstata, essa procura deve ocorrer a partir dos 45 anos (SBU; SBPC/ML, 2018).

Sendo o câncer de próstata uma neoplasia que ataca toda a população masculina e o número de casos vem aumentando com o tempo em vários locais do mundo, a identificação e a triagem dos pacientes é um grande desafio para as nações (CASSELL et al., 2019; REGO et al., 2020).

Mas, apesar de políticas como a lei brasileira 10.289/2001, que vislumbram atender esta demanda de casos e capacitar os profissionais envolvidos no atendimento e prevenção do câncer de próstata, o primeiro passo será a identificação dos casos, que poderia ser feita através da definição do perfil do homem com o câncer de próstata.

No caso do câncer de próstata, de acordo com Mota e Barros (2019), é de suma importância a definição do perfil do paciente. Essa definição do perfil do paciente que possui o câncer pode facilitar o diagnóstico, o encaminhamento de exames e o tratamento adequado. Ademais, segundo Araújo et al. (2015), o campo da pesquisa que define o perfil dos pacientes ainda precisa ser desenvolvido.

Atualmente, as informações que podem ser utilizadas para formar o perfil dos pacientes estão sendo armazenadas em base de dados. E as novas tecnologias têm produzido grande quantidade de dados nas mais diversas áreas, no campo da saúde, não é diferente (OLIVEIRA, et al., 2007).

Ademais, se for necessária uma análise das informações armazenadas, de forma manual, é impossível, pois esbarra no cálculo e associação de informação de grandes

quantidades de dados (PREISSLER, 2016). Logo, os dados da saúde para serem estudados necessitam de um processo de análise automatizada.

1.2 Limitação das Abordagens Atuais

Considerando a quantidade de dados e o território brasileiro com suas dimensões continentais, a realização de um estudo para formulação de um perfil do homem com o câncer de próstata se torna difícil e implicará em um estudo com resultado custoso para sua realização manual.

Contudo, algumas ações vêm acontecendo no território nacional, por exemplo, os estudos de Araújo et al. (2015), em que os autores caracterizaram o perfil dos pacientes diagnosticados com câncer de próstata em um hospital em Ribeirão Preto no estado de São Paulo. Neste estudo, os autores analisaram os dados através da estatística descritiva.

Utilizando o *Microsoft Office Excel*, Matos e Barros (2019) realizaram o levantamento do perfil dos pacientes diagnosticados com câncer de próstata no hospital de câncer de Pernambuco, situado na cidade de Recife. Contudo, a análise ficou restrita ao percentual e à frequência, que novamente é um exemplo de outra análise utilizando estatística descritiva.

No caso de Rego et al. (2020), em que se fez o uso de estatística descritiva, gráficos e usando o software SPSS 23.0, realizou-se a análise de dados obtidos no 9º Mutirão de Prevenção ao Câncer no município de Montes Claros no estado de Minas Gerais. Essa análise ficou restrita à frequência absoluta e à frequência relativa dos dados obtidos durante o evento.

Contudo, há outras experiências, como a relatada por Hussain et al. (2019), o qual aplicou Redes Bayesiana para quantificar a associação entre fatores morfológicos presentes em exames de imagem de pacientes com suspeitas de câncer de próstata. Tal procedimento foi feito com o objetivo de treinar essa rede Bayesiana para verificar a força de relacionamento entre os fatores.

Mas, o trabalho de Hussain et al. (2019) aplicou a rede Bayesiana, mostrando um avanço com o uso de ferramenta computacional, porém ficando interessado na análise de dados de imagens e não se preocupou em analisar a situação socioeconômica do paciente para formação do perfil do paciente.

Assim, percebe-se que, quando se trata de formar o perfil do paciente com câncer de próstata no campo brasileiro, ainda há muitas restrições e poucas ações. E, quando ocorrem ações, há o destaque do uso da estatística descritiva, sendo impossível aplicar tal tática em uma base com milhares de dados para serem levantados e analisados.

1.3 Solução proposta

Segundo Furlan (2018), a mineração de dados tem o objetivo de reunir padrões, previsões, associações e erros para descobrir conhecimento escondido em grandes quantidades de dados. Assim, para a situação da saúde, essa é uma tática considerável, haja vista a grande produção de dados criados nos últimos anos com o processo de informatização do setor.

De acordo com Araújo (2007), a mineração de dados possibilita decisões governamentais melhores. Desta forma, a mineração pode auxiliar na leitura destas bases de dados da saúde, e permitir, com as suas tarefas e ferramentas, a compreensão dos dados e facilitação na hora de tomar uma decisão quando se tratar de políticas de auxílio à população na área da saúde.

Ao aplicar a mineração de dados nas bases da saúde, tal procedimento permite verificar as necessidades da população, como também fazer o levantamento de registros que possibilitem a formação do perfil dos homens com o câncer de próstata. Isso é possível, pois os dados necessários vêm do processo de informatização, que viabiliza o levantamento de informações de cada região e seus habitantes.

Melhorar a qualidade de vida da população, organizar os serviços de saúde e aprimorar os existentes são uns dos benefícios que a mineração de dados pode proporcionar quando aplicada nas bases da saúde. Em FEUSER, 2017, há a demonstração da aplicação de mineração de dados em vários pontos da saúde, e os resultados são positivos, por exemplo, a automatização e melhoramento do suporte aos pacientes com a doença de chagas em Bambuí em Minas Gerais, entre outros trabalhos, que são relacionados em sua pesquisa.

Mas, deve-se lembrar de que a mineração de dados faz parte de um processo de Descoberta de Conhecimento em Base de Dados ou *Knowledge Discovery in Databases* (KDD), o qual é composto por três etapas: pré-processamento, mineração de dados e pós-processamento. Portanto, para realizar e obter os resultados esperados, ainda executam-se mais duas etapas.

No pré-processamento, na base de dados, serão analisados os dados a fim de se verificar as inconsistências, como exemplo, campos vazios, valores incoerentes para os campos e valores fora dos limites para o preenchimento. Para assim eliminar qualquer viés que possa influenciar o resultado final.

No pós-processamento, serão recolhidas as regras de associação, que representam a ligação entre os itens, para definir se um item aparece ou não conforme com a presença do

outro de acordo com a manipulação feita pelo algoritmo de mineração de dados, a partir dos restantes do processo de pré-processamento.

Portanto, dentre as táticas presentes na mineração de dados, a associação será a tarefa adotada, por facilitar o entendimento dos dados presentes em grandes bases de dados, pois conforme Silva et al. (2016), ela trabalha com regras que são de fácil interpretação e podem ser traduzidas até para linguagem natural.

Para a formação das regras de associação, é necessário a definição de um algoritmo com tal função. Neste trabalho, far-se-á uso do algoritmo *Apriori*, uma vez que Feuser (2017) destaca como sendo um algoritmo utilizado para medir o grau de confiança da regra de associação. Ademais, segundo Silva (2016, p.341), com o algoritmo *Apriori* é possível montar uma estratégia de descoberta dos itens frequentes de forma simples e eficiente.

Além disso, o algoritmo *Apriori* é de fácil utilização e experimentação e pode ser utilizado em grandes conjuntos de dados como fica destacado no trabalho de EFFIOK; LIU; HITCHCOCK (2017), que comparou esse com os algoritmos *FP-GROWTH e Dynamic Itemset Counting*.

O algoritmo *Apriori* será aplicado na base do Instituto Nacional do Câncer José Alencar Gomes da Silva (INCA), que possui os dados dos pacientes em tratamento de câncer no Brasil. A partir da análise dessa base, serão filtradas as regras de associações para a formação do perfil de pacientes com câncer de próstata a fim de montar o mapeamento do brasileiro com o câncer de próstata.

Ademais, para a validação das regras geradas, ao final, o algoritmo *Apriori* tem utilizado, durante o processamento, dois índices; o primeiro é o suporte que mede a frequência de repetição de um item e suas associações dentro da base; e o segundo é a confiança que calcula a frequência das combinações dos itens que passaram pela fase do suporte. Tais medidas são realizadas com a finalidade de possibilitar validade às regras geradas.

Portanto, o algoritmo *Apriori* faz consultas sucessivas às bases de dados devido ao cálculo dos índices, o que poderá deixar lenta a geração dos resultados, mas, segundo Feuser (2017), esse algoritmo possui propriedades que permitem um bom desempenho, como a antimonotonia da relação, que diz que, para um item ser frequente, todo o seu subconjunto também deve ser. Outrossim, usa recursos da memória principal e a estrutura *hash*.

A motivação para esta pesquisa surgiu com a possibilidade da extração das informações da base do INCA através da mineração de dados. Para, deste modo, formar o perfil do paciente com regras relevantes e confiáveis, a fim de que os governantes tenham

uma direção na hora de formular as políticas de assistência à saúde do homem no território nacional.

1.4 Objetivos

1.4.1 Objetivo Geral

Determinar o perfil sociodemográfico do brasileiro com o câncer de próstata por meio do algoritmo *Apriori*.

1.4.2 Objetivos Específicos

- Aplicar processo de pré-processamento de dados na base de dados do INCA com o uso de estatística descritiva;
- Implementar o algoritmo *Apriori* para capturar as regras de associação na base de dados do INCA;
- Identificar as informações de homens diagnosticados com câncer de próstata;
- Determinar o nível de influência dos fatores como idade, raça, tabagismo e alcoolismo na formação do perfil sociodemográfico do homem com câncer de próstata;
- Mapear o perfil sociodemográfico do homem que possui câncer de próstata no Brasil;
- Apresentar os resultados e análises com os perfis encontrados.

2 REVISÃO DE LITERATURA

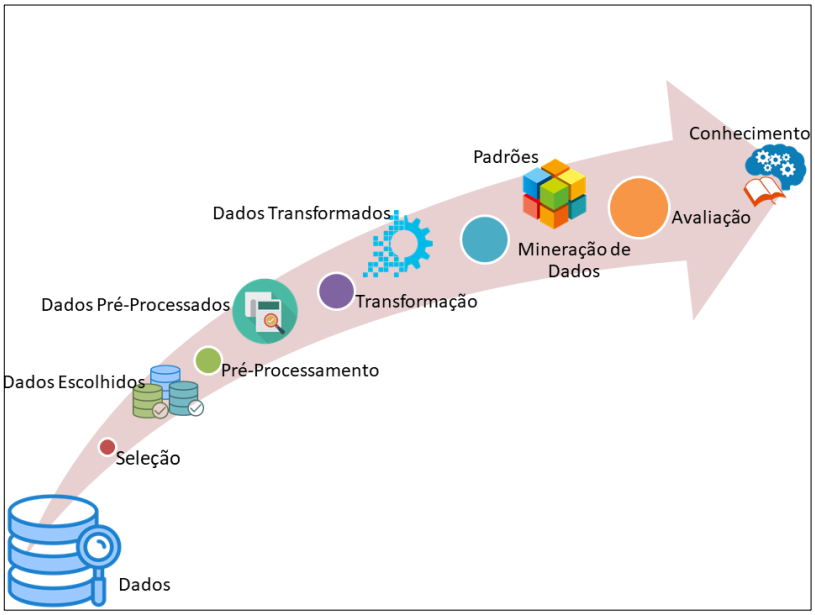
Neste ínterim, será apontada a situação atual dos campos envolvidos nesta pesquisa. Em um primeiro momento, será composta por definições de conceitos básicos sobre mineração de dados; em seguida, o campo referente ao câncer de próstata no Brasil; e, por último, os trabalhos relacionados ao tema.

2.2 Mineração de Dados

A mineração de dados é um processo de análise de informações armazenadas em bases de dados, o qual vai desde banco de dados até grandes arquivos com informações binárias. E segundo Silva et al. (2016), a mineração de dados é considerada a aplicação de técnicas capaz de receber conjunto de dados e devolver padrões de comportamentos.

A mineração entra no meio do processo de descoberta de conhecimento nas bases de dados, que não passa de uma solução para capturar algum conhecimento presente em grandes quantidades de dados (CAMILO; SILVA, 2009). Este processo apresenta as atividades de seleção, pré-processamento, transformação, mineração de dados e avaliação, como podemos ver na Figura 1.

Figura 1 - Representação do Processo de KDD



Fonte: Elaborada pelo autor, 2021.

Os dados podem vir de diferentes fontes, portanto, deve-se estabelecer um objetivo para determinar quais informações farão parte do processo de análise de dados. A seleção de dados é o primeiro passo no processo de aquisição de conhecimento através do KDD. Segundo Campbell (2021, p.28), os resultados e conclusões do processo dependem dos dados selecionados.

Após a seleção, chega-se a uma fase importante para resultados fidedignos, esta fase é denominada de pré-processamento. Nessa fase, são encontradas táticas e ações que filtraram os dados com ruídos, ausentes, duplicados e inconsistentes, que, se não forem tratados, poderão influenciar negativamente nas conclusões do processo de análise de dados.

O pré-processamento é a fase em que será realizada uma limpeza, integração e transformação dos dados. Mas, segundo Campbell (2021, p.16), a primeira tarefa de todas deverá ser a limpeza, pois a análise é resultante de como os dados foram limpos.

No processo de limpeza de dados, encontra-se, entre os problemas, a ausência de valores, que, para sua resolução, pode-se adotar as estratégias de remover ou de substituições por valores padrões. Contudo, independente da estratégia que for adotada para resolução do problema, deve-se precaver para não adicionar um viés aos dados, uma vez que podemos estar eliminando uma quantidade considerável da amostra, ou podemos estar removendo dados que estão associados aos ausentes que podem ser importantes (SILVA et al., 2016; KLOSTERMAN, 2019; CAMPBELL, 2021).

Na remoção de valores, é preciso tomar os devidos cuidados, pois, caso a quantidade da amostra existente seja pequena e ocorra a remoção, pode haver a diminuição ao ponto que a análise não traga resultados que apresentem certo grau de precisão da realidade.

A estratégia de preenchimento automático dos dados ausentes pode ser feita através de atribuições de valores estatísticos da amostra como a média, a moda e a mediana. Além dessas, podem ser utilizados resultados de análises como as de regressão. Porém, como sempre Silva et al. (2016) enfatiza, deve-se tomar o maior cuidado para não criar viés dentro da pesquisa.

Não há só a possibilidade da existência de dados ausentes no meio da amostra, mas também há a presença de dados ruidosos ou conhecidos como *outliers*, que, na fase de limpeza dos dados, devem ser tratados. Os dados ruidosos são os casos de dados fora de intervalo de valores possíveis para o campo, ou caracteres fora do domínio do campo.

Para a remoção de dados ruidosos, pode-se utilizar as estratégias de inspeção e correção manual e a identificação e limpeza automática. Na primeira, com o uso de valores estatísticos como média e desvio-padrão, é possível encontrar a presença de ruídos nos dados

(JAMSA, 2021). Além disso, pode-se fazer uso do gráfico de caixa, que permite verificar graficamente esses valores. Se a quantidade de dados ruidosos for pequena, a correção manual pode ser feita. Mas, quando ocorre em grande quantidade, uma possibilidade seria a suavização dos mesmos.

Na segunda estratégia de limpeza de dados ruidosos, o processo de identificação e limpeza automática ocorre por meio de algoritmos, que, conforme Silva et al. (2016), eles conseguem identificar e organizar esses dados ruidosos em grupos. Assim, eles facilitariam a separação, diminuindo a influência destes na análise final do resultado, permitindo garantir acurácia, completude, consistência e conformidade dos dados, que são itens para definir a qualidade dos dados (JAMSA, 2021).

Logo após a limpeza de dados ruidosos, vem o processo de transformação dos dados, que, de acordo Silva et al. (2016) e Sharma et al. (2019), é o processo de conversão e normalização dos dados antes da realização da mineração dos dados. Esse processo é necessário, pois, além de bases possuírem dados de domínios e grandezas muito diversas, existe o problema de algoritmos que só trabalham com dados categóricos ou dados numéricos.

No processo de transformação, quando é executada a normalização os valores, são organizados em faixas de valores sumarizadas, deixando, deste modo, os valores dentro de um intervalo, podendo, em alguns casos, distinguir dados normalizados de ruidosos.

Logo, de acordo com Silva et al. (2016), há os métodos de discretização e codificação. O primeiro é o procedimento de passar os valores numéricos para categorias. Porém, no segundo, é realizada passagem de categorias para números, mas sem criar relação entre os números sem que inicialmente exista essa relação.

Ao finalizar a transformação, vem o passo de mineração de dados no processo KDD, que é o processo de encontrar padrões dentro de uma grande quantidade de dados, passando e, inclusive, indicando variáveis e tendências ocultas (SILVA et al., 2016; CAMPBELL, 2021; JAMSA, 2021).

Para encontrar esses padrões, são aplicadas técnicas que podem ser classificadas de forma geral como predição, agrupamento e associação (SILVA et al., 2016). Porém, não há impedimento em subdividir estas tarefas por outras.

A predição é o processo de relacionar os dados aos seus atributos dentro de uma organização de dados, e, de acordo com Silva et al. (2016), encaixa-se na técnica supervisionada, em que o procedimento de associação faz a união do dado ao rótulo da informação. Mas também, faz uso do histórico de dados encontrados na database analisada para tomar como instância e exemplo para predizer o futuro (CAMPBELL, 2021, p. 22).

A predição pode ser dividida em duas subtarefas, que são classificação e regressão. A primeira é o processo de analisar o conjunto de dados fornecidos com a database já classificados para poder prever em qual classificação um novo dado pertence, por este motivo ela é definida como supervisionada, pois existe um grupo de dados já treinados para garantir a classificação de forma correta (CAMILO; SILVA, 2009; JAMSA, 2021). Na segunda, há a associação dos dados a um conjunto de valores numéricos e contínuos, essa técnica mede o nível de relacionamento entre um conjunto de variáveis e a ocorrência de um dado.

No agrupamento, temos a análise de conjuntos de dados em que se encontram apenas as descrições conforme Silva et al. (2016). O objetivo desta tarefa é reunir os dados semelhantes entre si, e em diferentes entre grupos (CAMILO; SILVA, 2009). Ou seja, o agrupamento procura encontrar padrões em meio aos dados e não responder uma variável específica. Devido a isso, é classificada como técnica não supervisionada.

O agrupamento pode se assemelhar à classificação, mas, conforme Campbell (2021) diferencia-se pelo detalhe de juntar os dados pela as várias semelhanças existentes entre eles.

Por último, temos a associação, que tem o propósito de buscar ocorrências frequentes e simultâneas entre os elementos de um contexto (SILVA et al., 2016). Ou seja, uma primeira variável, a sua presença ou sua ausência, pode influenciar uma segunda variável, o que pode ser representado dentro da estrutura (SE *atributo* X ENTÃO *atributo* Y), no qual o primeiro atributo é o antecedente e o segundo o consequente.

De acordo com Camilo e Silva (2009), a associação é a técnica mais conhecida devido ao problema da análise da cesta de compras. E, apesar de ser uma técnica muito comum, tanto que se encontra muitos padrões triviais, a busca é sempre pelos padrões inusitados (SILVA et al., 2016).

No processo KDD, a última fase, após a mineração de dados, é a avaliação dos padrões ou resultados encontrados. Nessa fase, encontram-se os especialistas do assunto, os tomadores de decisões e conhecedores do negócio, haja vista que, após a mineração, os resultados deverão ser apresentados em gráficos, tabelas, relatórios, etc.

Seguindo Camilo e Silva (2009), é nesta fase que ocorre testes e validações que garantirão a confiabilidade e a precisão dos dados encontrados. Assim, podemos concluir que, nessa fase, quanto mais fácil e confiável for o entendimento dos resultados melhor para as pessoas que necessitem desses resultados para tomarem as devidas decisões nos seus negócios.

Entre tantas estratégias de mineração de dados, a que poderá trazer resultados mais legíveis para as pessoas é a regra de associação. Conforme Silva et al. (2016), tal estratégia

trabalha com regras que são de fácil interpretação e pode ser traduzida até para linguagem natural.

Essas regras de associações medem o nível de influência de *atributo X* em um determinado *atributo Y*. Além disso, a descoberta de dependências no banco de dados é considerada uma das áreas mais importantes da mineração, haja vista que pode confirmar a necessidade de vários atributos para ocorrência de outros (MIRHASHEMI; MIRZAEI, 2021).

Porém, ao utilizar a regra de associação, é necessário definir o algoritmo que pretende utilizar para construir essas relações dentro dos vários dados existentes na base de dados. Conforme os autores Hermaliani et al. (2020), Mirhashemi e Mirzaei (2021) e Sharma, Meena e Sharma (2021), o algoritmo que mais se destaca é o *Apriori*.

2.2.1 Algoritmo Apriori

Esse algoritmo se sustenta em dois fatores para definir se uma regra de associação é confiável ou não. O primeiro fator é o suporte, este é aplicado na base de dados para medir a frequência com que determinados itens ocorrem juntos em uma regra, ou seja, quando esses itens aparecem dentro do universo de todos os registros na base de dados. Já o segundo é a confiança que calcula a confiabilidade de uma regra, esse cálculo ocorre pela probabilidade condicional de ocorrência de uma regra dada que ocorra os itens antecedentes da regra (AGRAWAL; IMIELINSK; SWAMI; 1993).

Contudo, Mehta e Bura (2020) destaca a medida de *lift*, que, dentro do algoritmo *Apriori*, serve para medir o quanto o consequente é influenciado pelo antecedendo, pois ele determina o quanto os atributos antecedentes e consequentes são dependentes um do outro.

O procedimento para definição dos índices passa pela definição de algumas equações. O primeiro índice se refere ao suporte que mede a frequência relativa com que os itens que compõe a base aparecem juntos em transações individuais, que é obtido através da Equação (I):

$$suporte_{regra}\ (A\ U\ B) = \frac{cont\ (A\ U\ B)}{cont\ (T)} \qquad (I)$$

Onde, temos:

- *cont* é uma operação de contagem;
- *cont (A U B)* indica a união dos itens A e B da regra, ou seja, vai retornar à quantidade de transações nas quais tanto o item A, quanto o B aparecem juntos;
- *T* é o conjunto completo de transações existentes dentro da base.

O segundo índice analisado na formação de regras de associação utilizando o algoritmo *Apriori* é o índice confiança, que tem o objetivo de expressar a importância e a confiabilidade de uma regra, dada à probabilidade de sua ocorrência. Ou seja, como pode ser expressa em porcentagem. Logo, temos uma probabilidade condicional de a regra *(A U B)* ocorrer, dado que sua premissa A ocorre. A obtenção do valor de confiança da regra é feita pela Equação II:

$$Confiança_{regra}\ (AUB) = \frac{suporte_{regra}\ (AUB)}{Suporte_{item}(A)} \qquad (II)$$

A frequência de um item em uma base é determinada pelo usuário que, ao especificar um valor, está definindo um suporte mínimo e todo item que obtiver uma frequência igual ou maior que este suporte é considerado forte.

