J. Health Inform. 2024 Janeiro-Dezembro; 16

Artigo Original

DOI: https://doi.org/10.59681/2175-4411.v16.2024.1014

Mineração de dados aplicada sobre câncer relacionado ao trabalho

Data mining applied to cancer work-related

Minería de datos aplicada sobre el cáncer relacionado con el trabajo

Bruna Ferreira Pfeiffer[1], Silvia Regina Gralha[1], Giordani da Silva Ramos[1]

Descritores: Mineração de Dados; Câncer Ocupacional; Produto Químico.

RESUMO

Objetivo: Encontrar regras de associação entre a ocupação do trabalhador, o produto químico exposto e o câncer diagnosticado em 2019. Método: Foram aplicadas técnicas de Mineração de Dados, dentro do processo de Descoberta de Conhecimento em Bases de Dados. Para identificar padrões e correlações, foram utilizados arquivos sobre Câncer Relacionado ao Trabalho – disponíveis pelo Sistema de Informação de Agravos de Notificação –, o *software Weka* e o algoritmo *Apriori*. Resultados: Apresentamos 2 regras com a métrica "Confiança" e 4 regras com a métrica "Convicção", que indicaram fortes associações entre "Produtor agrícola polivalente", "Radiação solar", "Outras neoplasias malignas da pele e doenças relacionadas" e "Radiação não ionizante e Agrotóxico". Conclusão: Os resultados podem incentivar organizações a elaborarem estratégias de prevenção contra o câncer ocupacional, de forma a manter e garantir a qualidade de vida e segurança dos trabalhadores, sobretudo dos trabalhadores pertencentes às ocupações com maior risco de exposição ao câncer.

Keywords: Data Mining; Occupational Cancer; Chemical Products.

ABSTRACT

Objective: To find association rules between the worker's occupation, the chemical exposed and the cancer diagnosed in 2019. Method: Data Mining techniques were applied, within the Knowledge Discovery in Databases process. To identify patterns and correlations, files on Work-Related Cancer – available from the Notification Aggravities Information System –, the Weka software and the Apriori algorithm were used. Results: We present 2 rules with the "Confidence" metric and 4 rules with the "Conviction" metric, which indicated strong associations between "Multipurpose agricultural producer", "Solar radiation", "Other malignant skin neoplasms and related diseases" and "Non-ionizing radiation and Pesticide". Conclusion: The results may encourage organizations to develop prevention strategies against occupational cancer, in order to maintain and ensure the quality of life and safety of workers, especially workers belonging to occupations with higher risk of cancer exposure.

Descriptores: Minería de Datos; Cáncer Profesional; Productos Químicos.

RESUMEN

Objetivo: Encontrar reglas de asociación entre la ocupación del trabajador, el producto químico expuesto y el cáncer diagnosticado en 2019. Método: Se aplicaron técnicas de Minería de Datos, dentro del proceso de Descubrimiento de Conocimiento en Bases de Datos. Para identificar patrones y correlaciones, se utilizaron archivos sobre cáncer relacionado con el trabajo – disponibles por el Sistema de Información de Agravios de Notificación –, el programa informático Weka y el algoritmo Apriori. Resultados: Presentamos 2 reglas con la métrica "Confianza" y 4 reglas con la métrica "Convicción", que indicaron fuertes asociaciones entre "Productor agrícola polivalente", "Radiación solar", "Otras neoplasias malignas de piel y enfermedades relacionadas" y "Radiación no ionizante y pesticida". Conclusión: Los resultados pueden animar a las organizaciones a desarrollar estrategias de prevención contra el cáncer ocupacional, con el fin de mantener y garantizar la calidad de vida y la seguridad de los trabajadores, especialmente de los trabajadores pertenecientes a ocupaciones con mayor riesgo de exposición al cáncer.

[1] *Mestre(a) em Tecnologias da Informação e Gestão em Saúde do PPG-Tecnologias da Informação e Gestão em Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre – UFCSPA, Porto Alegre (RS), Brasil.*

Autor Correspondente: **Bruna Ferreira Pfeiffer**
e-mail: **bruna.pfeiffer@ufcspa.edu.br**

Artigo recebido: 17/01/2023
Aprovado: 29/01/2024
https://jhi.sbis.org.br/

INTRODUÇÃO

Segundo a Organização Mundial de Saúde (OMS), as intoxicações acidentais, ocupacionais ou intencionais são importantes causas de agravos à saúde. Estima-se que 1.5 a 3.0% da população mundial intoxicam-se todos os anos. Para o Brasil, isto representa aproximadamente 4.800.000 casos novos a cada ano, destes, 0.1 a 0.4% das intoxicações resultam em óbito[1].