A propriedade do algoritmo *Apriori* define que, se um item é frequente em uma base de dados, ou seja, a frequência dele é maior que o suporte mínimo, as suas associações com outros itens serão também frequentes (ZHAO; ZANG; CAO, 2009).

No entanto, caso uma regra de associação possuir os índices de suporte e confiança acima dos respectivos valores mínimos, é considerada uma regra forte (SILVA, 2016). Além disso, para fechar a validação da regra gerada, é calculado o índice *lift*, que calcula o quão mais frequente é uma ocorrência da regra B dado que ocorre A.

O cálculo do índice de *lift* é desenvolvido de acordo com a Equação III:

$$lift = \frac{confiança\ (AUB)}{suporte\ (B)} \qquad (III)$$

Ao final, serão geradas as regras de associação, que darão base para a formação dos gráficos para demonstrar, de forma prática e eficiente, o perfil do homem brasileiro com probabilidade a desenvolver o câncer de próstata. E, assim, serão mais bem visualizado os resultados pelos analisadores (CAMPBELL, 2021).

Segundo Romão et al. (1999); Ribeiro, Vieira e Traina (2005); Goldschimidt e Passos (2005, p.62), os valores de suporte e confiança dependem muito do tipo de estudo que se está conduzindo. Ademais, Silva (2016) indica que estes valores precisam ser estudados para cada cenário que é aplicado.

A aplicação de mineração de dados para a tomada de decisão é muito importante (SOUZA; ZAIA, 2015). Na área da saúde, uma decisão que possa assistir melhor a população deve vir de estudos, portanto, é importante uma decisão embasada.

Porém, a aplicação de mineração de dados em saúde será debatida com mais detalhes na seção seguinte.

2.3 Mineração de dados na saúde

A área da saúde produz uma grande quantidade de dados dos seus pacientes devido aos atendimentos que são feitos. Essa quantidade de dados armazena muito conhecimento que, se for estudado, pode melhorar o atendimento e os encaminhamentos de ações na área da saúde.

Como é explicitado por Guisi et al. (2016), as atividades corriqueiras da área da saúde (consultas, exames laboratoriais, prescrições médicas, diagnósticos, vacinações, entre outras) geram uma grande quantidade de dados, e permitem a obtenção de conhecimento inerente aos fatores que se relacionam.

Porém, a aplicação do processo de obtenção de conhecimento em base de dados (KDD) na área da saúde encontra desafios como a falta de preenchimento correto e completo de prontuários, informações contraditórias e campos importantes sem dados. Isso dificulta o trabalho de análise e estudo do cenário da saúde (SOUZA; ZAIA, 2015).

Contudo, é válido deixar claro que a saúde é um campo em que a mineração de dados pode auxiliar a tomada de decisões e facilitar a promoção de saúde, pois, com ela, é possível recolher dados e conhecimentos implícitos que poderão facilitar a tomada de investimento e direcionamento de atenção. Segundo Feuser (2017), a mineração de dados pode aumentar a eficiência do auxílio e do suporte à população na área da saúde.

O cenário brasileiro precisa de uma solução que melhore a eficiência de aplicação dos recursos tanto financeiros, quanto humanos para prevenir e combater doenças, pois, como Sousa e Zaia (2015, p. 15) destacam, o Brasil apresenta uma demanda crescente de pacientes, mas com recursos finitos. Além disso, deveria haver uma ferramenta que facilitasse o entendimento dos dados dos usuários da saúde para poder auxiliar os gestores.

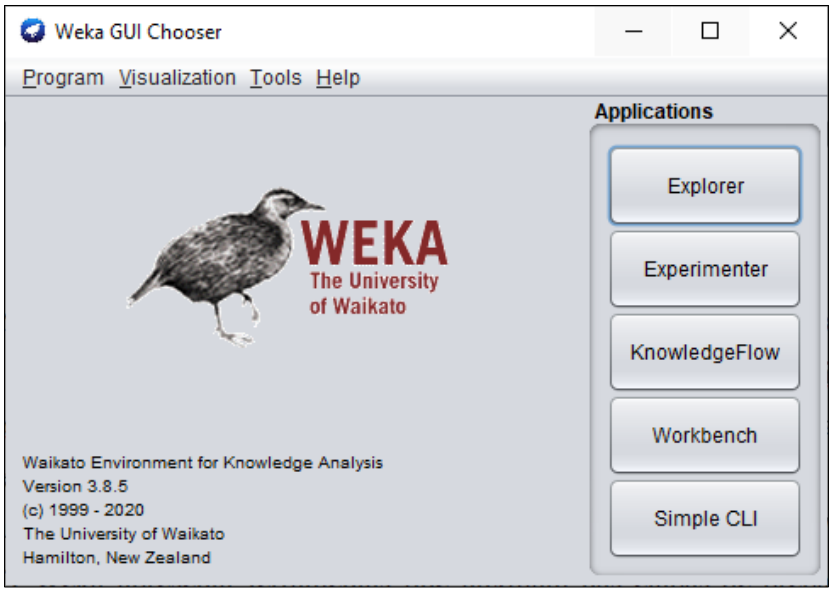
Essa situação não ocorre não só no âmbito da saúde pública, mas também tem ocorrido na área privada, que, por não disporem de meios automáticos de extração das informações de suas bases de dados, enfrentam dificuldades de prevenir doenças (CARVALHO; DALLAGASSA; SILVA, 2015).

De acordo com Carvalho, Dallagassa e Silva (2015) e Guisi et al. (2016), a aplicação de mineração de dados e o desenvolvimento de sistemas para a área da saúde fica restrita à parte administrativa, que controla os pagamentos, agendamentos de exames, encaminhamentos e prescrições. Isso demonstra que a preocupação até o momento não é focada no paciente e que, muitas das vezes, os recursos seriam mais bem administrados se houvesse um estudo dos dados do usuário do sistema de saúde.

Em seu trabalho, Guisi et al. (2016) afirma que, para que a descoberta de conhecimento em base de dados seja aplicada na área da saúde, ainda é incipiente a presença de sistemas comerciais integrados. Os mesmos autores destacam que, apesar da presença de sistemas de prontuário eletrônico, ainda não é possível fazer um estudo utilizando os dados de um país inteiro por falta de integração.

Na ausência de soluções que integrem os dados de todos os pacientes na área de saúde, estudos são realizados de forma individual, por exemplo, estudo desenvolvido com ferramentas *open source* como é o Weka (*Waikato Environment for Knowledge Analysis*), que é um projeto desenvolvido pela Universidade de *Waikato* na Nova Zelândia, o qual tem, em sua implementação, uma coleção de algoritmos de aprendizagem de máquina para a realização de tarefas de mineração de dados, e sua interface é de acordo com a Figura 2 (FRANK; HALL; WITTEN, 2016).

Figura 2 - Software Weka



Fonte: FEUSER, 2017.

O *Weka* apresenta ferramentas que auxiliam nas etapas de preparação dos dados, classificação, regressão, agrupamento, regras de associação e visualização dos dados. E, de acordo com Souza e Zaia (2015), ela é uma ferramenta mais amistosa para o uso por ser uma ferramenta gratuita, eficaz, rápida e efetiva.

Apesar de a mineração ser considerada importante para ser aplicada na área da saúde e a abertura de tecnologias e métodos para aplicar venha se tornando mais prática, há um obstáculo que vem com a disponibilização dos dados para serem analisada (FEUSER, 2017). Guisi et al. (2016) e Feuser (2017) demonstram que isso deve ser por falta de legislação ou devido à regra do sigilo dos dados dos pacientes.

Mas, é notória que a mineração de dados pode auxiliar a tomada de decisões e assim melhorar a assistência a população. Conforme Souza e Zaia (2015), a mineração pode direcionar melhor os recursos para campanhas como a de prevenção ao câncer de próstata, que será debatido na seção seguinte.

2.4 Câncer de Próstata

O câncer de próstata é um dos cânceres mais difícil de ser diagnosticado, já que é assintomático. Assim, se torna um desafio para saúde pública e para a população masculina no momento de identificar que está com o câncer (ARAÚJO et al., 2015). De modo análogo, os fatores cultural e social da população masculina interferem na situação, pois devido à forma como é feito o diagnóstico, os homens impõem dificuldades para ir ao médico (MENEZES, et al., 2019).

O câncer é uma das doenças que mais mata no mundo segundo o INCA. E, de acordo com Ferlay (2013, apud INCA, 2018, p.25), a tendência é só aumentar com o tempo. E entre os homens, o maior causador de óbitos, na escala mundial, fica a cargo do câncer de pulmão seguido de perto pelo câncer de próstata (INCA, 2018; MOTA; BARROS, 2019).

O câncer de próstata é uma enfermidade que acomete os homens normalmente em uma faixa de idade um pouco avançada e pode ser considerada da terceira idade, pois os casos ficam para homens maiores de 60 anos (ARAÚJO et al., 2015; MOTA; BARROS, 2019). Outrossim, existem fatores como histórico familiar, estilo de vida, alimentação e prática de exercícios que podem influenciar na ocorrência do mesmo.

Além das pré-disposição hereditária, o câncer de próstata pode ocorrer em homens com idades acima de 50 anos, com dieta rica em gordura animal, problemas com o tabagismo, problemas com o etilismo, e o estilo de vida (SILVA; NASCIMENTO, 2017; MOTA;

BARROS, 2019). Ou seja, esses fatores atingem parcela específica da população masculina, por isso, a necessidade de campanhas direcionadas para esclarecer detalhes deste câncer.

Mas, apesar de fatores como idade, raça negra e histórica familiar serem fatores já consolidados como de risco na incidência do câncer de próstata, é preciso que se faça uma análise de dados armazenados em banco de dados para identificar relações desconhecidas ou controversas que possam ser criadas com estudos (JUNIOR, et al., 2015; CASSELL, et al., 2019).

Assim, pode-se criar uma explicação para estudos controversos como o realizado por Cassell et al. (2019), o qual demonstrou que no Senegal os registros de parentes de primeiro grau de pacientes com câncer de próstata e o fator raça negra não influenciaram de modo algum a incidência do câncer. Desse modo, estudos são necessários para entendermos melhor essa doença, que atinge muitos homens de idade avançada.

Esse câncer é um dos mais comuns na população masculina e, como citado anteriormente, a tendência é só aumentar (PANIS, 2018). Diante disso, o quesito de detecção de forma precoce se torna um desafio. Porém, deve-se compreender que as formas de diagnóstico existentes podem causar um risco para quem se submete aos exames de diagnósticos devido a complicações que os mesmos podem acarretar (REGO et al., 2020).

Além dos fatores de riscos já bem estabelecidos dos homens que possuem câncer de próstata, existem fatores como inflamação crônica, dieta, baixa exposição ao sol e comportamento sexual (HUSSAIN et al., 2019). Assim, vários fatores deverão ser analisados para poder formar o perfil do homem com câncer de próstata.

Outro fator estudado para determinação de sua influência na presença ou não do câncer no homem é a atividade física, que, segundo Panis et al. (2018), é um dos fatores de risco para obtenção de câncer.

Apesar de a atividade física ser um fator de risco, e ajudar a melhorar o estado de saúde no geral de pacientes com câncer, inclusive os acometidos pelo câncer de próstata, estudos demonstram que os pacientes com câncer de próstata não têm incluído a atividade física nas suas práticas do dia a dia (STONE et al., 2019).

No guia para atividades físicas da Organização Mundial de Saúde (OMS), há uma recomendação para que todo adulto realize de 150 a 300 minutos de atividade física moderada, ou 75 a 150 minutos de atividades físicas vigorosas. Caso não ocorra nessa distribuição, pode ser feita uma mesclagem durante a semana (OMS, 2020).

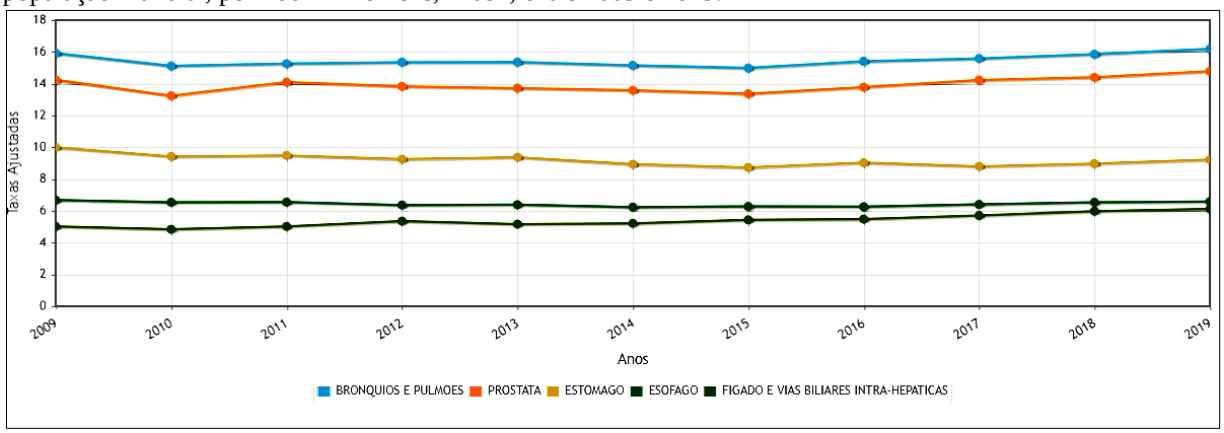
Mundialmente, o câncer de próstata apresenta uma taxa de morte em torno de 164.690, ou seja, cerca de 19% das mortes por câncer (HUSSAIN, 2019). Ademais, o próprio Hussain

(2019) determina que, se tiver dois ou mais parentes de primeiro grau com câncer de próstata, a probabilidade de ter o câncer sobe de 5 a 11 vezes.

Mas também particularizando os valores do cenário de câncer de próstata, na Nova Zelândia, Austrália, a taxa de homens com câncer de próstata fica no valor de 111,6 por 100 mil, enquanto, na América do Norte, esse valor é de 97,2. Acredita-se que esses valores só tendem aumentar, pois melhoraram as técnicas e os tratamentos do câncer de próstata (MENEZES et al., 2019; STONE et al., 2019). Assim, a Organização Mundial de Saúde estima que irá ter 499 mil mortes de câncer de próstata até 2030 (FERRÃO; BETTINELLI; PORTELLA, 2017).

Na realidade brasileira, o câncer de próstata é o que mais mata, tanto que o INCA realizou um estudo e demonstrou que, para o biênio 2018-2019, estipulando que seria 68 mil novos casos chegando ao total de 31,7% dos casos de câncer (INCA, 2018).

Gráfico 1 - Taxas de mortalidade das 5 localizações primarias mais frequentes, ajustadas por idade, pela população mundial, por 100 mil homens, Brasil, entre 2009 e 2019.



Fonte: INCA, 2021.

Segundo o Gráfico 1, a realidade brasileira apresenta dois tipos de câncer que mais atacam a população masculina, pulmão e próstata. Além disso, percebe-se que, com o passar dos anos, no período de 2009 a 2019, em que o ano de 2019 é o último ano disponível na ferramenta do Atlas da mortalidade do Ministério da Saúde, o câncer de próstata só vem crescendo.

Tabela 1 - Taxas de mortalidade por câncer de Próstata, brutas e ajustadas por idade, pelas populações mundial e brasileira de 2010, por 100 mil homens, no Brasil, entre 2009 e 2019

Localidade	**Taxas Específicas**											**Taxa Bruta**	**Taxa Padronizada**	
	00 a 04	05 a 09	10 a 14	15 a 19	20 a 29	30 a 39	40 a 49	50 a 59	60 a 69	70 a 79	80 +		P. Mundial	P. Brasil
Centro-Oeste	0,00	0,01	0,05	0,03	0,09	0,51	6,24	44,64	187,07	593,89		13,37	15,32	18,67
Nordeste	0,00	0,00	0,01	0,02	0,06	0,08	0,59	6,83	44,48	179,82	604,18	14,34	15,26	18,65
Norte	0,00	0,00	0,00	0,03	0,05	0,04	0,66	5,45	38,31	167,55	559,42	8,76	13,89	17,06
Sudeste	0,01	0,01	0,01	0,04	0,03	0,06	0,56	6,00	39,68	160,95	460,59	14,64	12,84	15,44
Sul	0,00	0,00	0,01	0,02	0,05	0,07	0,55	5,90	41,81	184,63	551,40	17,09	14,60	17,74
Brasil	0,00	0,00	0,01	0,03	0,04	0,07	0,57	6,16	41,43	171,72	525,00	14,30	13,94	16,91

Fonte: INCA, 2021.

De forma mais detalhada na Tabela 1, percebe-se que a realidade brasileira apresenta mais casos de próstata nas idades mais elevadas, acima de 60 anos, e as regiões que mais possuem casos é a região nordeste e sul. Mas também, pode se depreender dessa tabela, quando comparamos com a escala mundial, o Brasil apresenta valores maiores, fato que Panis et al. (2018) já destaca em seu trabalho.

O câncer de próstata é um caso de saúde pública que preocupa, pois vem aumentando nos últimos anos sua ocorrência (ARAÚJO et al., 2015; SACRAMENTO et al., 2019). No Brasil, uma iniciativa foi adotada em 20 de setembro de 2001, a partir da qual foi instituído o Programa Nacional de Controle do Câncer de Próstata, através da lei 10.289, vislumbrando atender esta demanda de casos e capacitar os profissionais envolvidos no atendimento e prevenção do câncer de próstata.

Além dessa lei, outra mais recente foi elaborada e aprovada pelo governo brasileiro, que ajuda ao diagnosticado com câncer, essa é a Lei 12.732, que institui um prazo de 60 dias para o início do tratamento após o diagnóstico para todos os tipos de neoplasias no Sistema Único de Saúde (SUS) (SACRAMENTO et al., 2019).

O diagnóstico do câncer de próstata apresenta como método ouro o exame de PSA, juntamente com o exame de toque realizado pelo urologista (REGO, et al., 2020), e poderia ser uma forma de realizar o rastreio do câncer de próstata na população masculina. Porém, segundo a Sociedade Brasileira de Urologia (SBU) e a Sociedade Brasileira de Patologia

Clínica Medicina Laboral (SBPC/ML) (2018), não é recomendado fazer o rastreio de modo geral, pois considera-se que a realização de biopsia de possíveis casos de próstata pode colocar em risco a vida de homens devido às complicações que a biopsia pode ocasionar.

Mas, na literatura, já encontramos outras ferramentas e técnicas para diagnosticar o câncer de próstata tais como a ultrassom transretal, exame retal digital e imagem de ressonância magnética, que podem ser métodos não invasivos a serem utilizados (HUSSAIN, et al., 2019).

O uso de tecnologias para extrair conhecimento de bases de dados é uma forma de identificar fatores de risco desse câncer, mas a literatura apresenta mais técnicas e modos. Tais procedimentos serão mais detalhados na seção seguinte com alguns trabalhos e os métodos utilizados para o diagnóstico do câncer de próstata.

2.5 Trabalhos Relacionados

No propósito de definir um perfil do acometido do câncer de próstata, a literatura ainda se encontra em estágio inicial e ainda é uma questão desafiadora para a ciência (ARAÚJO et al., 2015; REGO, et al., 2020).

Porém, a seguir, encontramos trabalhos como o de Araújo et al. (2015), em que os autores caracterizaram o perfil dos pacientes diagnosticados com câncer de próstata em um hospital em Ribeirão Preto no estado de São Paulo. Neste estudo, os autores ficaram restritos a 2620 homens diagnosticados com câncer de próstata. Os dados foram reunidos no software *Microsoft Office Excel* 2010, havendo análise através da estatística descritiva.

Utilizando-se novamente do *Microsoft Office Excel* (MATOS; BARROS, 2019), realizaram o levantamento do perfil dos pacientes diagnosticados com câncer de próstata no hospital de câncer de Pernambuco, situado na cidade de Recife. Para isto, foram utilizados os prontuários de 100 pacientes diagnosticados com adenocarcinoma da próstata no ano de 2013. Para a análise dos dados, consideraram o percentual e a frequência.

Seguindo a política de campanhas de prevenção do câncer, temos o trabalho de Rego et al. (2020), que, na campanha de Mutirão de Prevenção ao Câncer, que ocorreu no município de Montes Claros, no estado de Minas Gerais, quando fizeram o levantamento de variáveis: sociodemográficas (faixa etária, estado civil, e escolaridade), histórico familiar de câncer e hábitos de vida (tabagismo, etilismo e prática de atividade física). Esses dados foram levantados em uma amostra de 603 homens, que foram previamente estratificados baseados na idade, incluindo a faixa etária de 50 a 70 anos. Os dados foram tabulados e analisados de

forma descritiva, usando tabela de frequência e gráfico gerados pelo software SPSS na versão 23.0.

Com o uso da mineração de dados, Filho (2017), em seu trabalho, fez uso dessa técnica para analisar as informações contidas nos boletins epidemiológicos do portal do Ministério da Saúde. Tais boletins dizem respeito às informações de dengue em 18 municípios da Paraíba. Com o uso do *software* livre *Weka,* analisou-se os dados, através da aplicação do algoritmo *Apriori*, que já se encontra na ferramenta. Ao final, encontrou-se regras de associação que geraram padrões, os quais fizeram parte da proposta do trabalho direcionando ações de combate à Dengue nos munícipios analisados.

3 METODOLOGIA

Nessa seção, encontram-se todos os procedimentos para a realização do trabalho, a fim de tornar possível sua reprodução, como sugere o método científico, pois, conforme Wazlawick (2009), o não seguimento do método científico pode levar a conclusões erradas.

Esta pesquisa se caracterizará de acordo com os critérios definidos por Prodanov e Freitas (2013). Desse modo, ela pode ser considerada como pesquisa de natureza aplicada, de um método científico dedutivo, objetivos descritivos, um procedimento de um estudo de caso e uma abordagem quantitativa.

A caracterização da metodologia da pesquisa permite a sua reprodução em meio acadêmico. Assim, pode-se dizer que a pesquisa é aplicada em sua natureza, por tentar definir um perfil específico do homem com o câncer de próstata.

Quanto ao método científico que será utilizado, ela se apresenta, de forma dedutiva, uma vez que será, a partir da realidade nacional brasileira da população masculina, que se irá construir o perfil sociodemográfico dos homens com câncer de próstata.

A pesquisa, ainda, apresenta um objetivo descritivo, haja vista a análise dos dados presentes nas bases de dados do INCA, para expor as características dos indivíduos com câncer de próstata. Isto será possível através da técnica de associação da mineração de dados, pois serão criadas várias regras de associação a partir do algoritmo *Apriori*, que coletará os dados dessa base.

No caso do procedimento adotado, a presente pesquisa é caracterizada como estudo de caso, pois o propósito é analisar uma situação específica da população masculina brasileira para a formação de um perfil, que descreva uma parte dela que esteja propensa ao câncer de próstata.

E, por fim, a sua abordagem é definida como quantitativa, pois os dados serão analisados considerando sua frequência relativa e a confiabilidade das regras de associação encontradas nos dados da base de dados do INCA com as características dos homens com câncer de próstata.

Porém, antes de iniciarmos a pesquisa de fato, para o embasamento desta, em primeiro plano, foi realizada uma revisão sistemática sobre o tema do presente trabalho, com o propósito de verificar o estado da arte. Outrossim, foi seguido um protocolo, ver Apêndice A, para fosse possível sua reprodução no futuro.

Nas seções a seguir, é detalhada a descrição de todo o procedimento metodológico adotado nesta pesquisa com os seus detalhes.

3.1 Fonte de dados

A primeira etapa do processo de adquirir conhecimento por meio do KDD é a aquisição dos dados que serão analisados durante todo o processo. Para o estudo que se segue, foi adquirida a base de dados do INCA.

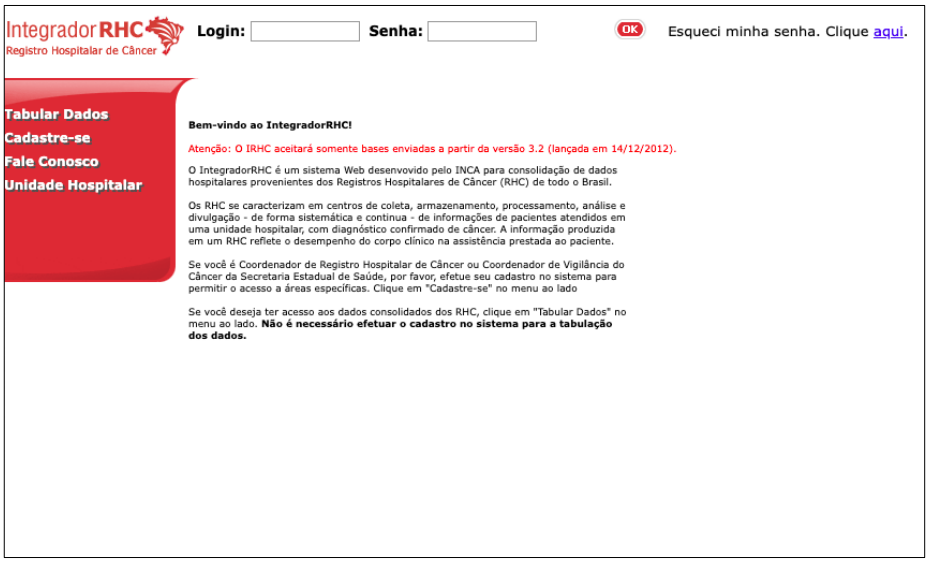
Essa base de dados surge dos registros hospitalares do câncer que, devido à sua importância, o Ministério da Saúde, em 1998, publicou a Portaria Nº 3535/GM, que estabeleceu requisitos para cadastramento de centros de atendimentos em oncologia, e tornou obrigatório o funcionamento do Registro Hospitalar de Câncer (RHC) (BRASIL, 1998).

Além disso, em 2005, o Ministério da Saúde, com o propósito de organizar a rede atenção oncológica, instituiu a Política Nacional de Atenção Oncológica, que tornou obrigatório os registros de câncer e define parâmetros para sua implantação e funcionamento.

Ademais, a Portaria MS/SAS Nº 741 de 2005, em seu parágrafo único, determina que "Arquivos eletrônicos dos dados anuais consolidados deverão, no mês de setembro de cada ano, a partir de 2007, ser encaminhados para o INCA, que deverá publicá-los e divulgá-los de forma organizada e analítica" (BRASIL, 2005). Isso faz do INCA o concentrador de das informações dos tratamentos de câncer do país.

Portanto, pertence ao INCA o local onde estão os dados de todos os cânceres e estão disponíveis para serem consultados. A aquisição dos dados para a análise foi possível pelo portal do Integrador RHC, Figura 3, que está disponível no sítio *https://irhc.inca.gov.br/RHCNet/*, e, através da opção tabular dados, irá se abrir uma janela onde se encontra a opção de *download*. Logo em seguida, temos a opção de fazer o *download* dos dados através de três modos: Base de dados de todos os estados, exceto SP; Base de dados de todos os estados; e Base de dados do estado.

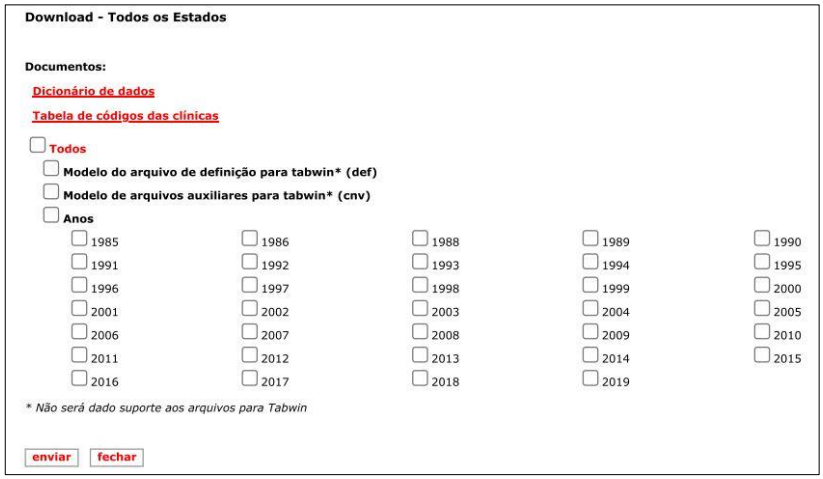
Figura 3 - Portal Integrador RHC



Fonte: Elabora pelo autor, 2021.

Para este estudo, foi realizado o *download* da base de dados de todos os estados, haja vista necessidade para formação do perfil nacional. Ao selecionar essa opção de *download*, há o encaminhamento para terceira janela, ver Figura 4, em que foi selecionado o período de 2010 a 2019, por ser a última década disponível para *download* no sistema.

Figura 4 - Download de dados do Portal Integrador RHC



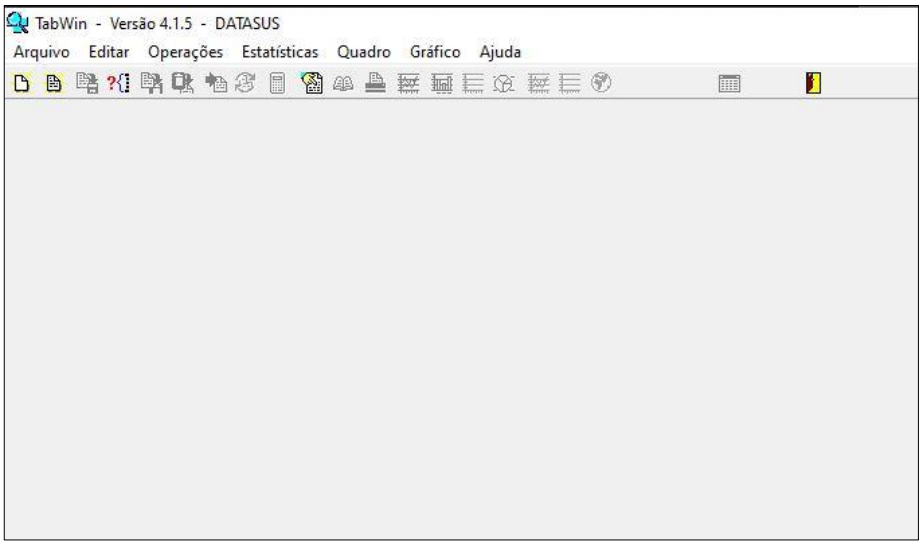
Fonte: Elaborada pelo autor, 2021.

Além dos dados em si, para realizar o estudo, foi necessário o *download* do dicionário da base de dados, em que se encontra disponível o significado de cada código que é salvo na base de dados do portal do Integrador RHC. Esse pode ser acessado no Anexo A.

3.2 Recursos de softwares

Para analisar os dados, foram utilizadas as ferramentas: *Tab* para *Windows* (*TabWin*) e S*pyder*. O *TabWin* é o sistema para facilitar a tabulação e tratar os dados gerados pelo aplicativo TABNET[1], foi desenvolvido pelo DATASUS. Ele se encontra disponível para *download* e com o manual de instalação no sítio do portal da saúde do DATASUS. Além disso, o *TabWin* permite operações aritméticas e estatísticas nos dados da tabela gerada ou importada, elaborar gráficos e mapas, e efetuar operações nas tabelas importadas, por exemplo, exportar para arquivos como *Comma-separated values* (CSV).

Figura 5 - *TabWin*



Fonte: Próprio Autor, 2021.

Após o *download* do arquivo exportado pelo portal Integrado do RHC, foram verificados que os arquivos com os dados de cada ano vinham na extensão dbf, que, para ser trabalhado na linguagem *Python*, foram convertidos para arquivos de extensão csv.

Assim, inicialmente, com ajuda do *TabWin*, cada arquivo dbf, que possui dados de cada ano, foi convertido em arquivo csv utilizando a opção ver arquivo dbf, que se encontra no menu arquivo. Ao abrir o arquivo, é permitido exportar para vários outros formatos inclusive csv.

Com os arquivos convertidos para o formato csv, utilizamos o *Spyder* versão 4.1.4, ver Figura 6, que, segundo Sharma e Bansal (2020), é uma ferramenta técnica poderosa de

[1] Tabnet: Tabulador genérico de domínio público que permite organizar dados de forma rápida e se encontra disponível no sítio: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/

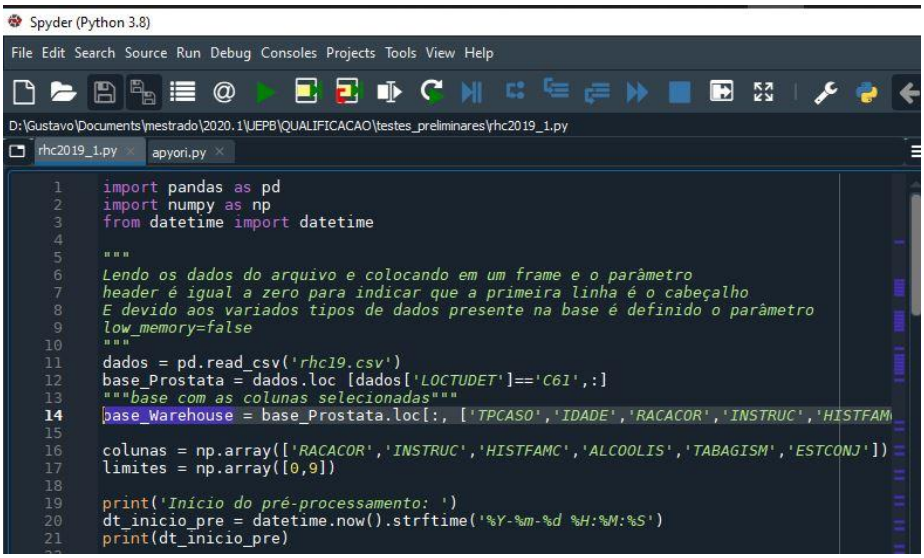
Python. Ela se encontra no ambiente de desenvolvimento integrado de *Python*, chamado de Anaconda, ou pode ser feito o download no seu sítio[2], que é um ambiente *open source*, ou seja, *software* livre e de código aberto e está disponível no sítio *www.anaconda.com/products/individual*. Para o nosso projeto, utilizamos a versão para *Windows*, tendo em vista que o computador disponível apresentava as configurações de acordo com a Tabela 2.

Tabela 2 - Computador disponível

Componente	Característica
Sistema Operacional	*Windows* 10
Processador	Core i7 2.0 GHz 3ª Geração
Memória RAM	12 GB
Capacidade de armazenamento (Disco Rígido)	1 TB

Fonte: Elaborada pelo autor, 2021.

Figura 6 - *Spyder*



Fonte: Elaborada pelo autor, 2021.

A escolha do ambiente Anaconda foi realizada por ser uma ferramenta *open source*, a qual, além de ser de código aberto, não exige pagamento de licenças, por ser um ambiente versátil, e por possuir vários pacotes pré-instalados como *numpy*, *mathplotlib*, *scipy*, *seaborn* e etc (SHARMA; BANSAL, 2020).

[2] https://www.spyder-ide.org

3.3 Transformação dos dados

Para possibilitar um resultado satisfatório e de confiança à fase de preparação dos dados, é importante que seja executada a seleção, integração e limpezas dos dados (CAMPBELL, 2021).

O processo de transformação consiste em preparar e organizar os dados para realização da mineração de dados, que pode ser realizada por meio da criação de consultas com a linguagem *Structured Query Language* (SQL), aplicações personalizadas, ou ferramentas de terceiros (JAMSA, 2021).

Para a aplicação da transformação neste projeto, primeiro foi realizado o processo de seleção dos dados, integração e limpeza, pois os dados brutos apresentavam dados de outros cânceres e com ruídos, que poderiam prejudicar a mineração de dados. Para o desenvolvimento deste projeto, houve o uso da linguagem de programação *Python*, e, com o auxílio da biblioteca *Pandas* foi possível a conversão de dados numéricos brutos da base para dados categóricos, como pode ser visto na Figura 7 o trecho do código onde é realizada a passagem do atributo raça numérico para categórico.

Figura 7 - Código de conversão do atributo numérico raça para categórico

```
raca = ['BRANCA','PRETA','AMARELA','PARDA','INDIGENA','S/I-RA']

classes_raca= [0,1,2,3,4,8,9]

column_raca = pd.cut(x=base_Warehouse['RACACOR'], bins=classes_raca,
labels=raca)
```

Fonte: Elaborada pelo autor, 2021.

3.4 Regras de Associação

Ao final do pré-processamento, os dados ficaram organizados de forma tabular que permitiu a aplicação do algoritmo *Apriori*, que é explicado na seção 3.4.1.

3.4.1 Definição de parâmetros do algoritmo Apriori

Para a aplicação do algoritmo *Apriori* na base em estudo, definimos os índices suporte, confiança e *lift*. O refinamento dos índices ocorreu com testes realizados em dados que representam dois anos da amostra estudada, e seguindo os procedimentos abaixo.

Para a definição do valor do suporte, realizou-se o cálculo através da Equação IV, que se baseia na Equação I.

$$Suporte_{regra}(CP) = \frac{qtde\ (CP)_{período}}{qtde\ C_{período}} \qquad (IV)$$

Onde, temos:

- *Qtde (CP)* $_{período}$ para a quantidade de casos de câncer de próstata ocorrida naquele período;
- *Qtde C* $_{período}$ para a quantidade de casos de câncer naquele período.

Já para definição do índice de confiança, devido à ausência de critérios específicos para o caso, que indicasse o valor, foram realizados testes onde seu valor se modifica a partir de 0,3 até o valor de 0,8 variando em 0,1; fazendo a seleção dos resultados únicos, ou seja, removendo os repetidos. Nesse intervalo, foi o momento em que apareceu a maior quantidade de resultados, haja vista que valores altos de suporte podem deixar de lado regras interessantes. Mas, para a confiança valores altos trazem a tendência de retornar às melhores regras (BALDOMIR, 2017).

Portanto, foram realizados testes com esses valores para filtrar os melhores resultados e, assim, conseguir extrair a maior quantidade de conhecimento possível que os dados da base de dados estudada pode oferecer.

Para a definição do *lift*, após realização de testes nos dois primeiros anos, a quantidade maior de resultados se deu quando o *lift* foi definido como 1,2; para confiança mínima da regra, ou seja, para que cada regra fosse considerada forte, teria de se indicar que, no mínimo, ocorreria um inteiro e dois décimos do consequente da regra de dados que ocorreu o antecessor da regra. Pois, de acordo Jamsa (2021), valores perto de um indicariam que a regra seria uma mera coincidência, e, acima disso, seria realmente uma associação.

3.5 Seleção dos atributos

Para análise, foram obtidos os dados do período de 2010 a 2019, pois era a última década disponível no portal do Integrador RHC do INCA, até o presente estudo. Para cada ano, desta década a ferramenta de exportação do portal do Integrador do INCA criou um arquivo dbf com colunas e linhas, em que as colunas representam os atributos e as linhas os casos.

Para a seleção dos fatores que ajudassem a desenvolver o perfil sociodemográfico do homem com câncer de próstata, foram levados em consideração os trabalhos analisados da literatura e os fatores presentes na base de dados do INCA. Assim, filtrou-se, entre todas as colunas presentes na base, as seguintes variáveis: o ano de diagnóstico, data do primeiro diagnóstico, estado de residência, idade, grau de instrução, raça/cor, e tipo de caso. Esses atributos são obrigatórios, ou seja, sempre terão algum valor armazenado.

Mas, foram selecionados alguns dados opcionais, como consumo de álcool, tabagismo, estado conjugal atual e histórico familiar. Porém, segundo o INCA (2010), a partir do momento em que a unidade hospitalar resolver preencher os dados opcionais, o preenchimento se tornará obrigatório para todo o registro que for cadastrado.

Como procedimento para selecionar os dados, foi consultado o dicionário da base de dados do SisRHC, atualizado e disponível para download a partir do dia 03 de março de 2020, ver Anexo A. Em seguida, foram aplicadas as etapas para o pré-processamento dos dados analisados:

1. Filtragem dos casos de câncer de próstata através do código de Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID), que nesse caso é C61, e o atributo fica na coluna LOCTUDET;
2. Utilizando a biblioteca Pandas do Python, foram selecionadas as colunas: TPCASO, IDADE, RACACOR, INSTRUC, HISTFAM, ALCOOLIS, TABAGISM, ESTADRES, ANOPRIDI, ESTCONJ, DTDIAGNO;
3. Seleção e remoção de registros com raça, escolaridade, histórico familiar, alcoolismo e tabagismo fora das opções 0 e 1, de acordo com o dicionário da base;
4. Seleção e remoção de registros com idades fora do intervalo de 0 a 150 anos;
5. Seleção e remoção de registros com ano de diagnóstico fora do intervalo de 1900 a 2021;
6. Conversão dos campos numéricos (RACACOR, ALCOOLIS, ESTCONJ, HISTFAM, INSTRUNC, TABAGISM e TPCASO) para campos categóricos.

Com os dados selecionados para o próximo passo, foi desenvolvido uma aplicação na linguagem *Python* e aplicado o algoritmo *Apriori* implementado por Yu Mochizuki denominado apyori[3] em *Python, que* se encontra na versão 1.1.1. Essa aplicação proporcionou a geração das regras de associação, e, deste modo, a formação do perfil sociodemográfico do homem brasileiro com o câncer de próstata.

3.6 Pós-Processamento dos dados

Com as regras de associação geradas, os dados foram analisados de acordo com fundamentação teórica já exposta para verificar se a realidade brasileira segue o mesmo perfil ou apresenta peculiaridades com relação ao encontrado na literatura consultada.

Ademais, para permitir um maior entendimento dos leitores e gestores que possam utilizar este trabalho para tomar decisão no campo das campanhas sobre o câncer de próstata, será feio o uso de gráficos, histogramas e planilhas com as regras de associação encontrada com a análise da base do INCA. Afinal, segundo Campbell (2021), as pessoas absorvem melhor a informação quando é representada visualmente.

[3] https://pypi.org/project/apyori/

4 RESULTADOS E DISCUSSÕES

O propósito de analisar uma década de dados para definir o perfil sociodemográfico do homem brasileiro com o câncer de próstata, a partir do período de 2010 a 2019 de dados do INCA, foi alcançado. Para embasamento da discussão, gráficos foram produzidos com uma aplicação desenvolvida em *Python*. Essa aplicação se encontra disponível no repositório do GitHub regras de associação[4].

A partir dos dados selecionados, foi possível montar uma amostra de 1.844.810 registros contidos na base de dados do estudo. Mas, depois do filtro C61, que corresponde ao CID da doença, realizado na coluna LOCTUDET para filtrar os dados sobre o câncer de próstata, restou um total de 222.951 registros. Com base nesses dados, em seguida, foram aplicados os filtros para excluir dados ruidosos e ausentes, que poderiam estar presentes.

No pré-processamento, foram removidos 144.237 registros, que equivalem a 64,69% da amostra, que são dados ruidosos, como, por exemplo, valores de idade com a atribuição de 9999 e Estados de residência declarados como 99 e 77, que não fazem parte do dicionário da base. Além disso, foram removidos os registros, que apresentam a declaração de "sem informação", que é selecionada nos quesitos do formulário de cadastro do paciente, a qual pode influenciar nas regras, haja vista que, em dados opcionais (Alcoolismo, Tabagismo, Histórico familiar e Estado Conjugal), muitos escolheram essa opção ou o atendente a selecionou na hora do preenchimento do formulário. Como pode ser visto nos Gráficos 2, 3, 4 e 5.

Ademais, os dados ausentes na base não ultrapassaram o valor de 119 registros, que corresponde ao percentual de 0,05% de todos os dados analisados. Porém, os valores referentes aos dados ruidosos de idade e opções 77 e 99 para o estado de residência do paciente chegaram ao valor de 1843 registros, isso corresponde a 0,8% da amostra. E, por último, dos valores referentes as opções que se declarou "não possuir informações" (S/I) sobre o quesito, a quantidade removida equivale a 142.394 registros que, no geral, corresponde a 64,4%, aproximadamente, dos dados restantes depois da aplicação dos filtros anteriores.