O Sistema de Informação de Agravos de Notificação (SINAN) foi desenvolvido no início da década de 90 com o objetivo inicial a coleta e processamento dos dados sobre agravos de notificação em todo o território nacional[2]. Dentre o grande grupo de agentes químicos a que a população se expõe diariamente, os agrotóxicos, gases tóxicos e metais pesados são elementos em destaque. Conforme definido pelo Ministério da Saúde (MS), os sistemas de vigilância nas secretarias de saúde são monitorados e controlados, nos três níveis federativos, para casos agudos de intoxicação por agrotóxicos[3].

Uma das bases de dados nacional, referente à saúde e disponível a população, provém do Departamento de Informática do Sistema Único de Saúde (DATASUS), do Governo Federal. Tal departamento é um grande provedor de soluções de *software* para as secretarias estaduais e municipais de saúde, sempre adaptando seus sistemas às necessidades dos gestores e incorporando novas tecnologias, na medida em que a descentralização da gestão se torna mais concreta. Este volume de dados em saúde é uma rica fonte da qual pode se extrair informações, desde que se utilize as técnicas adequadas[4].

O processo de Descoberta de Conhecimento em Bases de Dados (DCBD) é uma técnica que busca encontrar conhecimento a partir de um conjunto de dados, para que ele possa ser utilizado em um processo decisório, desde que o conhecimento descoberto seja útil e de interesse para os usuários e que seja apresentado de forma compreensível. Uma dessas etapas iniciais de busca de conhecimento é a Mineração de Dados (MD)[5-6]. A MD incorpora e investiga um grande volume de dados para adquirir exemplos significativos e padrões, sendo implementado em diversos segmentos: publicidade em banco de dados, gerenciamento de risco de crédito, identificação de declarações falsas, separar *e-mails* de *spam*, observar as perspectivas ou avaliações do cliente[7], entre outros. Na área da saúde as aplicações são cada vez mais diversas, desde a avaliação dos padrões de prescrição de medicamentos até a busca de correlações para traçar o risco de hospitalização de um determinado grupo de pacientes[8-9].

Este estudo teve por objetivo avaliar as associações, através da MD, existentes entre as ocupações dos trabalhadores, os tipos de produtos químicos expostos e os tipos de câncer diagnosticados e notificados em 2019. Este ano foi escolhido pelos autores por representar um período de normalidade de jornada de trabalho antes da pandemia da Covid-19. Foram identificados quais fatores impactam nos diagnósticos, ou quais resultados podem responder a tais questões, como: 1) Qual(is) produto(s) químico(s) o trabalhador esteve exposto?; 2) Qual a ocupação do trabalhador que se expôs ao(s) produto(s) químico(s)? e; 3) Apresentar os diferentes modelos de associação para identificar e entender as visões que este conjunto de dados pode mostrar.

MATERIAIS E MÉTODOS

Trata-se de uma pesquisa básica e experimental de MD baseada no modelo de DCBD, proposto por Parsaye[10] e difundido por Fayyad[5]. O modelo de DCBD apresenta as etapas de: 1) Seleção; 2) Pré-processamento; 3) Transformação; 4) Mineração de dados e; 5) Avaliação. Ao final das etapas espera-se obter o conhecimento desejado, sendo possível retornar em alguma etapa específica para ajuste (modelo com etapas lineares e cíclicas).

Empregar técnicas como, breve preparação do conjunto de dados, podem facilitar o trabalho do especialista de forma a minimizar o tempo despendido em tarefas de classificação e agrupamento[10]. Existe uma grande quantidade de ferramentas de MD, tanto comerciais quanto open source. Dentre elas, *Waikato Environment for Knowledge Analysis (Weka)* versão 3.8.6., selecionado pelos autores por atender aos objetivos do estudo. O *Weka* foi desenvolvido pela Universidade de Waikato, Nova Zelândia, em linguagem JAVA, na forma de algoritmos de aprendizado supervisionado e não supervisionado de máquina para a realização de tarefas de MD(11).

A base de dados intitulada "Câncer Relacionado ao Trabalho" refere-se a informações que documenta casos de câncer nos quais a exposição a fatores, agentes e situações de risco no ambiente e processo de trabalho é identificada como um elemento causal. Essa base de dados é construída para rastrear e analisar casos em que a incidência de câncer está diretamente ligada às condições laborais e à exposição a determinados elementos presentes no local de trabalho. A definição destaca que a exposição aos fatores de risco no ambiente de trabalho pode persistir mesmo após a cessação da exposição, isso significa que a base de dados inclui casos nos quais a influência do ambiente de trabalho na ocorrência do câncer pode continuar a se manifestar mesmo depois que o trabalhador tenha deixado de estar exposto a esses fatores. Tal base é disponibilizada para a população a partir da plataforma DATASUS[12].

RESULTADOS

O presente estudo seguiu as 5 etapas de DCBD proposto por Fayyad e estão detalhadas nos subtópicos a seguir:

Seleção da Base de Dados

Para atingir o objetivo deste estudo, foi realizado o *download* de 3 arquivos: *software Tabwin* – para tabular e associar os arquivos de tabulação e de dados –, arquivo de tabulação do SINAN e o arquivo de dados sobre câncer relacionado ao trabalho, ambos disponíveis em <https://datasus.saude.gov.br/transferencia-de- arquivos/>. Após tabulação e associação dos arquivos, o arquivo "CANCBR19.csv" foi gerado para análise dos dados, contendo 57 atributos – classificação qualitativa sobre os dados, disposto em colunas – e 788 instâncias – cada instância (linha) corresponde a dados de uma notificação sobre câncer relacionado ao trabalho, sendo qualitativo ou quantitativo, preenchido conforme o atributo. O Quadro 1 detalha os atributos pertencentes ao arquivo obtido. Todo processo ocorreu via *software Tabwin*.

Quadro 1 - Atributos pertencentes ao arquivo de dados "Câncer Relacionado ao Trabalho" notificados no ano de 2019

Atributos					
ALCATRAO	AMINA	ASBESTO	BENZENO	BERILIO	CADMIO
CAT	CNAE	CROMO	CS_ESCOL_N	CS_GESTANT	CS_RACA
CS_SEXO	CS_RACA	DIAG_ESP	DT_DIGITA	DT_NOTIFIC	DT_OBITO
EVOLUCAO	FUMA	HIDROCARBO	HORMONIO	ID_AGRAVO	ID_MN_RESI
ID_MUNICIP	ID_OCUPA_N	ID_PAIS	ID_REGIONA	ID_RG_RESI	ID_UNIDADE
IONIZANTES	MUN_EMP	NAO_IONIZA	NEOPLASICO	NIQUEL	NU_ANO
NU_IDADE_N	NUTEMPO	NUTEMPORIS	OLEOS	OUT_EXP_DE	OUTRO_EXP
SEM_DIAG	SEM_NOT	SG_UF	SG_UF_NOT	SILICA	SIT_TRAB
TEMPO_FUMA	TERCEIRIZA	TP_NOT	TP_TEMP_FU	TPTEMPO	TPTEMPORIS
TRAB_DOE	REGIME	UF_EMP			

O arquivo "CANCBR19.csv" foi aberto no *software Excel* e convertido para extensão ".xlsx" com intuito de executar a etapa de pré-processamento (subtópico 3.2.) e a etapa de transformação (subtópico 3.3.). Posteriormente o arquivo foi aberto no *Weka* para aplicar a MD e por fim salvo na extensão ".arff". Além dos 3 arquivos, foram realizados *downloads* dos arquivos "Dicionário de Dados" e "Ficha de Notificação/Investigação" para compreender as informações resumidas e abreviadas, disponíveis em <http://portalsinan.saude.gov.br/drt--cancer-relacionado-ao-trabalho>.

Pré-processamento

A fase de limpeza garante a confiabilidade dos dados que serão utilizados na fase da mineração, pois caso não sejam corrigidos, contornados ou minimizados, eventuais erros nos dados podem comprometer a eficácia da MD[13]. Durante a fase de limpeza, houve a exclusão manual de 95 instâncias com *missing data*.