Contudo, o INCA (2020), em sua análise, destacou que, no período de 2012 a 2016, esses dados opcionais são coletados pelas instituições informantes do RHC de forma inconsistente, e não tem como distinguir em qual instituição os itens passaram ser

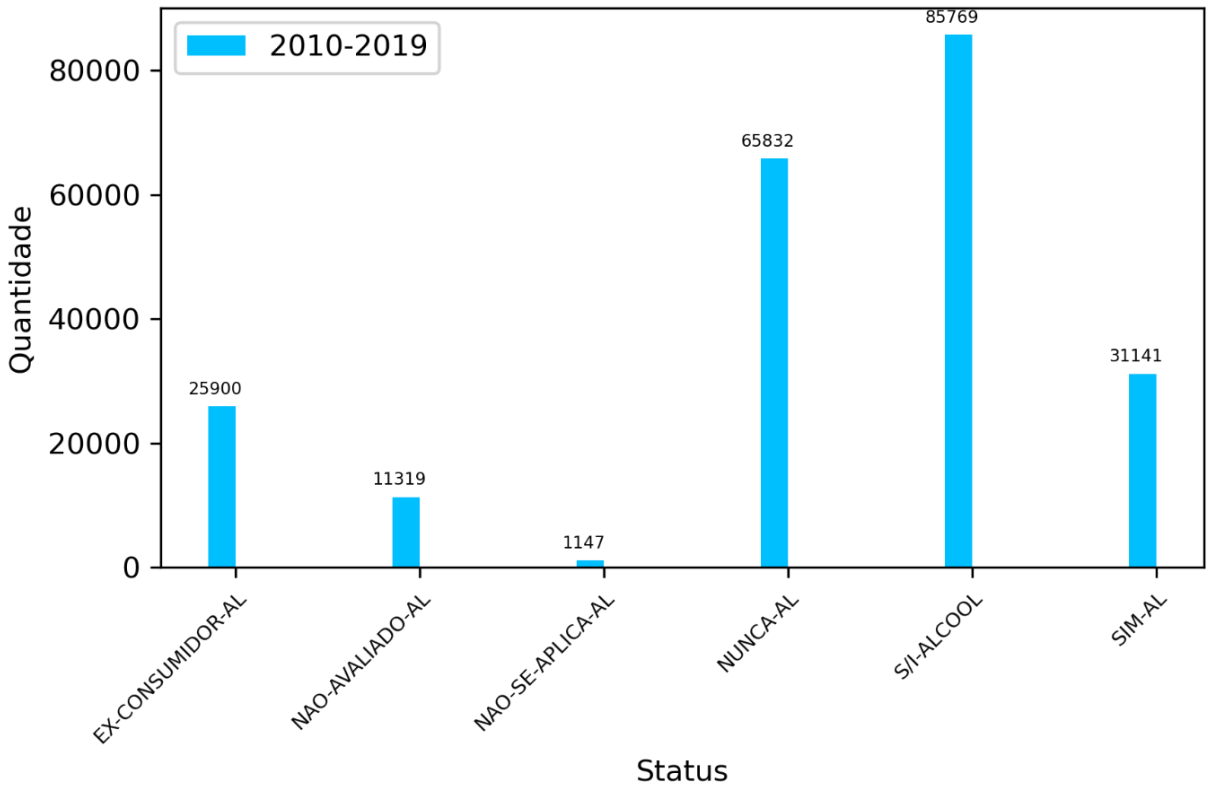
[4] https://github.com/DiasGustavo/regrasdeassociacao

obrigatórios, haja vista que, a partir do momento que a instituição coleta esses dados pela primeira vez, para os demais casos que venham a ser registrados, esses itens passam a ser obrigatórios. Ou seja, muitos dos dados opcionais não foram preenchidos. Ademais, segundo INCA (2010), quando é selecionada a opção “sem informação” ou o dado foi coletado incorretamente ou não há informações.

Esses tipos de informações podem criar tendências para resultados incoerentes tanto com a realidade quanto com os dados presentes na base. Nesse sentido, segundo Frank, Hall e Witten (2016, p.63), os erros e omissões em dados podem adquirir importância dentro da base de dados. Deste modo, foi optado pela exclusão dos dados incoerentes ou ausentes, pois pode gerar resultados equivocados.

Contudo, inicialmente, foram analisados os campos opcionais, pois trazem informações que caracterizam a população masculina em fatores de risco. O primeiro ponto analisado será referente ao consumo de álcool que fiou distribuído de acordo com o Gráfico 2.

Gráfico 2 – Quantitativo de respostas para o quesito sobre o consumo de álcool.



Fonte: Elaborada pelo autor, 2021.

Como pode ser apreendido do Gráfico 2, o total de registros declarados “sem informação” é a opção mais encontrada dentro dos prontuários dos pacientes. Assim,

traduzindo em números, os registros chegam ao valor de 85.769, que corresponde ao percentual de 38,8 % da amostra.

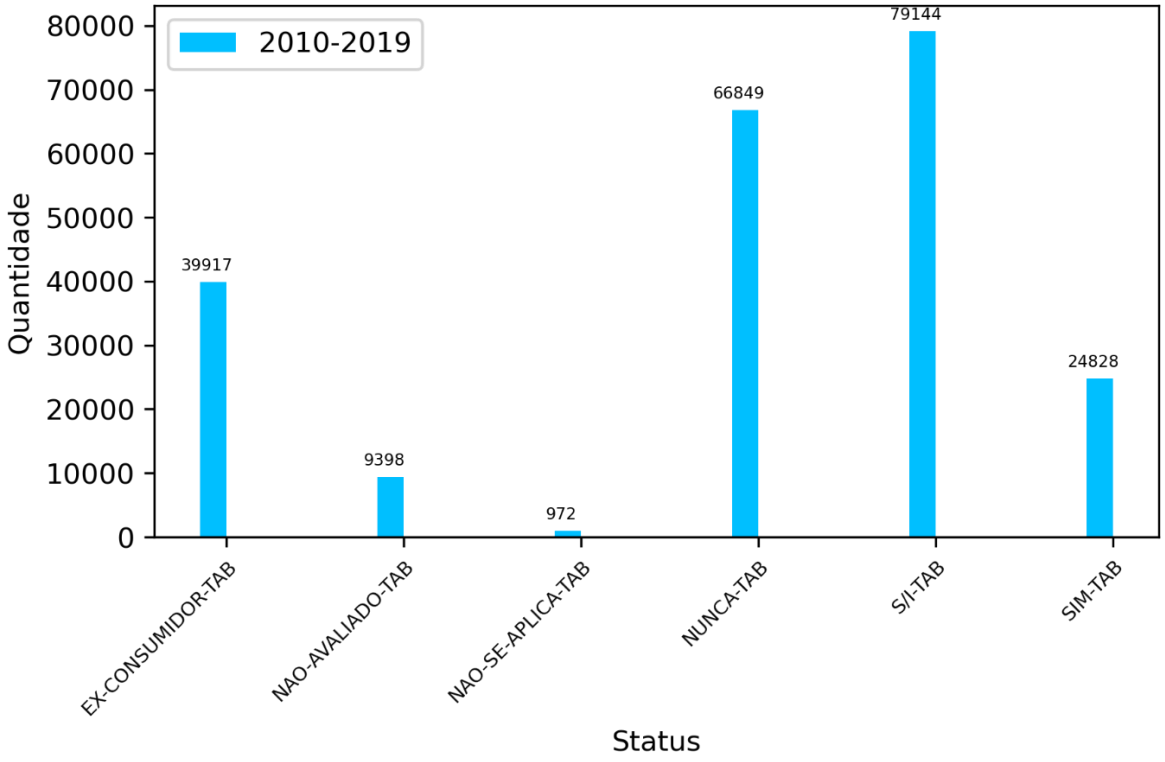
Ao eliminar esses dados inconsistentes, encontra-se a informação de que a maioria declarou nunca ter bebido, isso chegou ao valor de 65.832 registros. Assim, o resultado aproxima-se ao da pesquisa de Rego et al. (2020), que apresentaram índices maiores para os homens que possuem câncer de próstata e não bebem.

De acordo com a Pesquisa Nacional de Saúde (PNS), realizada em 2019, pelo Instituto Brasileiro de Geografia e Estatísticas (IBGE), a população masculina acima de 18 anos, formada, aproximadamente, 76,69 milhões, o índice de pessoas que bebem pelo menos uma vez ou mais por semana é 37,1%; apesar da maioria não beber, o consumo excessivo de álcool é causa de 5,3% das mortes daquele ano no mundo (IBGE, 2020c). E em comparação aos dados do IBGE, os consumidores e ex-consumidores, encontrados na base do INCA, representam apenas 0,2% da população masculina.

Assim, depreende-se que, apesar de o consumo de álcool ser considerado como um fator de risco para a saúde, a população masculina brasileira, em sua maioria, não faz uso regular de álcool de acordo com os dados do INCA e do IBGE.

Não só o alcoolismo é um fator impactante no câncer, como também o uso do tabaco contribui para o desenvolvimento do câncer (HUSSAIN, et al., 2019). Com o auxílio do Gráfico 3, foram analisadas as respostas referentes ao consumo do tabaco por pacientes com câncer de próstata.

Gráfico 3 – Quantitativo de respostas para o quesito sobre o consumo de tabaco.



Fonte: Elaborada pelo autor, 2021.

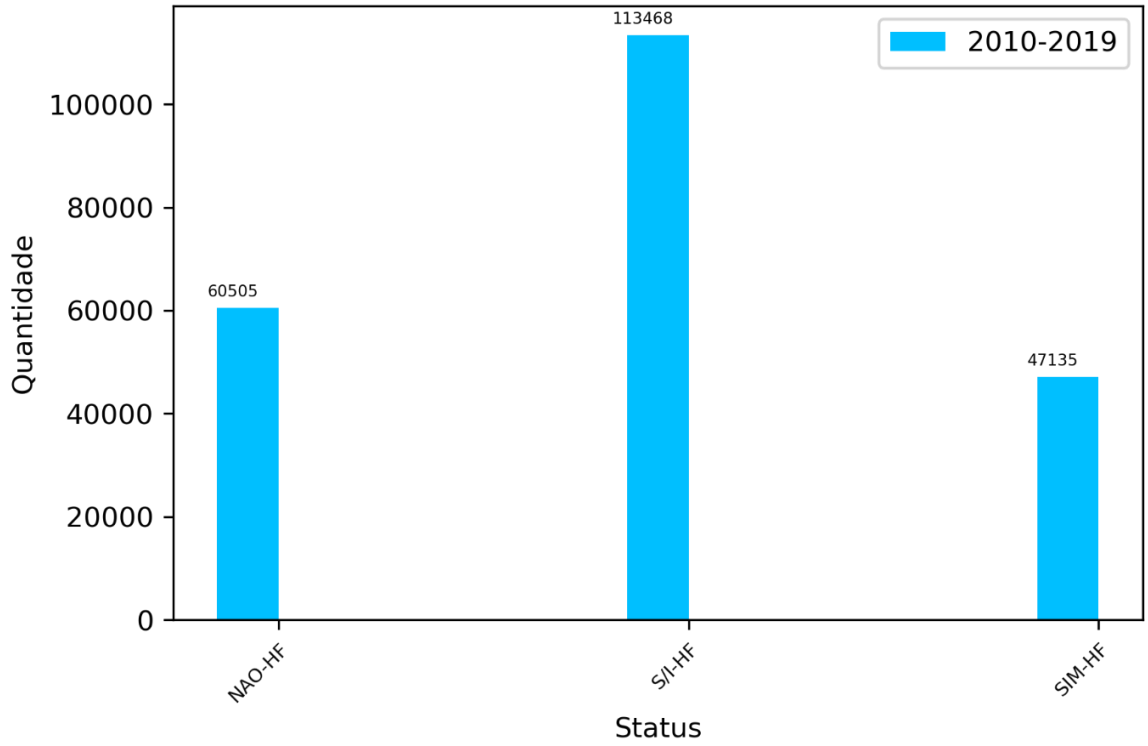
Neste Gráfico 3, também se identifica um valor maior para opção “sem informação”, que contabilizou o número de 79.144 registros. Assim, temos um percentual de 35,8% da amostra. Contudo, ao eliminarmos esses registros, a opção que ficou em destaque é referente aos pacientes que nunca fumaram, chegando a um total de 66.849 registros.

Porém, não se pode deixar de realçar o número de consumidores e ex-consumidores de tabaco, que se apresentam em 64745 registros (29,28%). Tais dados caracterizam que o brasileiro com câncer de próstata tem ou teve contato com o tabaco, que é um fator de agravo ao câncer.

No cenário brasileiro, segundo o IBGE (2020c), 16,2% da população masculina faz uso de produtos derivados do tabaco, totalizando, aproximadamente, 12 milhões, sendo a minoria. Mas, em proporção ao resultado PNS 2019, a amostra do INCA representa apenas 0,5%.

Em seguida, no Gráfico 4, verifica-se o índice de casos de câncer de próstata na família, ou seja, quanto ao fator de hereditariedade, apresentamos o que aparece nos casos registrados pelo INCA.

Gráfico 4 – Quantitativo de respostas para o quesito sobre caso de câncer na família.



Fonte: Elaborada pelo autor, 2021.

O Gráfico 4 representa os valores referentes ao quesito da existência de caso de câncer de próstata na família. Como observa-se, não foi diferente dos que já foram expostos, haja vista que a opção que mais se destaca é referente a “sem informações”, que contabiliza 113.468 registros, registrando um percentual de 51,3% da amostra.

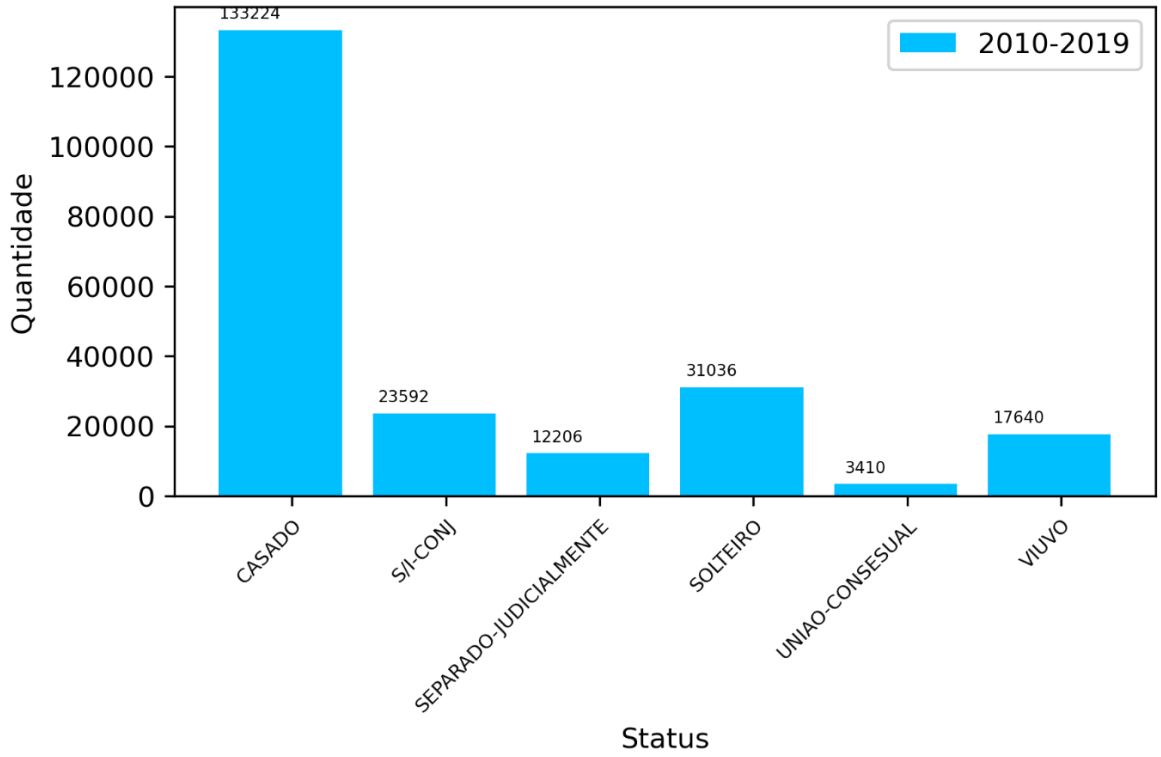
No entanto, quando excluídos os dados sobre a opção “sem informação”, as opções ficaram niveladas, havendo 60.505 (27,3%) registros para não ocorrência de câncer e 47.135 (21,3%) para ocorrência de casos na família. Assim, foi percebido o percentual a mais para a não ocorrência de câncer de próstata na família.

Apesar de Silva e Nascimento (2017), em sua pesquisa, na cidade de Parintins, no Amazonas, ter chegado ao entendimento que o histórico familiar pode aumentar de 3 a 10 vezes a possibilidade de câncer, e Rego et al. (2020), na sua pesquisa, mais de 60% dos participantes do evento avaliados possuíam histórico familiar de câncer. A realidade nacional de um país com dimensões continentais, encontra-se quase que dividida segundo os registros do INCA para o período de 2010 a 2019.

Mas também, entre outros fatores, que foram analisados, tem-se o estado conjugal de cada indivíduo, que se torna importante no tratamento da doença, uma vez que, segundo

Ferrão, Bettinelli e Portella (2017), a esposa possui um papel fundamental no apoio ao paciente, fazendo, deste modo, o papel de cuidadora e companheira do paciente.

Gráfico 5 – Quantitativo de respostas para o quesito sobre estado conjugal.

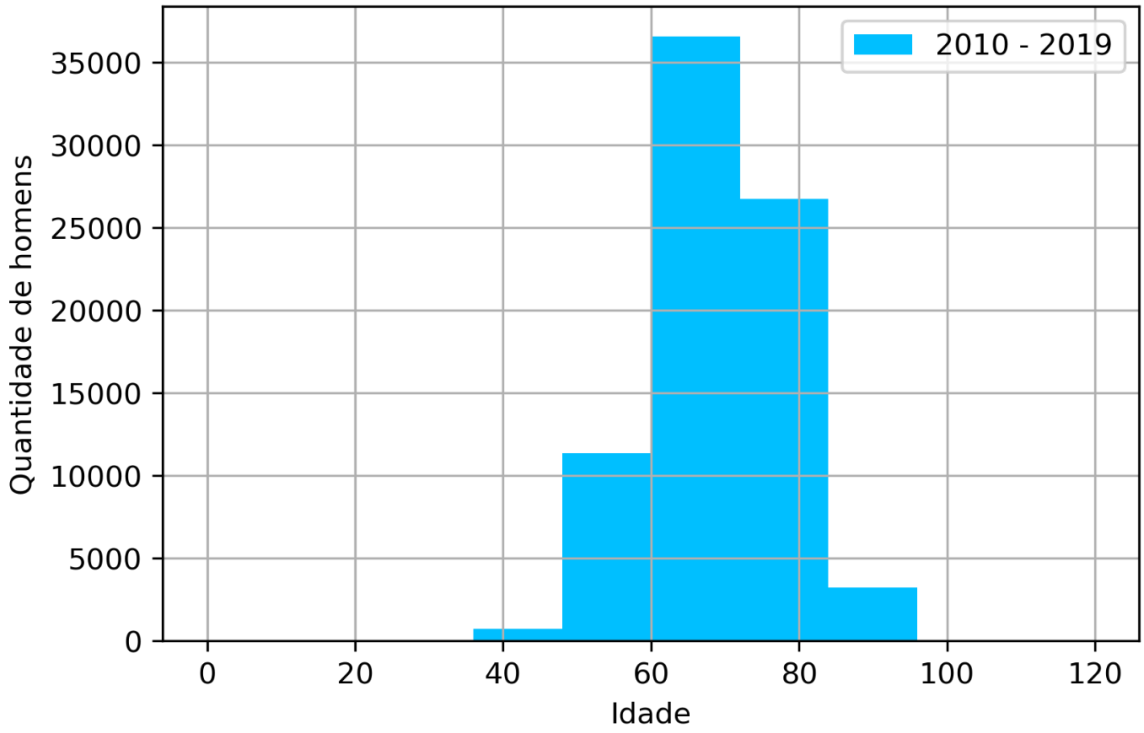


Fonte: Elaborada pelo autor, 2021.

No Gráfico 5, é exposta a situação com relação ao estado conjugal dos pacientes. Entretanto, foi percebido que, entre os campos opcionais, é o único em que a opção de "sem informação" foi menor em relação a uma parte dos dados. Essa opção ficou com um total de 23.592 registros, que representa 10,7% da amostra, enquanto a opção "casado", que apresenta o maior resultado, ficou com um valor de 133.224 registros (60,3%) da amostra. Isso só vem corroborar o resultado encontrado nos trabalhos de Araújo et al. (2015), Menezes et al. (2019), e Rego et al. (2020).

Ao considerarmos o atributo idade, no cenário brasileiro, os homens com câncer de próstata, neste período, apresentaram uma média de idade em torno de 68,66 anos; e desvio-padrão de 9,04, demonstrando um certo nível de variância desse conjunto de registros. A distribuição dos pacientes que tiveram seus dados cadastrados na base de dados do INCA se encontra exposta no histograma do Gráfico 6. Esse tipo de gráfico, de acordo com Shmueli et al. (2020), é uma das formas de expressar uma distribuição de uma população.

Gráfico 6 - Histograma das idades dos homens propensos ao câncer de próstata no período de 2010 e 2019.



Fonte: Elaborada pelo autor, 2021.

O resultado do Gráfico 6 só vem confirmar que o câncer de próstata atinge os homens com idade avançada. Ou seja, acima de 60 anos, e, muitas das vezes, é raro em homens abaixo dos 50 (CASSEL et al., 2019; MOTA; BARROS, 2019; REGO et al., 2020). Assim, percebe-se que, no período de 2010 a 2019, a idade ficou concentrada entre 60 e 80 anos.

De acordo com as projeções do IBGE (2020a) para a população brasileira masculina no ano de 2019 disponível no sítio de projeções[5], foi montada a Tabela 3 em que é disposta a proporcionalidade entre os números de casos de câncer de próstata e a população.

Tabela 3 – Proporcionalidade dos casos de câncer de próstata e a população masculina por faixa etária.

Faixa Etária (ANOS)	População 2019 (IBGE, 2020a)	Casos de Câncer de Próstata (INCA)	Proporcionalidade (%)
40 – 49	13.872.925	1260	0,0091
50 – 59	11.152.139	10.775	0,0966
60 – 69	7.431.729	29.971	0,4033
70 – 79	3.781.955	28.497	0,7533
80 – 89	1.371.137	7.572	0,5522
90 +	252.916	515	0,2036

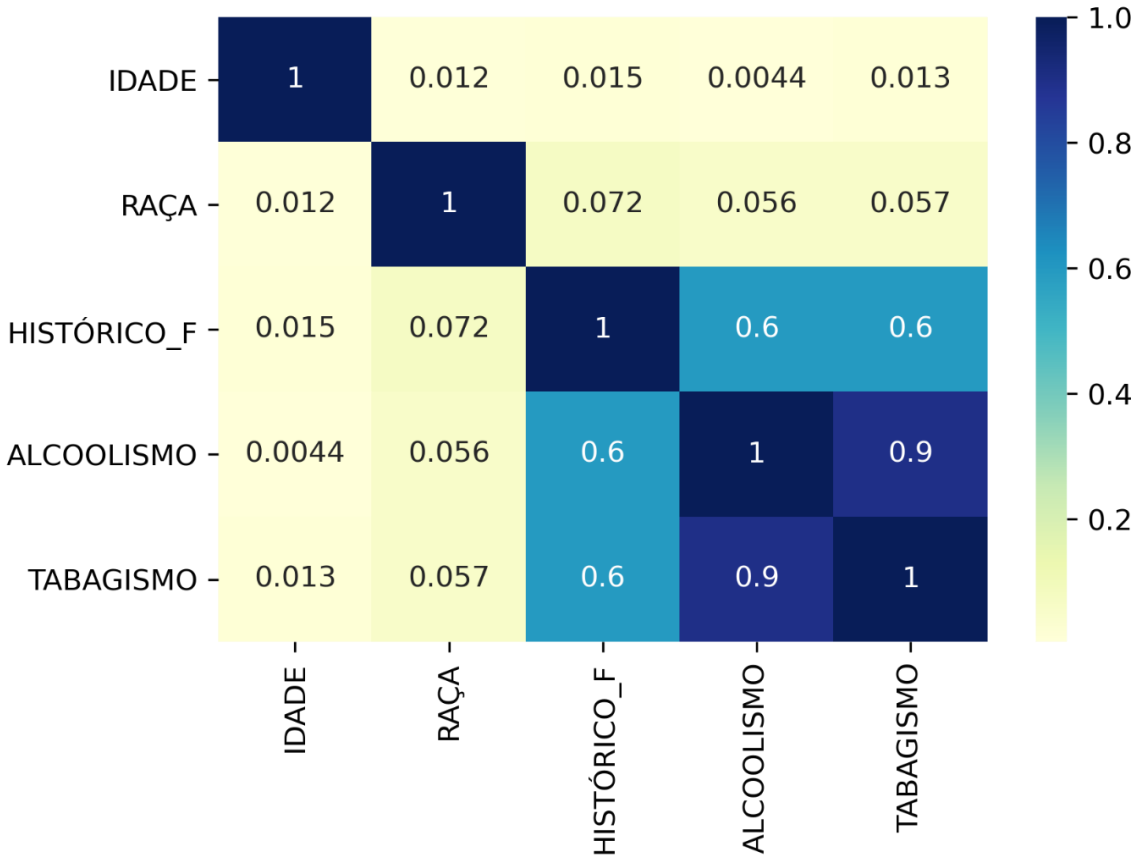
Fonte: Elaborada pelo autor, 2022.

[5]https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?edicao=21830&t=resultados

Assim, a Tabela 3 confirma que quanto maior a idade do homem, maior a possibilidade de câncer de próstata. Em conformidade com a observação feita por Mota e Barros (2019), os casos de próstata ocorrem mais nos homens acima de 60 anos.

Esses últimos cinco atributos analisados da base do INCA formam os principais fatores de risco para os homens que possuem câncer de próstata, pois podem contribuir para o surgimento e aumento das possibilidades do homem possuir esse tipo de doença durante a sua vida (SILVA; NASCIMENTO, 2017; CAVALCANTI; KRUGER, 2018).

Gráfico 7 – Mapa de calor das correlações entre os atributos idade, raça, histórico familiar, alcoolismo e tabagismo.



Fonte: Elaborada pelo autor, 2021.

A correlação mede o nível de relação entre uma variável e outra, ou seja, o quanto ao aumento de uma variável impacta no valor de aumento ou diminuição de outra variável. Além disso, o seu valor varia de -1 a +1, quanto mais próximo de -1 correlação negativa e quanto mais próximo de +1, temos correlação positiva.

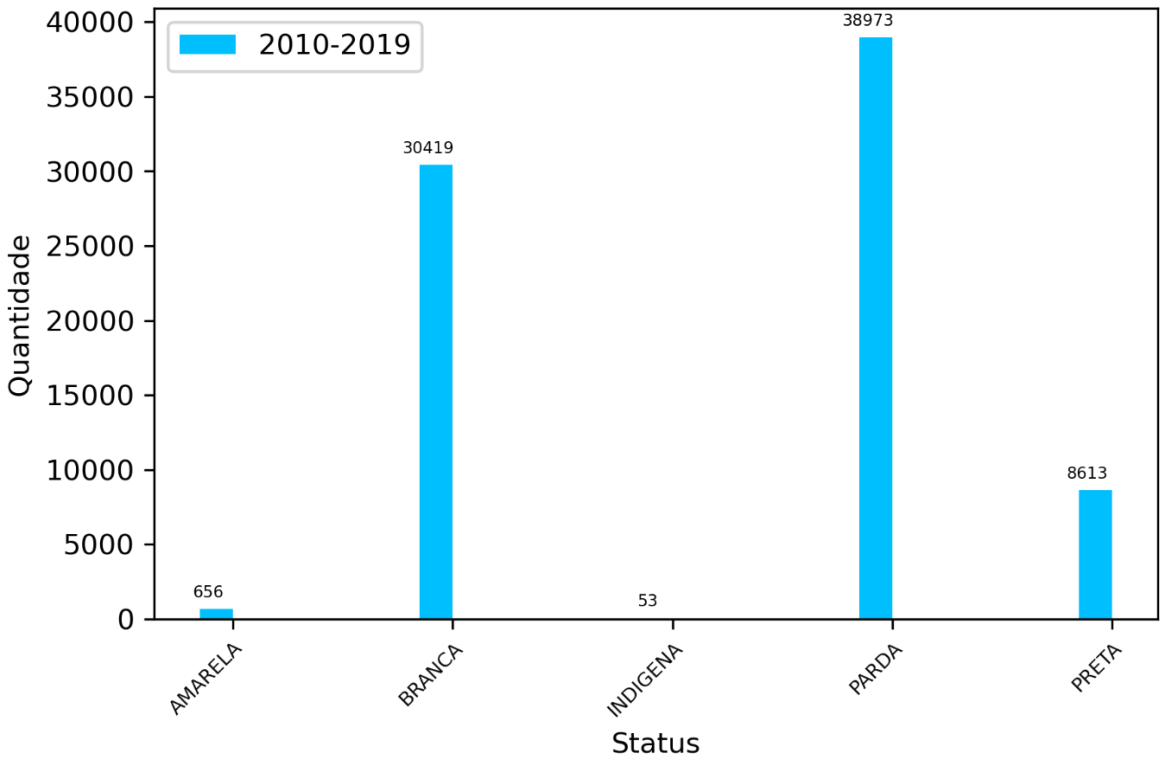
Assim, como se vê no Gráfico 7, é possível verificar que o histórico familiar, o alcoolismo, e o tabagismo apresentam uma correlação moderada positiva, no valor de 0,6.

Desse modo, pode-se depreender que esses fatores se relacionam dentro dos dados do INCA, principalmente, do tabagismo e alcoolismo, que, ao serem analisados juntos, apresentam uma correlação forte.

Além disso, Silva e Nascimento (2017), Junior et al. (2015), e Mota e Barros (2019), em suas pesquisas em três cidades diferentes do Brasil, Boa Vista-RR, Parintins-AM e Recife-PE, respectivamente, confirmam que a idade, raça, etilismo e tabagismo são fatores que influenciam no câncer de próstata.

Apesar da raça não aparecer como fator influenciador nos dados de idade, alcoolismo e tabagismo, mas é um fator de risco (ARAUJO et al., 2015). Desse modo, em seguida foram analisadas as raças dos homens encontradas na amostra, e os resultados colocados no Gráfico 8. Contudo, o resultado demonstra que a realidade do homem brasileiro apresenta um destaque para a raça branca nos casos, o que foi visualizado nos trabalhos de Araújo et al. (2015) e Menezes et al. (2019). Contudo, a raça parda foi a que mais se destacou com 38.973, isso corresponde ao percentual de 49,5% da amostra. Esse resultado é acompanhado de perto pela raça branca com o valor de 30.419 registros, percentual de 38,6%, como pode ser percebido através da leitura do Gráfico 8.

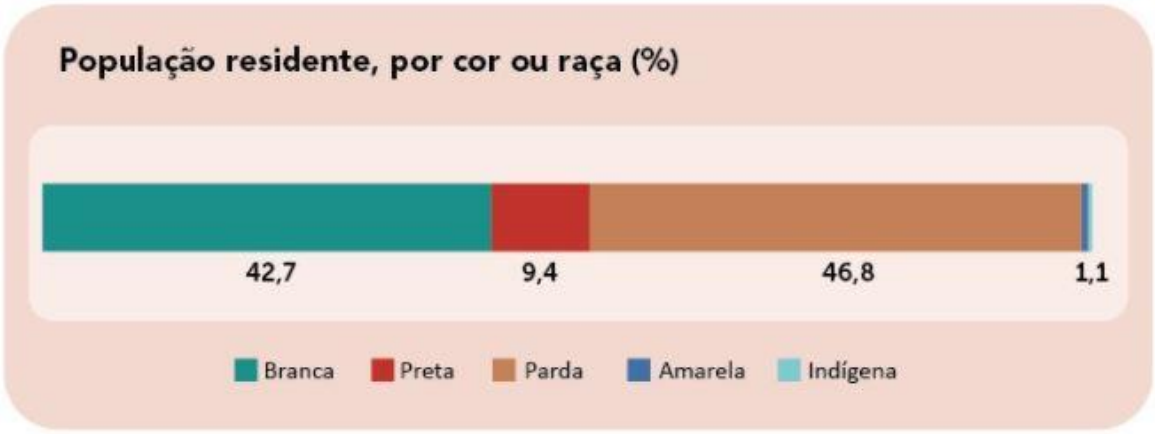
Gráfico 8 - Comparativo entre as raças dos homens com câncer de próstata.



Fonte: Elaborada pelo autor, 2021.

Deste modo, de acordo com o Gráfico 8, percebe-se que a realidade brasileira foge dos estudos que apontam a raça negra como maior incidência do câncer de próstata, uma vez que a mesma é considerada um fator de risco (MOTA; BARROS, 2019; CASSELL et al., 2019). Esse fato é compreensível, pois a maior parte da população brasileira é composta por pardos, a qual corresponde a 46,8% da população no ano de 2019, como pode ser visto no Gráfico 9 (IBGE, 2019).

Gráfico 9- População residente, por cor ou raça

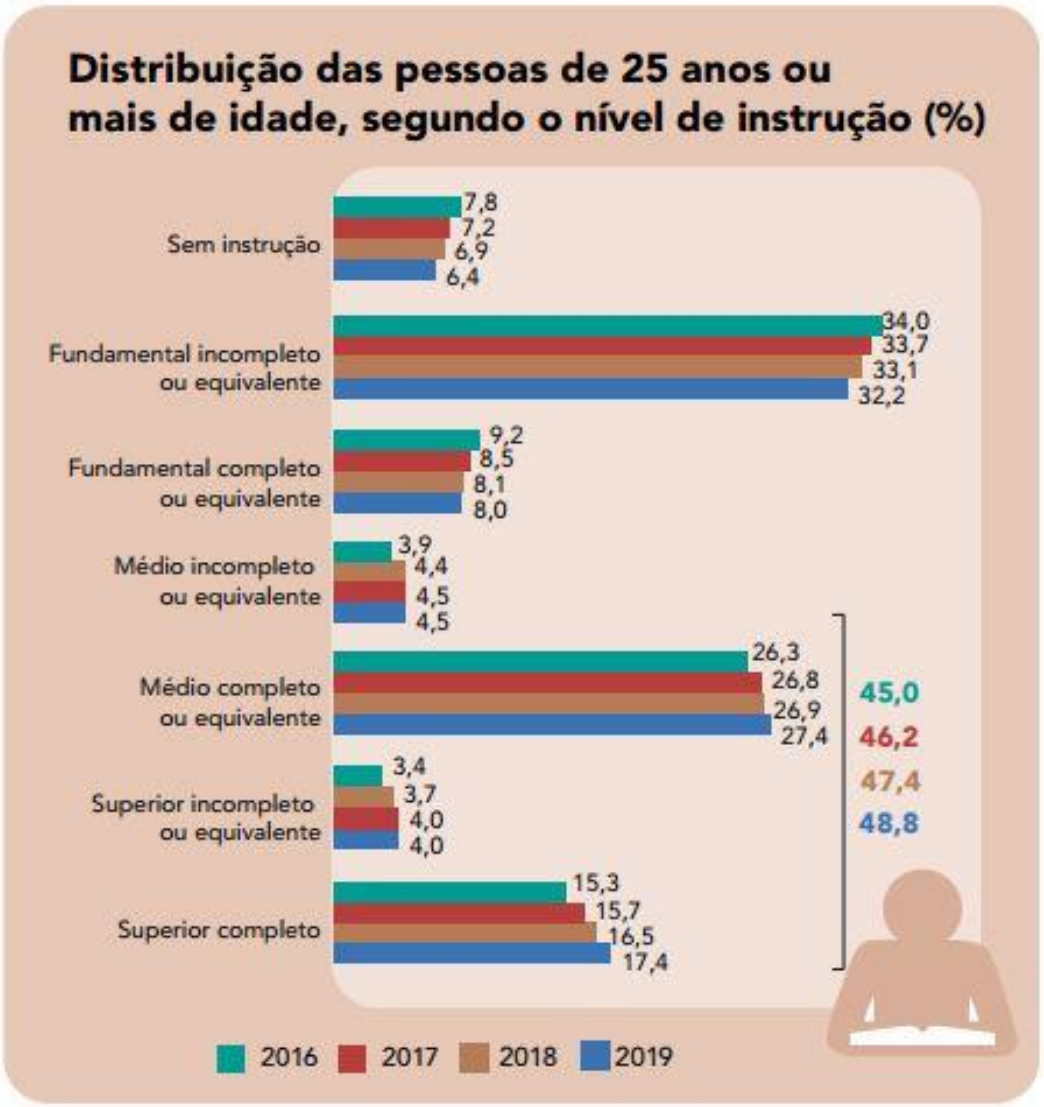


Fonte: IBGE, 2019.

No Gráfico 11, encontra-se o estudo do grau de instrução da população masculina brasileira, que possuiu câncer de próstata no período estudado. Como é notório, o número maior se deu ao fundamental incompleto, chegando ao número de 41.741 registros, que forma o percentual de 53% da amostra. Porém, um destaque tem que ser feito com relação à taxa de homens que declararam não ter formação nenhuma, 11.623 registros (14,8%), pois entende-se, deste modo, que a população masculina brasileira não possui um grau de instrução elevado. Para este cenário, de acordo com Panis et al. (2018) a baixa escolaridade é um fator socioambiental associado aos casos de câncer.

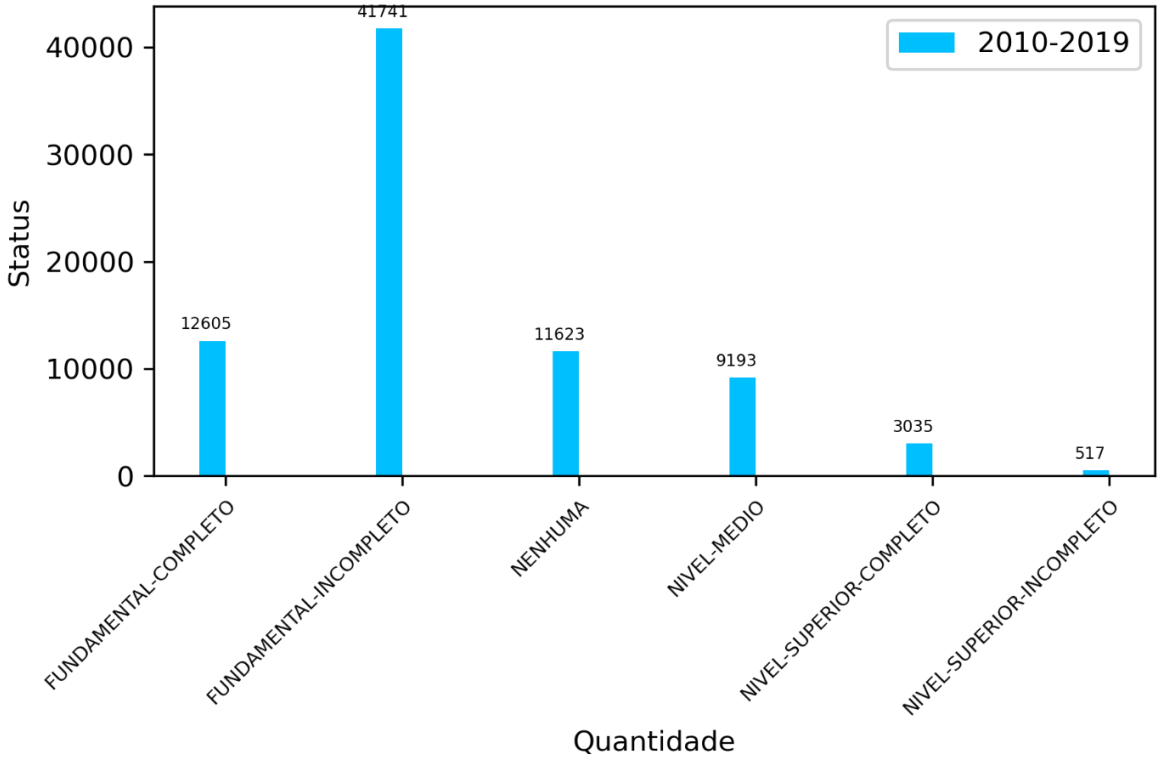
Apesar desse nível de instrução encontrado, de acordo com a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD) de 2019 com dados da educação, o brasileiro fica em média 9,4 anos estudando, a taxa de analfabetismo para as pessoas 15 anos ou mais 6,6% e sem instrução de 6,4% da população (IBGE, 2020b). Essa pesquisa apresenta dados que demonstra a diminuição nos últimos anos dos índices, como pode ser visto no Gráfico 10, contudo ainda existe um número considerável de pessoas analfabetas ou sem instrução.

Gráfico 10 – Distribuição das pessoas de 25 anos ou mais de idade, segundo o nível de instrução.



Fonte: IBGE, 2020b.

Gráfico 11 - Comparativo entre os graus de instrução dos homens com câncer de próstata.



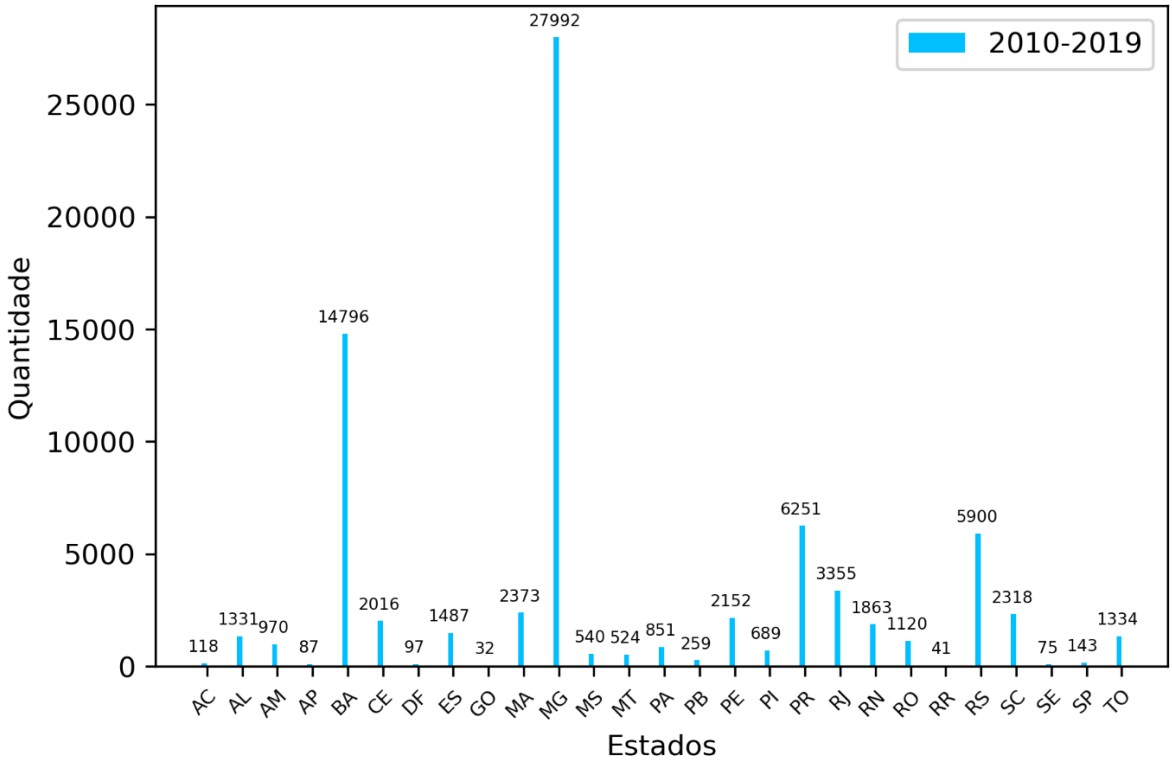
Fonte: Elaborada pelo autor, 2021.

Contudo, entendam que, apesar deste trabalho apresentar uma visão geral do Brasil, estudos de locais específicos do país já refletem este perfil como é apresentado no trabalho de Ferrão, Bettinelli e Portella (2017) e Rego et al. (2020), no qual, o primeiro analisou o atendimento de homens com câncer de próstata em um hospital, do Rio Grande do Sul, e o segundo analisou pacientes atendidos em um evento de prevenção, no estado de Minas Gerais.

Os casos de câncer de próstata vêm crescendo ano a ano no Brasil, assim se torna um caso de saúde pública, que precisa ser cuidado (KRUGER; CAVALCANTI, 2018). No Gráfico 12, pode-se notar que o Estado de Minas Gerais se destaca entre todos nessa década avaliada, chegando a 27.992 registros, que equivale a 35,6% de toda a amostra. E em segundo lugar temos o estado da Bahia com 14.796 registros (18,8%).

No âmbito regional, a região que se destaca é a região sudeste com 32.997 registros, que equivalem a 41,9% de todos os dados. Em segundo lugar, temos a região nordeste com 25.554 registros ou 32,5% dos casos. Em terceiro, temos a região sul com 14.469 registros ou 18,4%. Logo em seguida, em quarto lugar, temos a região norte com 4.521 registros, ou 5,7% da amostra. E, por último, a região centro-oeste com 1.193 registros, ou 1,5%. Essa disposição já vem acontecendo e é prevista pela análise do período de 2012 a 2016 do INCA (2020).

Gráfico 12 - Distribuição das ocorrências de casos de próstata no Brasil no período de 2010 a 2019.



Fonte: Elaborada pelo autor, 2021.

Ademais, as maiores proporcionalidades entre a população e os casos de câncer de próstata ainda destacam os Estados de Minas Gerais e Bahia conforme a Tabela 4. Todavia, é possível perceber que os Estados do Tocantins, Rondônia, Paraná, Rio Grande do Norte, e Rio Grande do Sul estão entre as maiores proporcionalidades.

Tabela 4 - Proporcionalidade dos casos de câncer de próstata e a população masculina por Estado.

Estado	Projeção de População de 2019 (IBGE, 2020a)	Casos de Câncer de Próstata (INCA)	Proporcionalidade (População/Casos) (%)
AC	441.216	118	0,0267
AL	1.601.112	1331	0,0831
AM	2.081.262	970	0,0466
AP	423.498	87	0,0205
BA	7.233.509	14.796	0,2045
CE	4.432.035	2016	0,0455
DF	1.447.284	97	0,0032
ES	1.978.483	1487	0,0752
GO	3.481.598	32	0,0009
MA	3.479.859	2373	0,0682
MG	10.422.468	27.992	0,2686
MS	1.379.290	540	0,0391

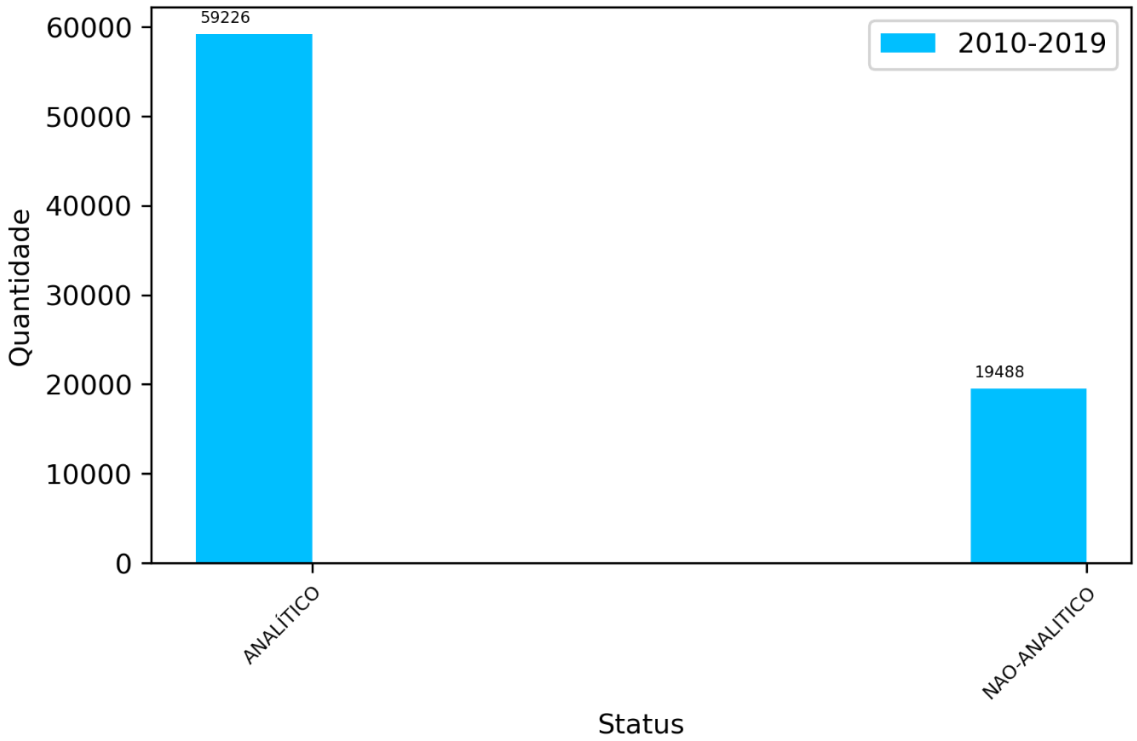
MT	1.767.091	524	0,0296
PA	4.315.587	851	0,0197
PB	1.939.480	259	0,0133
PE	4.588.206	2152	0,0469
PI	1.586.538	689	0,0434
PR	5.602.812	6.251	0,1116
RJ	8.255.502	3.355	0,0406
RN	1.709.856	1.863	0,1089
RO	899.108	1.120	0,1246
RR	312.110	41	0,0131
RS	5.536.738	5.900	0,1065
SC	3.554.814	2.318	0,0652
SE	1.110.281	75	0,0067
SP	22.388.135	143	0,0006
TO	792.423	1.334	0,1683

Fonte: Elaborada pelo autor, 2022

Quando os pacientes procuram um hospital ligado ao RHC, o seu caso pode ser classificado como caso analítico ou não analítico. No caso analítico, o hospital vinculado ao RHC diagnosticou e realizou o tratamento, ou acompanhou o tratamento do paciente. Mas, no caso não analítico, o diagnóstico e o tratamento muitas vezes não são realizados pelo hospital vinculado ao RHC, ou o paciente faleceu nas primeiras 48 após a matrícula no hospital (INCA, 2010).

No Gráfico 11, tem-se a informação de que os casos analíticos se destacam mais do que os casos não analíticos na rede brasileira. Mas, deve-se dar ênfase ao fato de que a classificação depende do atendente que a realiza na primeira entrevista do paciente (INCA, 2010).

Gráfico 11 - Classificação dos tipos de casos dos pacientes no período de 2010 e 2019.



Fonte: Elaborada pelo autor, 2021.

Apesar desses dados iniciais já retornarem informações que passam a construir o perfil do brasileiro com o câncer de próstata, ainda existem muitos dados ainda para serem analisados, uma vez que, segundo Frank et al. (2016), é no meio dos dados que se encontram informações importantes que raramente são reveladas.

Desse modo, com o objetivo de revelar o conhecimento intrínseco dentro da base do INCA, logo após o levantamento dos dados anteriormente analisados, foi executado o algoritmo *Apriori* para a formação das regras de associação e para descobrir todo o conhecimento que possa ainda estar faltando ser demonstrado.

O suporte foi o primeiro índice definido com o valor de 0,129, e de acordo com a Equação IV, uma vez que o total de casos de câncer no período de 2010 a 2019 foi de 1.844.810 registros, e, entre eles, os referentes ao câncer de próstata ficaram com 238.795 registros. Além disso, a confiança ficou variando de 0,3 a 0,8, e, por último, o *lift* mínimo de 1,2. Na implementação *Apyori*, o total de elementos mínimos na regra também podem ser definidos. Assim foi definido que as regras deveriam ser formadas por no mínimo 2 itens, pois as regras são formadas no padrão *X* então *Y*. Logo, garante-se que as regras sempre serão formadas por antecessores e consequentes.

Outrossim, destaca-se os principais resultados encontrados com as regras que mais se repetiram, apresentando um grau de confiança acima de 30%, e não apresenta a ausência de informação no quesito, ou seja, não foi envolvido a opção sobre o local em que o atendente ou o paciente declarou “sem informação”. Dessa maneita, a Tabela 5 apresenta os resultados do período estudado, em que se tem representado por “c” o índice de confiança e o *lift* da regra encontrada.

Tabela 5 - Regras de associação com dados do período de 2010 a 2019.

1. Mora na Bahia, então nunca fumou, c=0,71, lift: 1,367
2. Mora na Bahia, então é da raça parda, c=0,78, lift: 1,587
3. Ex-consumidor de álcool, então é ex-consumidor de tabaco, c=0,68, lift=2,476
4. Ex-consumidor de tabaco, então é ex-consumidor de álcool, c=0,49, lift=2,476
5. Nunca bebeu, então nunca fumou, c=0,74, lift=1,430
6. Nunca fumou, então nunca bebeu, c=0,75, lift=1,430
7. Nunca bebeu e o caso é analítico, então nunca fumou, c=0,74, lift=1,427
8. Caso analítico e nunca fumou, então nunca bebeu, c=0,75, lift=1,424
9. Nunca bebeu e a raça é branca, então nunca fumou, c=0,70, lift=1,342
10. Raça branca e nunca fumou, então nunca bebeu, c=0,77, lift=1,456
11. Nunca bebeu e casado, então nunca fumou, c=0,74, lift=1,431
12. Casado e nunca fumou, então nunca bebeu, c=0,76, lift=1,440
13. Possui fundamental incompleto e nunca bebeu, então nunca fumou, c=0,73, lift=1,413
14. Possui fundamental incompleto e nunca fumou, então nunca bebeu, c=0,77, lift=1,457
15. Nunca bebeu e não possui histórico familiar, então nunca fumou, c=0,76, lift=1,459
16. Nunca fumou e não possui histórico familiar, então nunca bebeu, c=0,77, lift=1,478
17. Nunca bebeu e a raça é parda, então nunca fumou, c=0,77, lift=1,493
18. Raça parda e nunca fumou, então nunca bebeu, c=0,75, lift=1,419
19. Nunca bebeu e possui histórico familiar, então nunca fumou, c=0,72, lift=1,390
20. Possui histórico familiar e nunca fumou, então nunca bebeu, c=0,72, lift=1,364
21. Nunca bebeu, casado e o caso é analítico, então nunca fumou, c=0,74, lift=1,425
22. Casado, o caso analítico e nunca fumou, então nunca bebeu, c=0,75, lift=1,433
23. Possui fundamental incompleto, nunca bebeu e o caso é analítico, então nunca fumou, c=0,73, lift = 1,410
24. Possui fundamental incompleto, caso analítico e nunca fumou, então nunca bebeu, c=0,76, lift=1,445
25. Nunca bebeu, caso analítico, e não possui histórico familiar, então nunca fumou, c=0,75, lift=1,452
26. Nunca fumou, caso analítico e não possui histórico familiar, então nunca bebeu, c=0,77, lift=1,467
27. Nunca bebeu, caso analítico e raça parda, então nunca fumou, c=0,77, lift=1,485
28. Raça parda, caso analítico e nunca fumou, então nunca bebeu, c=0,74, lift=1,406
29. Possui fundamental incompleto, nunca bebeu e casado, então nunca fumou, c=0,73, lift=1,412
30. Possui fundamental incompleto, casado e nunca fumou, então nunca bebeu,

c=0,77, lift=1,473 31. Nunca bebeu, casado, não possui histórico familiar, então nunca fumou, c=0,76, lift=1,467 32. Nunca fumou, casado e não possui histórico familiar, então nunca bebeu, c=0,78, lift=1,493

Fonte: Elaborada pelo autor, 2021.

Além do que foi caracterizado nos parágrafos iniciais, as regras de associação apresentaram dados envolvendo Estado de residência, estado conjugal, grau de instrução, raça, alcoolismo, tabagismo, presença de histórico familiar e tipo de caso. Nesses casos, a variação da confiança alterou apenas na exclusão das regras 3, 4 e 9 presentes na Tabela 5 à medida que definimos um nível maior.

Ademais, pode-se destacar alguns pontos dentre as regras. O primeiro ponto destacado diz respeito ao termo de consumo de álcool, pois a maioria declara que nunca bebeu. Desse modo, nas regras geradas com relação a esta opção, houve uma maior significância na regra 32 da Tabela 5. Essa regra apresenta uma confiança de 78% e um *lift* de 1,493; esses dados garante que esse consequente (nunca bebeu) é mais frequente em regras contendo os antecedentes de nunca ter fumado, ser casado e não apresentar histórico familiar dentro da base de dados estudada.

Esses dados só confirmam mais ainda a pesquisa realizada por Menezes et al. (2019), na qual aplicaram um questionário em que um dos quesitos traz a afirmativa de que é falsa segundo os autores, “As bebidas alcoólicas contribuem para o desenvolvimento do câncer de próstata. ” (MENEZES et al., 2020, p. 1176).

Além disso, outro fator interessante que apareceu várias vezes no consequente das regras foi o fator de consumo de tabaco, em que os pacientes, em sua maioria, declararam que nunca usaram. Tal consequente sofreu mais influência, de acordo com as regras geradas, quando os antecedentes da regra eram “nunca bebeu e a raça é parda”. Essa regra apresentou o índice de confiança de 77% e *lift* de confiança de 1,493, como pode ser visto na regra 17. Isso no cenário de Montes Claros, em Minas Gerais, já ocorria, como é exposto na pesquisa feita por Rego et al. (2020), segundo o qual, 52,6% dos participantes do estudo declararam ser não fumantes.

Esses dois últimos fatores analisados estão, frequentemente, associados quando ocorrem, assim podemos dizer que a ausência de consumo também estaria (STONE et al., 2019).

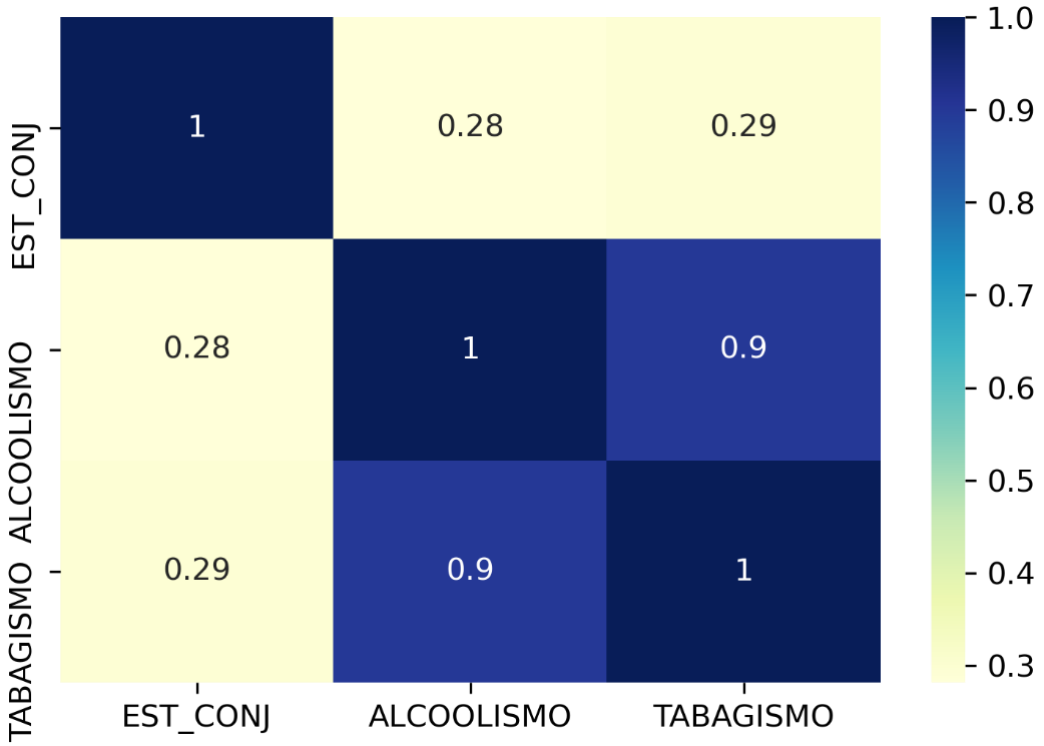
Quando se considera o histórico familiar, pode-se concluir, de acordo com as regras geradas, que a maior parte é influenciada por não possuir histórico familiar, cenário que foi

encontrado também nos trabalhos de Rego et al. (2020) e Stone et al. (2019). Mas, nota-se que ainda no meio dos dados estudados, há influência do histórico familiar, haja vista que apareceram duas regras ligadas ao consumo de álcool e tabaco, como pode ser visto nas regras 19 e 20. Isso é validado pela correlação positiva moderada, que ficou demostrada no Gráfico 7.

Não só esses últimos fatores influenciaram nos dados, mas também o estado conjugal dos pacientes interferiu no resultado das regras de associação. Esse item ratificou que a maioria dos homens são casados, e se relaciona, na maioria das regras, com o consumo de álcool e tabaco na formação das mesmas. Porém, como podemos ver no Gráfico 12, a correlação entre esses três fatores é fraca.

Ademais, quando se envolve o estado conjugal, é importante destacar o apoio familiar ao paciente, que, segundo Ferrão, Bettinelli e Portella (2017), é importante a participação da esposa, pois ela é a pessoa que irá auxiliar o paciente no tratamento da doença.

Gráfico 12 – Correlação dos atributos estado conjugal, alcoolismo e tabagismo.



Fonte: Próprio Autor, 2021.

De acordo com as regras geradas, a raça parda como antecedente, ocorre entre as regras que possuem um dos maiores *lift*, como pode ser visto na regra 17. Ou seja, essa raça possui um nível de influência maior. Assim, depreende-se que a raça parda é a que mais se

destaca entre os pacientes com câncer de próstata. Além do mais, destaca-se que os pacientes da Bahia eram pardos, isso com uma confiança de 78%.

Esse resultado para a raça diverge do que ocorre no mundo, onde se dá destaque de ocorrência de câncer de próstata em indivíduos da raça negra. Esse fato ocorreu também na pesquisa de Araújo et al. (2015) e Stone et al. (2019)

Seguindo a análise detalhada das regras, a única característica quanto ao grau de instrução que se destaca entre os antecedentes é a opção de "fundamental incompleto", que, entre as regras, encontra-se em uma que apresenta um dos maiores índices de *lift*, como pode ser visto na regra 30, haja vista que pesquisas, no território nacional, já demonstrava que os brasileiros com câncer de próstata não possuíam nível de escolaridade maior do que o fundamental completo (FERRÃO; BETTINELLI; PORTELLA, 2017; MENEZES et al., 2019; REGO et al., 2020)

Apesar de que os dados brutos destacaram o estado de Minas Gerais como residência da maioria dos pacientes, o estado que apareceu dentro das regras foi o estado da Bahia, assim, depreendeu-se que os baianos que possuíam câncer de próstata eram pardos e declararam que nunca fumaram.

Portanto, a partir das regras obtidas e da moda das confianças que formam as regras, pode-se formar o perfil sociodemográfico do homem brasileiro com câncer de próstata no período de 2010 a 2019. Esse perfil ficaria como sendo de homens que nunca fumaram, com confiança de 74%; nunca beberam, com confiança de 77%; de raça parda, em sua maioria, com a confiança de 77%; casados, de quem destacamos a confiança de 76%; com nível de ensino de fundamental incompleto, cujo dado aparece com confiança de 73%; e que não possuem histórico familiar da doença, com confiança de 77%.

Além disso, como a maioria dos casos foram considerados analíticos, entende-se que o atendimento na rede pública de hospitais de oncologia prevalece em relação à rede privada no tratamento de câncer de próstata.

5 CONCLUSÕES

A ausência de um perfil do homem com o câncer de próstata dificulta o diagnóstico, pois o câncer desse tipo é assintomático no começo (MOTA; BARROS, 2019). Além disso, é importante a definição desse perfil, haja vista a possibilidade de facilitar o diagnóstico e o encaminhamento do tratamento.

Não só há a escassez de literatura que trate do desenvolvimento de técnicas que possam definir esse perfil, mas também há o agravante da grande quantidade de dados armazenada em bases dados que dificulta o estudo e a análise por pessoas (OLIVEIRA, et al., 2007; ARAUJO et al., 2015).

Deste modo, este trabalho uniu os registros armazenados na base de dados do INCA com homens diagnosticados com câncer de próstata, para poder traçar o perfil sociodemográfico do homem brasileiro com câncer de próstata no período de 2010 a 2019 através do uso do algoritmo *Apriori*. Assim, através da análise dos dados, formaram-se as regras de associação para ser projetado o possível perfil do homem propenso a esse tipo de câncer.

Com os gráficos e regras geradas, ao final do processo de aplicação do algoritmo *Apriori*, foi percebido que os fatores de tabagismo, alcoolismo, raça e estado conjugal são os que mais se destacaram por aparecerem nas regras com os maiores índices de confiança. Porém, fugindo um pouco da literatura encontrada, que dá o destaque à raça negra como maior incidência de casos, a realidade brasileira é destacada pelos casos na raça parda.

Ademais, o destaque dos fatores de tabagismo e de alcoolismo declarados no preenchimento das fichas dos pacientes, em sua maioria, como nunca beberam ou nunca fumaram apresentam uma forte correlação. Assim, resultados demonstram que esses dois tópicos são bastantes interligados.

Apesar da incompletude dos dados opcionais, na base de registros do INCA, é importante destacar que a análise foi feita a nível nacional e pode ser utilizada para definição de campanhas de informação e acompanhamento dos homens com câncer de próstata. Assim, torna-se um instrumento norteador no contexto da saúde do homem no Brasil.

5.1 Dificuldades encontradas

Para elaboração deste estudo, surgiram algumas dificuldades que atrasaram o andamento das etapas do projeto. Além disso, houve dificuldades para elaboração dos resultados finais.

O primeiro empecilho ocorreu quando foi utilizado o portal Integrador do RHC para baixar os dados que seriam estudados, pois os dados vieram todos quebrados em um arquivo para cada ano, não permitindo, deste modo, ter-se uma visão geral dos dez anos de dados.

O segundo ponto de dificuldade encontrado deu-se quando foi verificado que os arquivos exportados pela ferramenta do INCA vieram no formato DBF, que, para aplicarmos a metodologia elaborada, foi necessária conversão dos arquivos para o formato CSV.

O terceiro ponto de dificuldade ocorreu por falta de parâmetros para definir a confiança, haja vista que, na literatura, não foi encontrado apenas a indicação do valor, que varia de acordo com o caso estudado (SILVA, 2016). Assim, foram necessárias várias execuções do algoritmo em fase de treinamento para poder definir os valores, aumentando o tempo para aplicação do algoritmo nos dados do estudo.

E, por último, uma comparação com dados do censo de 2020 não foi possível devido à não realização do referido censo, uma vez que o novo coronavírus impediu que o IBGE realize o mesmo no seu tempo normal.

5.2 Trabalhos futuros

Com o propósito de um aprofundamento sobre o câncer de próstata no Brasil, pretende-se, para trabalhos futuros, desenvolver pesquisas que possam garantir a análise de dados em comparativo com a base de registro de câncer de base populacional – RCBP, também mantido pelo INCA; Aplicar análise de Componentes Principais (ACP) para constatar o nível de influência entre os fatores de risco cor/raça, tabagismo, alcoolismo e histórico familiar; e implementar algoritmos de previsão nos dados da base do INCA para assim auxiliar os profissionais da saúde no diagnóstico e tratamento da doença.

Outrossim, devido à não realização do censo 2020 pelo IBGE no período normal por motivos do distanciamento social, que era uma das medidas de contenção do novo coronavírus, outro trabalho que ficará para o futuro será a realização da parametrização dos dados INCA com os dados do novo censo feito pelo IBGE em 2022.

REFERÊNCIAS

AGRAWAL, R; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. **ACM SIGMOD Record**, New York, v. 22, n. 2, p. 207-216, 1993.

AGUIAR, J. S.; PEREIRA, L. D. A.; THOMAS, C. A. B. Assistência Oncológica no Estado do Espírito Santo, a partir do Sistema Integrador dos Registros Hospitalares de Câncer; 2000 – 2014. **Informativo Vigilância do Câncer**, Espírito Santo; 2017. Disponível em: <https://saude.es.gov.br/Media/sesa/DANTS/INFORMATIVO%20VIGILANCIA%20DO%20CANCER%20-%20RHC%2020%2012%202017.pdf>. Acesso em: 20 jan. 2021.

ARAÚJO, T. S.; OLIVEIRA, T. P. S.; SILVA, E. R. G. Sistemas Inteligentes de Apoio à Tomada de Decisão na Gestão Pública Municipal: Uma Abordagem Conceitual. In: Conferência Sul-Americana em Ciência e Tecnologia Aplicada ao Governo Eletrônico, 4., 2007, Palmas. **Anais**... Florianópolis: Digital Ijuris, 2007. p. 17-29.

ARAÚJO, J. S. et al. Caracterização social e clínica dos homens com câncer de próstata atendidos em um hospital universitário. **Revista Mineira de Enfermagem**, Belo Horizonte; V. 19.2, 2015.

BALDOMIR, R. A. **Aplicação do Algoritmo Apriori para Detectar Relacionamento entre Empresas nos Processos Licitatórios do Governo Federal. Brasília**, 2017. Disponível em: <https://bdm.unb.br/handle/10483/19987>. Acesso em: 24 Ago. 2021.

BRASIL. Ministério da Saúde. Portaria nº 3535/GM. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 1998. Disponível em:< https://bvsms.saude.gov.br/bvs/saudelegis/gm/1998/prt3535_02_09_1998_revog.html>. Acesso em: 05 jan. 2021.

BRASIL. Ministério da Saúde. Portaria nº 741/MS/SAS. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 2005. Disponível em:< https://bvsms.saude.gov.br/bvs/saudelegis/sas/2005/prt0741_19_12_2005.html>. Acesso em: 05 jan. 2021.

CAMILO, C. O.; SILVA, J. C. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Goiânia, 2009.

CAMPBELL, A. **Data Visualization Guide: Clear Introduction to Data Mining, Analisys and Visualization**. [S.I]: [S.N.], 2021. 56p.

CARVALHO, D. R.; DALLAGASSA, M. R.; SILVA, S. H. Uso de Técnicas de Mineração de Dados para a Identificação Automática de Beneficiários Propensos ao Diabetes Mellitus Tipo 2. **Inf. Inf**., Londrina, v. 20, n. 3, p. 274 – 296, set./dez. 2015. Disponível em: < http://dx.doi.org/10.5433/1981-8920.2015v20n3p274>. Acesso em: 10 mai. 2021.

CASSELL, A. et al. A Review of Localized Prostate Cancer: Na African Perspective. **World Journal Of Oncology**. V. 10, 2019. Disponível em: < https://doi.org/10.14740/wjon1221 >. Acesso em: 05 jan. 2021.

CAVALCANTI, G.; KRUGER, F. P. G. Conhecimento e Atitudes sobre o Câncer de Próstata no Brasil: Revisão Integrativa. **Revista Brasileira de Cancerologia**, Rio de Janeiro, V. 64, n. 4, p. 561-567. 2018. Disponível em: < https://doi.org/10.32635/2176-9745.RBC.2018v64n4.206 >. Acesso em: 05 jan. 2021.

EFFIOK, E. E.; LIU, E.; HITCHCOCK, J. Analyse Lifestyle Related Prostate Cancer Risk Factors Retrieved from Literacy. In: IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber. **Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)**, 2017. Disponível em: <https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.174>. Acesso em: 17 Jul. 2020.

FERRÃO, V.; BETTINELLI, L. A.; PORTELLA, M. R. Vivências de Homens com Câncer de Próstata. **Revista de Enfermagem UFPE On Line**, Recife; V. 11, 2017. Disponível em: <https://periodicos.ufpe.br/revistas/revistaenfermagem/article/download/231178/25153> . Acesso em: 10 jan. 2021.

FEUSER, R. J. **Mineração de Dados com Regras de Associação Aplicada em Dados de Unidade Saúde de Pronto Atendimento**. Pato Branco, 2017. Disponível em: < http://repositorio.utfpr.edu.br/jspui/handle/1/22182>. Acesso em: 05 jan. 2021.

FILHO, F. M. **Regras de Associação para Análise e Casos de Dengue em Municípios Paraibanos**. 2017. 72 f. Dissertação (Mestrado Ciência e Tecnologia em Saúde) Universidade Estadual da Paraíba, Campina Grande, 2017.

FRANK E.; HALL, M. A.; WITTEN, I. H. **The WEKA Workbench**. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition, 2016.

FURLAN, M. B. Algoritmos e Técnicas para Mineração de Dados. In: Assis. **Mineração de dados**, 2018. Disponível em: < https://cepein.femanet.com.br/BDigital/arqTccs/1511420203.pdf>. Acesso em: 05 jan. 2021.

GOLDSCHIMIDT, R; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GONÇALVES, I. R.; PADOVANI, C.; POPIM, R. C. Caracterização epidemiológica e demográfica de homens com câncer de próstata. **Associação Brasileira de Ciência e Saúde Coletiva**, Rio de Janeiro, n 13, p. 1337-1342, 2008. Disponível em: <https://doi.org/10.1590/S1413-81232008000400031>. Acesso em: 05 jan. 2021.

GUISI, D. et al. Técnicas de Classificação em Problemas Relacionados a Doenças Cardíacas. In: Workshop de Pesquisa em Computação dos Campos Gerais, 16., 2016. **Anais eletrônicos**... Ponta Grossa: UTFP, 2016. Disponível em:< http://www.wpccg.pro.br/volumes/volume001/proceedings/wpccg-2016>. Acesso em: 10 mai. 2021.

HERMALIANI, E. H. et al. Data Mining Technique to Determine the Pattern of Fruits Sales & Supplies Using Apriori Algorithm. **Journal of Physics: Conference Series**, 2020. Disponível em: <doi:10.1088/1742-6596/1641/1/012070>. Acesso em: 27 fev. 2021.

HUSSAIN, L. et al. Applying Bayesian Network Approach to Determine the Association Between Morphological Features Extracted from Prostate Cancer Images. **IEEE Access**, V. 7. 2019. Disponível em: < https://ieeexplore.ieee.org/document/8579592 >. Acesso em: 05 jan. 2021.

Instituto Brasileiro de Geografia e Estatísticas (IBGE). **Conheça o Brasil – População Cor ou Raça**. Rio de Janeiro, 2019. Disponível em: < https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18319-cor-ou-raca.html>. Acesso em: 23 mar. 2022.

Instituto Brasileiro de Geografia e Estatísticas (IBGE). **Projeção da população do Brasil por sexo e idade para o período 2010-2060**. Rio de Janeiro, 2020a. Disponível em:< https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?edicao=21830&t=resultados>. Acesso em: 23 mar. 2022.

Instituto Brasileiro de Geografia e Estatísticas (IBGE). **Educação: 2019**. Rio de Janeiro: IBGE, 2020b. Disponível em: < https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf>. Acesso em: 23 mar. 2022.

Instituto Brasileiro de Geografia e Estatísticas (IBGE). **Pesquisa nacional de saúde: 2019: percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal: Brasil e grandes regiões / IBGE, Coordenação de Trabalho e Rendimento**. Rio de Janeiro: IBGE, 2020c. Disponível em: < https://www.pns.icict.fiocruz.br/wp-content/uploads/2021/02/liv101764.pdf>. Acesso em: 23 mar. 2022.

Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA). **Registros Hospitalares de Câncer: Planejamento e Gestão**. 2ed. Rio de Janeiro: INCA, 2010.

Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA). Estimativa 2018 – Incidência de Câncer no Brasil. **Rev Bras Cancerol**. 2018.

Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA). **Informativo Vigilância do Câncer: Perfil da Assistência Oncológica no Brasil entre 2012 e 2016**. Rio de Janeiro: INCA, 2020. Disponível em: < https://irhc.inca.gov.br/RHCNet/>. Acesso em: 20 jun. 2021.

Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA). **Integrador RHC – Registro Hospitalar de Câncer**. Rio de Janeiro: INCA, 2021. 1 base de dados. Disponível em: < https://irhc.inca.gov.br/RHCNet/>. Acesso em: 10 mai. 2021.

JAMSA, K. **Introduction to Data Mining and Analytics with Machine Learning in Rand Python**. Burlington: World Headquartes Jones & Bartlett Learning, 2021.

JUNIOR, M. M. L. et al. Unraveling Brazilian Indian Population Prostate Good Health: Clinical, Anthropometric and Genetic Features. **IBJU**, Rio de Janeiro, v. 41, abr. 2015. Disponível em: <10.1590/S1677-5538.IBJU.2015.02.23 >. Acesso em: 10 mai. 2021.

KLOSTERMAN, S. **Projetos de Ciência de Dados com Python**. Tradução de Aldir Coelho Côrrea da Silva. São Paulo: Novatec, 2019.

MEHTA A.; BURA D. Mining of Association Rules in R Using Apriori Algorithm. In: Hura G., Singh A., Siong Hoe L. (eds) Advances in Communication and Computational Technology. **Lecture Notes in Electrical Engineering**, vol 668. Springer, Singapore, 2020 Disponível em: <https://doi.org/10.1007/978-981-15-5341-7_14>. Acesso em: 27 fev. 2021

MENEZES, R. et al. Knowledge, Behaviour and Health Practices of Men Concerning the Prostate Cancer. **Revista Online de Pesquisa Cuidado é Fundamental**, v. 11. 2019. Disponível em: <http://www.seer.unirio.br/index.php/cuidadofundamental/article/download/7001/pdf>. Acesso em: 05 Jan. 2021.

MIRHASHEMI, S.H.; MIRZAEI, F. Extracting association rules from changes in aquifer drawdown in irrigation areas of Qazvin plain, Iran. **Groundwater for Sustainable Development**, v 12. 2021. Disponível em:< https://doi.org/10.1016/j.gsd.2020.100495>. Acesso em: 05 jan. 2021.

MOTA, T. R.; BARROS, D. P. O. Perfil dos Pacientes com Câncer de Próstata em Hospital de Referência no Estado de Pernambuco. **Revista Brasileira de Análises Clínicas**, Recife; 2019.

NOTARI, I. A. C. L. M. **Gerenciamento de Projetos: Aplicação do Planejamento no Setor de Óleo e Gás**. 2017. 120 f. Trabalho de Conclusão de Curso de Engenharia Química. Universidade Federal Fluminense, Niterói, 2017.

PANIS, C. et al. Revisão crítica da mortalidade por câncer usando registros hospitalares e anos potenciais de vida perdidos. **Einstein**, São Paulo, v. 16, n. 1, 2018. Disponível em: < https://doi.org/10.1590/S1679-45082018AO4018>. Acesso em: 15 jan. 2021.

OLIVEIRA, M. C.; MARQUES-AZEVEDO, P. M.; FILHO, W. C. C. Grades Computacionais na Otimização da Recuperação de Imagens Médicas Baseadas em Conteúdo. **Revista Radiologia Brasileira**, v. 40, 255-261; 2007.

PREISSLER, A. **Data Mining para Definição dos Perfis de Pacientes com Câncer de Estômago**. 2016. 62 f. Trabalho de Conclusão de Curso de Ciência da Computação, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Santa Rosa, 2016.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. Novo Hamburgo: Feevale, 2013. Disponível em: < https://www.feevale.br/cultura/editora-feevale/metodologia-do-trabalho-cientifico---2-edicao >. Acesso em: 20 jan. 2014.

REGO, R. F. N. B. et al. Perfil Clínico Epidemiológico da População Atendida Num Programa de Rastreamento de Câncer de Próstata. **Revista de Atenção à Saúde**, São Caetano do Sul, v. 18, n. 65, jul/set. 2020. Disponível em: < https//doi.org/10.13037/ras.vol18n65.6647 >. Acesso em: 10 mai. 2021.

ROMÃO, Wesley et al. Extração de regras de associação em C&T: O algoritmo Apriori. **XIX Encontro Nacional em Engenharia de Produção**, v. 34, p. 37-39, 1999.

SACRAMENTO, R. S. et al. Associação de variáveis sociodemográficas e clínicas com os tempos para início do tratamento do câncer de próstata. **Associação Brasileira de Ciência e Saúde Coletiva**, V. 24, n. 9; 2019. Disponível em:< https://doi.org/10.1590/1413-81232018249.31142017 >. Acesso em: 05 Jan. 2021.

SHARMA, S.; BANSAL, M. Real-Time Sentiment Analysis Towards Machine Learning. **International Journal of Scientific & Technology Research**, V. 9; 2020. Disponível em: <https://www.researchgate.net/profile/Mamta-Rajshree/publication/340934268_Real-Time_Sentiment_Analysis_Towards_Machine_Learning/links/5ea5bc3d299bf1125611b0ad/Real-Time-Sentiment-Analysis-Towards-Machine-Learning.pdf>. Acesso em: 10 mai. 2021.

SHARMA, P.; MEENA, U.; SHARMA, G.K. Application of Data Mining Algorithms for Tourism Industry. In: **Dash S.S., Das S., Panigrahi B.K. (eds) Intelligent Computing and Applications**. Advances in Intelligent Systems and Computing, vol 1172. Springer, Singapore, 2021. Disponível em: <https://doi-org.ez15.periodicos.capes.gov.br/10.1007/978-981-15-5566-4_43>. Acesso em: 03 mar 2021.

SILVA, L. A. et al. **Introdução à Mineração de Dados: com aplicações em R**. Rio de Janeiro: Elsevier, 2016.

SILVA, J. S.; NASCIMENTO, L. P. **Fatores Culturais Associados a não Adesão aos Exames Preventivos de Câncer de Próstata em Parintins**. Manaus, 2017. Disponível em: < http://repositorioinstitucional.uea.edu.br/bitstream/riuea/759/1/Fatores%20culturais%20associados%20a%20n%C3%A3o%20ades%C3%A3o%20aos%20exames%20preventivos%20de%20C%C3%A2ncer%20de%20Pr%C3%B3stata%20em%20Parintins.pdf>. Acesso em: 30 dez. 2019.

SOCIEDADE BRASILEIRO DE UROLOGIA (SBU); SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA/MEDICINA LABORATORIAL (SBPC/ML). **Rastreio de Câncer de Próstata**. 2018. Disponível em: < https://portaldaurologia.org.br/medicos/noticias/nota-oficial-sbu-e-sbpc-ml-rastreio-de-cancer-de-prostata/>. Acesso em: 05 jan. 2021.

SOUZA, A. M. P.; ZAIA, J. E. O uso do Data Mining na Promoção de Saúde. **Atas de Saúde Ambiental**, V. 3, n. 1; 2015. Disponível em: <https://revistaseletronicas.fmu.br/index.php/ASA/article/view/685/851>. Acesso em: 27 fev. 2021.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. Rio de Janeiro: Elsevier, 2009.

SHMUELI, G. et. al. **Data Mining for Business Analytics: Concepts, Tecniques, and Applications in Python**. John Wiley & Sons: Hoboken-USA, 2020.

ZHAO, Y.; ZHANG, C.; CAO, L. **Post-Mining of Association Rules: Techniques for Efective Knowledge Extraction**. New York: Hershey, 2009.

APÊNDICE A – PROTOCOLO DE REVISÃO SISTEMÁTICA APLICADA AO ESTUDO

1 PLANEJAMENTO E EXECUÇÃO DA REVISÃO SISTEMÁTICA

Nesse primeiro momento, ocorre a definição do protocolo de revisão que será utilizado, com os objetivos e questões de pesquisa; em seguida, tem-se as estratégias de buscas adotadas; logo após, os critérios de seleção dos resultados encontrados e se faz a extração dos dados dos resultados.

O protocolo surgiu da união de pontos fortes encontrados nos protocolos utilizados em Biolchini et. al (2005), Levac et. Al (2010) e Moraes et. al (2011) e, seguindo esse protocolo, chega-se a um resultado, logo, obedecendo-se aos mesmos procedimentos poder-se-á encontrar resultados semelhantes.

1.1 Descrição do problema

Com a aplicação do processo de informatização nos setores governamentais, inclusive a saúde, a quantidade de informação armazenada em banco de dados vem crescendo de forma considerável. Isto dificulta uma possível análise manual por qualquer agente (PREISSLER, 2016).

As políticas de saúde que possam ajudar a população seriam mais bem direcionadas se tivessem os perfis dos pacientes assistidos por elas, porém ainda é um campo de pesquisa com lacunas (ARAUJO et al., 2015). Ademias, isto não difere na formação do perfil de quem possui o câncer de próstata.

A mineração de dados apresenta-se como uma tática que pode ser utilizada para analisar de forma automática os dados, e permitir a formação dos perfis dos pacientes para assistir a população de forma mais efetiva e precisa (ARAUJO, 2007).

1.2 Objetivos

Com os estudos preliminares e com o propósito de levantar conteúdo bibliográfico para a compreensão do tema e lacunas do respectivo trabalho os objetivos são:

- Analisar fontes bibliográficas com o propósito de caracterizar a aplicação de mineração de dados em bases de dados na formação de perfis de pacientes com câncer de próstata;
- Analisar publicações científicas para delimitar as lacunas presentes no uso de regras de associação na análise de bases de dados para a formação de perfis de pacientes com câncer de próstata;
- Levantar publicações no estado da arte sobre a aplicação de regras de associação, como técnica de mineração de dados, adotada para a formação de perfis de pacientes com câncer de próstata.

1.3 Questão de pesquisa

Alcançados os objetivos anteriormente definidos, retorna-se aos fundamentos para responder o questionamento se o uso de regras de associação, como técnica de mineração de dados, é mais eficaz do que análise estatística para a formação do perfil de saúde do homem brasileiro propenso ao câncer de próstata. Entretanto, divide-se esta questão em quesitos mais específicas, como:

1. Há como se apontar a existência do uso recorrente de fatores como idade, raça, hábitos alimentares, hábitos de exercícios físicos nas regras de associação para a formação do perfil de saúde do paciente com câncer de próstata?
2. Qual o estado da arte em técnicas de análise de dados para formação de perfis de homens propensos ao câncer de próstata?
3. A ocorrência do uso de mineração de dados para a formação perfis de saúde de pacientes propensos ao câncer de próstata?

Ambas as perguntas têm valor fundamental no desenvolvimento do entendimento sobre o uso de regras de associação para a formação do perfil de pacientes com câncer de próstata.

1.4 Estratégias de Busca

Nessa seção, serão abordados todos os meios utilizados para realização das buscas. A definição das fontes de pesquisa para o levantamento bibliográfico, em que teremos a escolha das fontes de publicações, as definições das palavras-chave, os termos técnicos principais na área, a definição da *string* de busca que será aplicada nas bases de artigos e as línguas que serão utilizadas nas buscas.

1.4.1 Fontes de Pesquisa

Foram selecionadas as bibliotecas digitais: *IEEE XPLORE*, *ACM Library*, *ScienceDirect*, *Scopus*, *PubMed* e *Scielo* para a realização das consultadas em suas bases devido à grande indexação de trabalhos na área de informática e de saúde tanto nacional quanto internacional.

1.4.2 Palavras-Chave

Através das questões de pesquisa definidas na subseção 1.3, pode-se retirar as seguintes principais palavras-chave: análise de dados (*data analysis*), mineração de dados (*data mining*), perfil de saúde (*health profile*) e neoplasias da próstata (*prostatic neoplasms*).

1.4.3 String de Busca

Utilizando-se das palavras-chave anteriores e com a finalidade de remover os dados supérfluos e incluir os que deem embasamento para responder as questões de buscas a *string* de busca, as quais ficaram de acordo com o

Quadro 1, foram construídas questões utilizando os sinônimos dos termos chaves para garantir que, nos resultados, encontrássemos trabalhos de acordo com o tema mesmo sendo escrito com termos semelhantes.

Quadro 1- *String* de busca em Inglês

String de busca
((“data mining” OR “data analysis”) AND (“prostatic neoplasms” OR “cancer of the prostate”) AND (“health profile” OR “Epidemiological Profile”)) OR ((“prostatic

neoplasms" OR "cancer of the prostate") AND ("health profile" OR "Epidemiological Profile"))

Fonte: Elaborado pelo autor, 2021.

Porém, para a consulta na base scielo, que é uma base que adota a língua portuguesa esta *string* de busca ficou escrita de acordo com o

Quadro **2**.

Quadro 2 - String de busca em português

String de busca
(("mineração de dados" *OR* "análise de dados") *AND* ("neoplasias da próstata" *OR* "câncer de próstata") *AND* ("perfil de saúde" *OR* "perfil epidemiológico")) *OR* (("neoplasias da próstata" *OR* "câncer de próstata") *AND* ("perfil de saúde" *OR* "perfil epidemiológico"))

Fonte: Elaborado pelo autor, 2021.

1.4.4 Língua

As línguas adotadas para escolha na busca são o inglês, por se tratar da mais adotada nas bases de periódicos, e a portuguesa devido à presença da base *Scielo*.

1.5 Critérios de seleção

Para realizar o processo de seleção das publicações que serão analisadas na revisão, serão considerados os critérios de inclusão e de exclusão dispostos a seguir.

Os critérios de inclusão são:

- O resultado deve estar no idioma inglês ou português;
- O resultado deve estar disponível integralmente na web;
- O resultado deve conter no título e/ou no resumo alguma relação com o tema deste trabalho.

Os critérios de exclusão são:

- O resultado ter publicação superior a cinco anos a contar do ano de 2020;
- É um resumo ou Revisão Sistemática de Literatura ou pesquisa bibliográfica;

- Artigos iguais ou que possuam versão atualizada (deve ser mantido apenas o mais recente).
- O resultado obter nível inferior ao nível *low* na avaliação de qualidade descrita no item 3.6.

1.6 Procedimento da seleção

Para realização do processo de seleção dos resultados, será utilizada a ferramenta *Start* (*State of the Art through Systematic Review*)[6], software desenvolvido por um laboratório (o LaPES) da Universidade Federal de São Carlos, do Estado de São Paulo no Brasil, para gerenciar todas as referências e citações encontradas na busca.

No início, foi realizada a busca em cada fonte de pesquisa, depois o seu resultado foi exportado para um arquivo BibTex[7], que é utilizado na ferramenta para o gerenciamento das citações. Além disso, depois de importado os arquivos das pesquisas, o *Start* lista todos os resultados, e primeiramente questiona se desejamos classificar os trabalhos duplicados automaticamente ou de forma manual.

Diante dos resultados das fontes, o processo de seleção ocorrerá em duas fases. A primeira será referente ao levantamento dos trabalhos que obedecem aos critérios de inclusão, caso o resultado não obedeça, pelo menos, um item esse será excluído. Na segunda fase, haverá a consulta aos títulos e resumos para constatar se algum trabalho ainda atende aos critérios de exclusão, caso não satisfaça pelo menos um, o trabalho será rejeitado.

Logo após a realização dos passos de seleção, restarão apenas os trabalhos que serão utilizados para extração dos dados relevantes para o trabalho em questão, que terão os seus passos descritos nas próximas seções.

1.7 Análise de Qualidade

Com a finalidade de considerar trabalhos com um teor de qualidade considerável na extração dos dados, o último critério de exclusão menciona uma avaliação de qualidade, que conterá fatores que darão esse teor de qualidade.

---

[6] Start encontra-se disponível no sítio lapes.dc.ufscar.br

[7] BibTex: é um arquivo que normalmente é utilizado em conjunto com o sistema tipográfico LaTex, porém se tornou bastante utilizado por outras ferramentas no controle da bibliografia.

A avaliação de qualidade tem o objetivo de enquadrar o trabalho em alguns níveis de prioridade para a leitura do trabalho, que variarão de acordo com a sua elaboração. Os níveis de avaliação serão quatro, oferecidos pela ferramenta *Start*, sendo eles: *very low*, *low*, *high* e *very high*, para cada um, consideramos a forma com que trata o tema proposto nesse trabalho. Para esta etapa, os níveis da avaliação se encontram de acordo com o

Quadro **3**.

Quadro 3 – Avaliação de qualidade

Níveis	**Critérios**
Very low	O trabalho faz menção a apenas mineração de dados ou câncer de próstata.
Low	O trabalho faz menção à mineração de dados e câncer de próstata.
High	O trabalho faz menção à mineração de dados e/ou perfil de paciente com câncer de próstata.
Very High	O trabalho faz menção à técnica de regras de associação e perfil de paciente com câncer de próstata.

Fonte: Elaborado pelo autor, 2021.

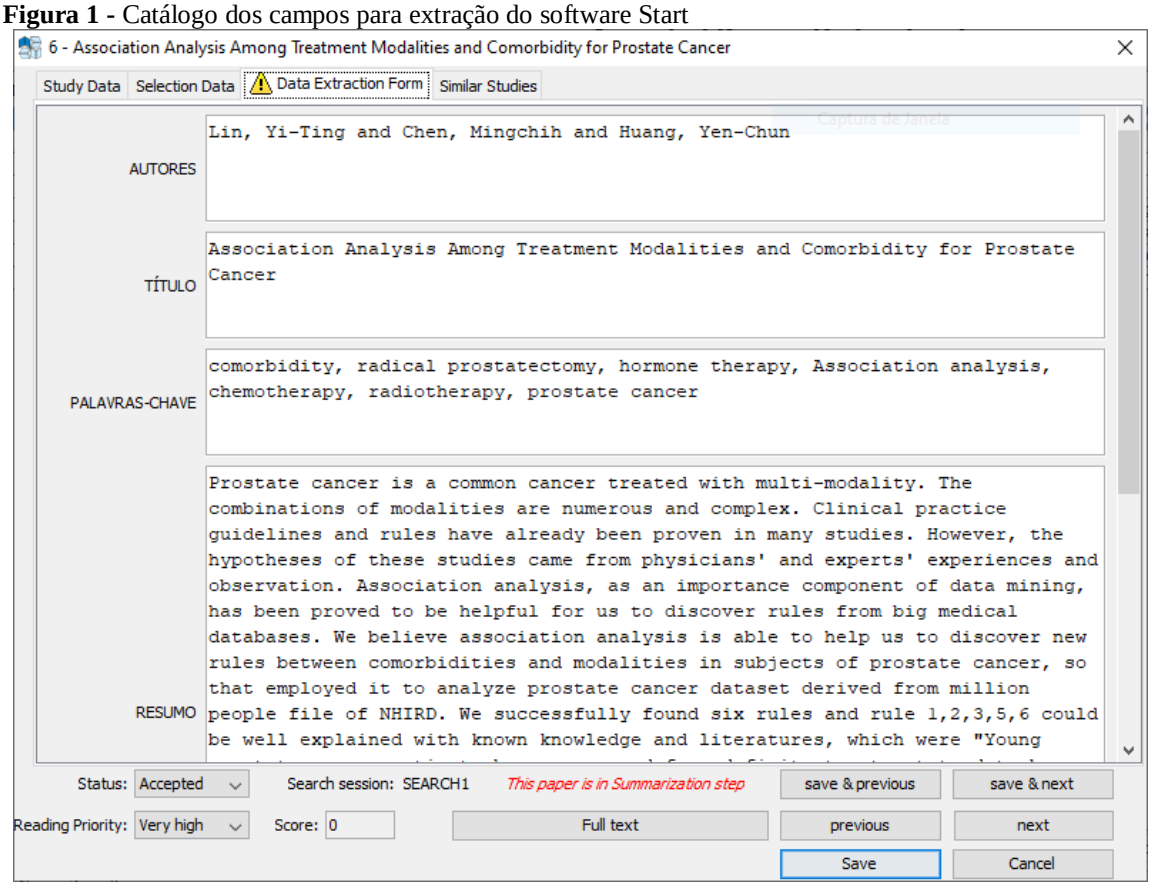
O processo de qualificação ocorrerá com a leitura das partes de introdução e conclusão dos trabalhos, para assim ter o nível de tratamento do estudo com relação à adoção de regras de associação como técnica de mineração de dados para a formação do perfil do paciente com câncer de próstata. Ao fim dessa etapa, os trabalhos que restarem serão repassados para a etapa de extração de dados.

1.8 Extração dos Resultados

Para a realização da extração dos dados, realizou-se a sua catalogação dos trabalhos na ferramenta, como pode ser visto na ficha presente na Figura 1, em que, dentre todos os dados considerados para fins de extração, temos: autores, título, palavras-chave, ano de publicação, base de indexação, fatores de influência do câncer de próstata, técnicas de formação de perfis de pacientes com câncer de próstata, técnicas de análise de dados, ferramentas de armazenamento de dados utilizadas, objetivo do estudo, e resultados estatísticos (se houver). Estes itens serão catalogados na guia *Data Extraction Form*, como pode ser visto na Figura 1.

.

Figura 1 - Catálogo dos campos para extração do software Start



Fonte: Elaborada pelo autor, 2021.

2 CONDUÇÃO DA REVISÃO SISTEMÁTICA

No término do planejamento da revisão, as atividades previstas começaram a ser desenvolvidas. Com a definição das fontes de pesquisa e da *string* de busca, foram realizadas as consultas. Em seguida, ocorreram as realizações dos filtros de seleção passando primeiramente pelos de inclusão e depois de exclusão. Ao final dessa etapa, foram retiradas as informações relevantes dos trabalhos selecionados. Todo esse processo será detalhado nas seções que seguem.

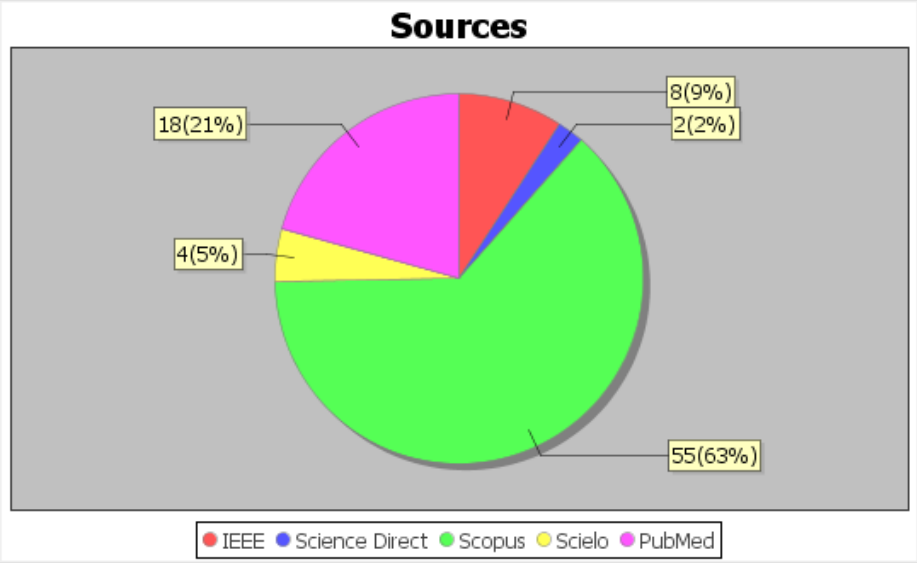
2.1 Realização da Seleção

Nesta seção, serão abordadas as buscas realizadas nos portais selecionados. Para ter um resultado de relevância, foi utilizada, em cada portal, uma forma avançada da busca.

Porém, na base *ACM Library*, não se retornou nenhum resultado, na *Sciense Direct*, foi necessária a divisão da *string* de busca divido a uma limitação da ferramenta de busca do portal. Na primeira parte, ficou o "*prostatic neoplasms" OR "cancer of the prostate" AND "health profile" OR "Epidemiological Profile*"; e, na segunda, o "*data mining" OR "data analysis" AND "prostatic neoplasms" OR "cancer of the prostate" AND "health profile" OR "Epidemiological Profile*". Contudo, nas demais, não houve a necessidade de alteração na *string* de busca.

Além disso, a base de artigos *Scielo*, com o uso da *string* em Inglês e português, apresentou o mesmo resultado. No contexto geral, foram encontrados 87 artigos, cujos artigos se encontram distribuídos nas bases de acordo com a Figura 2, que, em seguida, foram submetidos aos critérios de inclusão e exclusão.

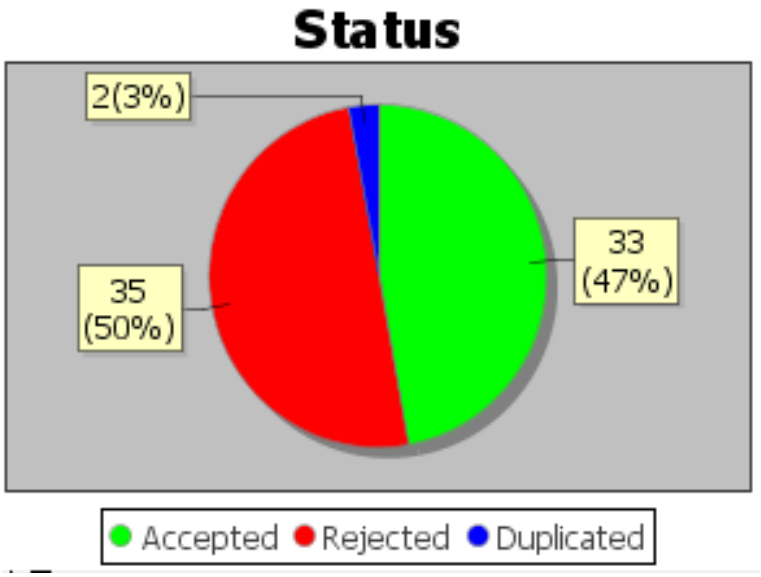
Figura 2 - Distribuição dos trabalhos retornados na busca por cada portal



Fonte: Elaborada pelo autor, 2021.

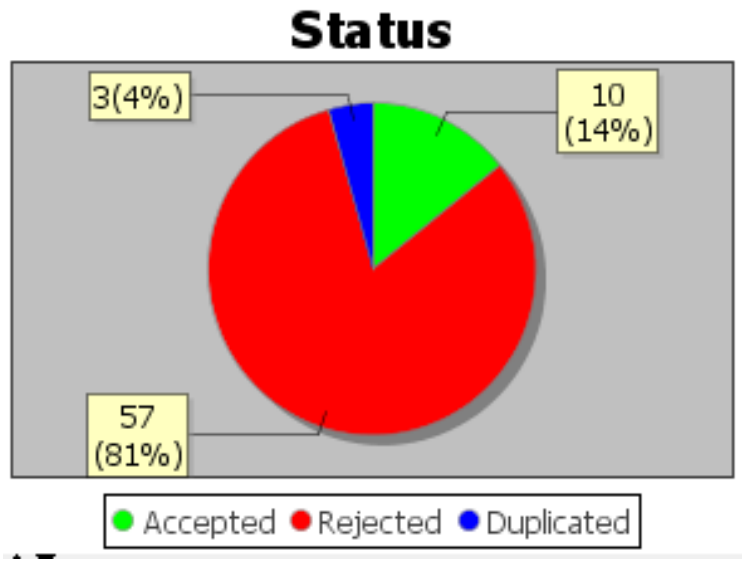
As etapas de inclusão e exclusão foram, logo em seguida, aplicadas à amostra de trabalhos encontrados. Na seleção, todos os trabalhos passaram pelos critérios de inclusão e, logo após, foram aplicados os critérios de exclusão, os quais são representados na Figura 3 e na Figura 4, e expõem os resultados dessas etapas respectivamente.

Figura 3 - Resultado após aplicação dos critérios de inclusão.



Fonte: Elaborada pelo autor, 2021.

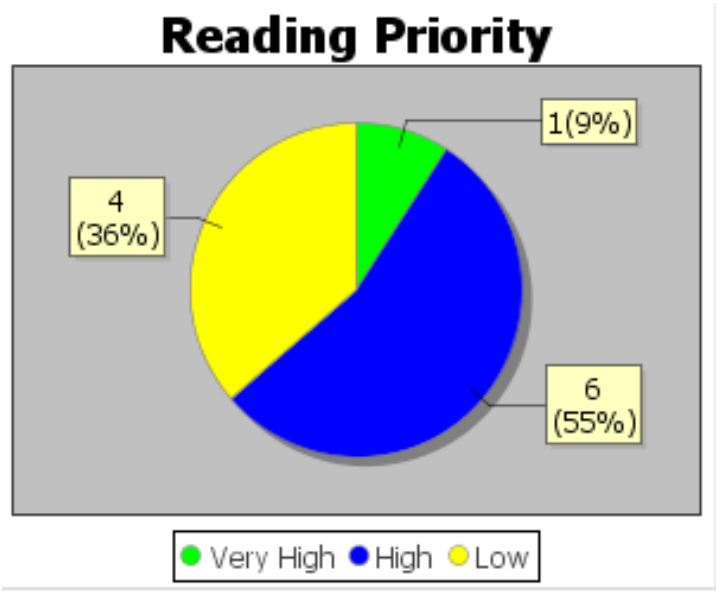
Figura 4 - Resultado após aplicação dos critérios de exclusão.



Fonte: Elaborada pelo autor, 2021.

Entre os critérios de exclusão se encontra a avaliação do nível de correlação do artigo encontrado com o tema da revisão, que fica definido no item de definição de prioridade de leitura. Para isso, foram lidos e classificados os trabalhos que passaram pelo processo de inclusão, de modo que o resultado gerou o seguinte gráfico exposto na Figura 5.

Figura 5 - Resultado da avaliação da Prioridade de Leitura.



Fonte: Elaborada pelo autor, 2021.

Ao término da fase de seleção dos trabalhos encontrados nas buscas e com a garantia que dentre todos restaram apenas os trabalhos que mais se aproximam do que se deseja pesquisar, e possa ajudar a responder as questões de pesquisa, será desenvolvida a fase da extração dos dados. Esta fase será melhor descrita na seção seguinte.

2.2 Execução da extração dos dados

Toda a catalogação dos trabalhos selecionados pode ser feita través das informações que se encontram nas guias de detalhes de cada artigo presentes no software *Start*, porém deve-se frisar que a presença destes dados depende das ferramentas de buscas dos portais consultados. Outrossim, os dados podem ser obtidos através do relatório gerado pela própria ferramenta.

Porém, o processo de extração dos dados dos trabalhos selecionados se deu através da extração objetiva, a partir da qual foram colhidas as informações de Título do trabalho, Autor (es), Palavras-chave, Resumo, Fonte de pesquisa e Ano de publicação. Para visualização dos trabalhos selecionados foi elaborada a Tabela 1.

Tabela 1 - Artigos selecionados

Título	Autores	Ano	Revista	Volume	Páginas	DOI
Determinants of changes in physical activity from pre-diagnosis to post-diagnosis in a cohort of prostate cancer survivors	Stone, C.R. and Courneya, K.S. and McGregor, S.E. and Li, H. and Friedenreich, C.M.	2019	Supportive Care in Cancer	27	2819-2828	10.1007/s00520-018-4578-2
Applying Bayesian Network Approach to Determine the Association Between Morphological Features Extracted from Prostate Cancer Images	L. {Hussain} and A. {Ali} and S. {Rathore} and S. {Saeed} and A. {Idris} and M. U. {Usman} and M. A. {Iftikhar} and D. Y. {Suh}	2019	IEEE Access	7	1586-1601	10.1109/ACCESS.2018.2886644
Towards the Identification of Histology Based Subtypes in Prostate Cancer	A. {Chatrian} and K. {Sirinukunwattana} and C. {Verrill} and J. {Rittscher}	2019	2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)		948-952	10.1109/ISBI.2019.8759199
A Review of Localized Prostate	Cassell A ; Yunusa B ;	2019	World Journal of			

Cancer: An African Perspective.PG - 162-8	Jalloh M ; Mbodji MM ; Diallo A ; Ndoye M ; Kouka SC ; Labou I ; Niang L ; Gueye SM		Oncology	
Vivências De Homens Com Câncer De Próstata	Ferrão, Luana ; Antonio Bettinelli, Luiz ; Rodrigues Portella, Marilene	2017	J Nurs UFPE on line	10.5205/reuol.10712-95194-3-SM.1110sup201720
Knowledge, Behaviour and Health Practices of Men Concerning the Prostate Cancer.	Menezes, R ; Menezes, M ;¦, E Ferraz Teston - Revista de Pesquisa ; 2019, undefined	2019	core.ac.uk	
Perfil clínico epidemiológico da população atendida num programa de rastreamento de câncer de próstata Epidemiological clinical profile of the population served in a prostate cancer screening program	Nunes, Renata Furletti ; Rego, Barros ; Barros, Rodrigo Alencar	2020	seer.uscs.edu.br	
Association of	Raone Silva	2019	SciELO Brasil	

sociodemographic and clinical variables with initial staging in men with prostate cancer	SacramentoLuana de Jesus SimiãoKátia Cirlene Gomes VianaMaria Angélica Carvalho AndradeMaria Helena Costa AmorimEliana Zandonade	

Fonte: Elaborada pelo autor, 2021.

ANEXO A – DICIONÁRIO DA BASE DE DADOS DO INCA

Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC (atualizado 30/03/2020)

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
44	ALCOOLIS	Histórico de consumo de bebida alcoólica	1. Nunca; 2.Ex-consumidor; 3.Sim; 4.Não avaliado; 8.Não se aplica; 9.Sem informação
22	ANOPRIDI	Ano do diagnóstico	aaaa
42	ANTRI	Ano da triagem	dd/mm/aaaa
base de dados SP	BASDIAGSP	Base mais importante para o diagnóstico do tumor	1.Exame clínico 2.Recursos auxiliares não microscópicos 3.Confirmação microscópica 4.Sem informação
24	BASMAIMP	Base mais importante para o diagnóstico do tumor	1.Clínica; 2.Pesquisa clínica; 3.Exame por imagem; 4.Marcadores tumorais; 5.Citologia; 6.Histologia da metástase; 7.Histologia do tumor primário; 9. Sem informação
47	CLIATEN	Clínicas do primeiro atendimento - entrada do paciente	Codificação segundo Tabela de Clínicas do SisRHC
31	CLITRAT	Clínica de início do tratamento	Codificação segundo Tabela de Clínicas do SisRHC
SIS	CNES	Número do CNES do Hospital	Codificação segundo tabela do Cadastro Nacional de Estab. de Saúde
32	DATAINITRT	Data do início do primeiro tratamento específico para o tumor, no hospital	dd/mm/aaaa
36	DATAOBITO	Data do óbito	dd/mm/aaaa
21	DATAPRICON	Data da 1ª consulta	dd/mm/aaaa
23	DIAGANT	Diagnóstico e tratamento anteriores	1.Sem diag./Sem trat.; 2.Com diag./Sem

			trat.; 3.Com diag./Com trat.; 4.Outros; 9. Sem informação
22	DTDIAGNO	Data do primeiro diagnóstico	dd/mm/aaaa
32	DTINITRT	Ano do início do primeiro tratamento específico para o tumor, no hospital	aaaa
21	DTPRICON	Ano da 1ª consulta	aaaa
42	DTTRIAGE	Data da triagem	dd/mm/aaaa
28a	ESTADIAG	Estadiamento clínico do tumor (TNM) - Grupo	Codificação do grupamento do estádio clínico segundo classificação TNM
28a	ESTADIAM	Estadiamento clínico do tumor (TNM)	Codificação do grupamento do estádio clínico segundo classificação TNM
17	ESTADRES	UF de procedência (residência)	Sigla da UF de procêdencia
41	ESTCONJ	Estado conjugal atual	1.Solteiro; 2.Casado; 3.Viúvo; 4.Separado judicialmente; 5.União consensual; 9.Sem informação
35	ESTDFIMT	Estado da doença ao final do primeiro tratamento no hospital	1.Sem evidência da doença (remissão completa); 2.Remissão parcial; 3.Doença estável; 4.Doença em progressão; 5.Suporte terapêutico oncológico; 6. Óbito; 8. Não se aplica; 9. Sem informação
48	EXDIAG	Exames relevantes para o diagnóstico e planejamento da terapêutica do tumor	1.Exame clínico e patologia clínica; 2.Exames por imagem; 3.Endoscopia e cirurgia exploradora; 4.Anatomia patológica; 5.Marcadores tumorais; 8.Não se aplica; 9. Sem informação
43	HISTFAMC	Histórico familiar de câncer	1.Sim; 2.Não; 9.Sem informação
SIS	IDADE	Idade na 1ª consulta (calculada pela diferença entre a data da 1ª consulta e a data do nascimento)	Idade, em anos; valor igual a zero para crianças menores de 1 ano
11	INSTRUC	Escolaridade	1.Nenhuma; 2.Fundamental incompleto; 3.Fundamental completo; 4.Nível médio;

			5.Nível superior incompleto; 6.Nível superior completo; 9.Sem informação
50	LATERALI	Lateralidade do tumor	1.Direita; 2. Esquerda; 3.Bilateral; 8.Não se aplica; 9.Sem informação
9	LOCALNAS	Local de nascimento	Sigla da UF de nascimento
25	LOCTUDET	Localização primária (Categoria 3d)	Código da CID-O, 3 dígitos
25	LOCTUPRI	Localização primária detalhada (Subcategoria 4d)	Código da CID-O, 4 dígitos
49	LOCTUPRO	Localização provável do tumor primário (somente para os casos em que a localização primária do tumor é desconhecida)	CID-O, 4 dígitos
51	MAISUMTU	Ocorrência de mais um tumor primário	1.Não; 2.Sim; 3.Duvidoso
	MUUH	Município da unidade hospitalar	Tabela de municípios do IBGE
12	OCUPACAO	Ocupação principal	Codificação pela Tabela da Código Brasileiro de Ocupações; mais de três 9 representa Ocupação ignorada
46	ORIENC	Origem do encaminhamento	1.SUS; 2.Não SUS; 3.Veio por conta própria;8.Não se aplica; 9. Sem informação
28b	OUTROESTA	Outros estadiamentos clínicos do tumor	Codificação do grupamento do estádio clínico segundo outras classificações que não a TNM
34	PRITRATH	Primeiro tratamento recebido no hospital	1.Nenhum; 2. Cirurgia; 3.Radioterapia; 4.Quimioterapia; 5.Hormonioterapia; 6.Transplante de medula óssea; 7.Imunoterapia; 8.Outras; 9.Sem informação
13	PROCEDEN	Código do Município de procedência (residência)	Tabela de municípios do IBGE
10	RACACOR	Raça/cor	1.Branca; 2.Preta; 3.Amarela; 4.Parda; 5.Indígena; 9.Sem informação

33	RZNTR	Principal razão para a não realização do tratamento antineoplásico no hospital	1.Recusa do tratamento; 2.Tratamento realizado fora; 3.Doença avançada, falta de condições clínicas ou outras doenças associadas; 4.Abandono do tratamento; 5.Complicações de tratamento; 6.Óbito; 7.Outras razões; 8.Não se aplica; 9. Sem informação
6	SEXO	Sexo	1. Masculino; 2. Feminino
45	TABAGISM	Histórico de consumo de tabaco	1.Nunca; 2.Ex-consumidor; 3.Sim; 4.Não avaliado; 8.Não se aplica; 9.Sem informação
26	TIPOHIST	Tipo histológico do tumor primário	Codificação da morfologia do tumor pela CID-O
27	TNM	TNM	Codificação do estádio clínico segundo classificação TNM
38	TPCASO	Tipo de caso	1. Sim (Analítico); 2. Não (Não analítico)
SIS	UFUH	UF da unidade hospitalar	Sigla da Unidade da Federação da unidade hospitalar (IBGE)
SIS	VALOR_TOT	Text	-