A seleção de atributos é uma técnica utilizada com o intuito de reduzir a quantidade dos dados, facilitando a aplicação de algoritmos de mineração. Esta redução visa eliminar atributos que não agregam informações para a análise, produzindo assim uma representação mais compacta, mais facilmente interpretável do objetivo a ser alcançado, focalizando a atenção do usuário sobre os atributos mais relevantes[13]. Ao aplicar a seleção de atributos houve a exclusão manual de 39, restando 18 atributos de interesse que contribuíram para responder as questões centrais e atingir o objetivo principal do estudo: ID_OCUP_N, DIAG_ESP, ASBESTO, SILICA, AMINA, BENZENO, ALCATRAO, HIDROCARBO, OLEOS, BERILIO, CADMIO, CROMO, NIQUEL, IONIZANTES, NAO_IONIZA, HORMONIO, NEOPLASICO e OUT_EXP_DE. Ao final da etapa de pré-processamento, o arquivo de dados contava com 18 atributos e 693 instâncias.

Transformação

Ao final dos 18 atributos de interesse para análise, 17 apresentavam dados numéricos e 1 apresentava um campo livre para digitação no qual o responsável pela notificação pudesse informar se o trabalhador se expôs a outro(s) tipo(s) de produto(s) químico(s) não mencionado(s) na "Ficha de Notificação/Investigação". Portanto, os atributos estavam distribuídos como:

- ID_OCUP_N: atributo numérico com valores preenchidos de acordo com a Classificação Brasileira de Ocupações (CBO);
- DIAG_ESP: atributo numérico com valores preenchidos de acordo com a Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID);
- ASBESTO, SILICA, AMINA, BENZENO, ALCATRAO, HIDROCARBO, OLEOS, BERILIO, CADMIO, CROMO, NIQUEL, IONIZANTES, NAO_IONIZA, HORMONIO e NEOPLASICO: atributos numerais, sendo: 1 (sim), 2 (não) e 9 (ignorado);
- OUT_EXP_DE: campo livre para digitação.

Para melhor disposição e interpretação dos dados gerados pela ferramenta *Weka*, informações redundantes no atributo OUT_EXP_DE foram unificados, por exemplo: agrotóxico, organoclorado, organofosforado, herbicida, pesticida, tordon, produto clorado, veneno para lavoura, gramofone e roundup, solupan e intercap foram unificados como "Agrotóxico", assim como radiação solar e exposição solar foram unificados como "Radiação solar". Após processo de unificação, os atributos ASBESTO, SILICA, AMINA, BENZENO, ALCATRAO, HIDROCARBO, OLEOS, BERILIO, CADMIO, CROMO, NIQUEL, IONIZANTES, NAO_IONIZA, HORMONIO, NEOPLASICO e OUT_EXP_DE foram agrupados para o atributo criado pelos autores denominado "QUIMICO". Por fim, todos os atributos foram transformados manualmente para dados nominais, detalhados no Quadro 2.

Nos casos em que o trabalhador esteve exposto a mais de um elemento químico, por exemplo: NAO_IONIZA (Radiação não ionizante) e OUT_EXP_DE (Agrotóxico), esses dois elementos foram consolidados como uma análise. Tal procedimento foi decorrente da impossibilidade de discernir se o desenvolvimento do câncer resultou da exposição de um elemento químico ou se advém da associação de dois ou mais elementos químicos. Ao final da etapa de Transformação, o arquivo de dados final era composto por 3 atributos e 693 instâncias. Possui 160 registros com valores distintos e 98 ocupações únicas, 133 produtos químicos e 61 diagnósticos. Posteriormente, o arquivo foi exportado para o Weka e então gerado o arquivo "CANCBR19.arff" para o processo de MD propriamente dito.

Quadro 2 - Atributos reservados para MD e seus respectivos dados originais e transformados

Atributos	Dados originais	Dados transformados
ID_OCUPA_N	Exemplos: 314705, 715210 e 622010.	Exemplos: Técnico de acabamento em siderurgia, Pedreiro e Jardineiro.
DIAG_ESP	Exemplos: C91, C00 e C56.	Exemplos: Leucemia linfoide, Neoplasia maligna do lábio e Neoplasia maligna do ovário.
QUIMICO	Atributo nominal criado pelos autores.	Exemplos: Chumbo, Cimento, Cloro, Cola, Verniz e Tinta.

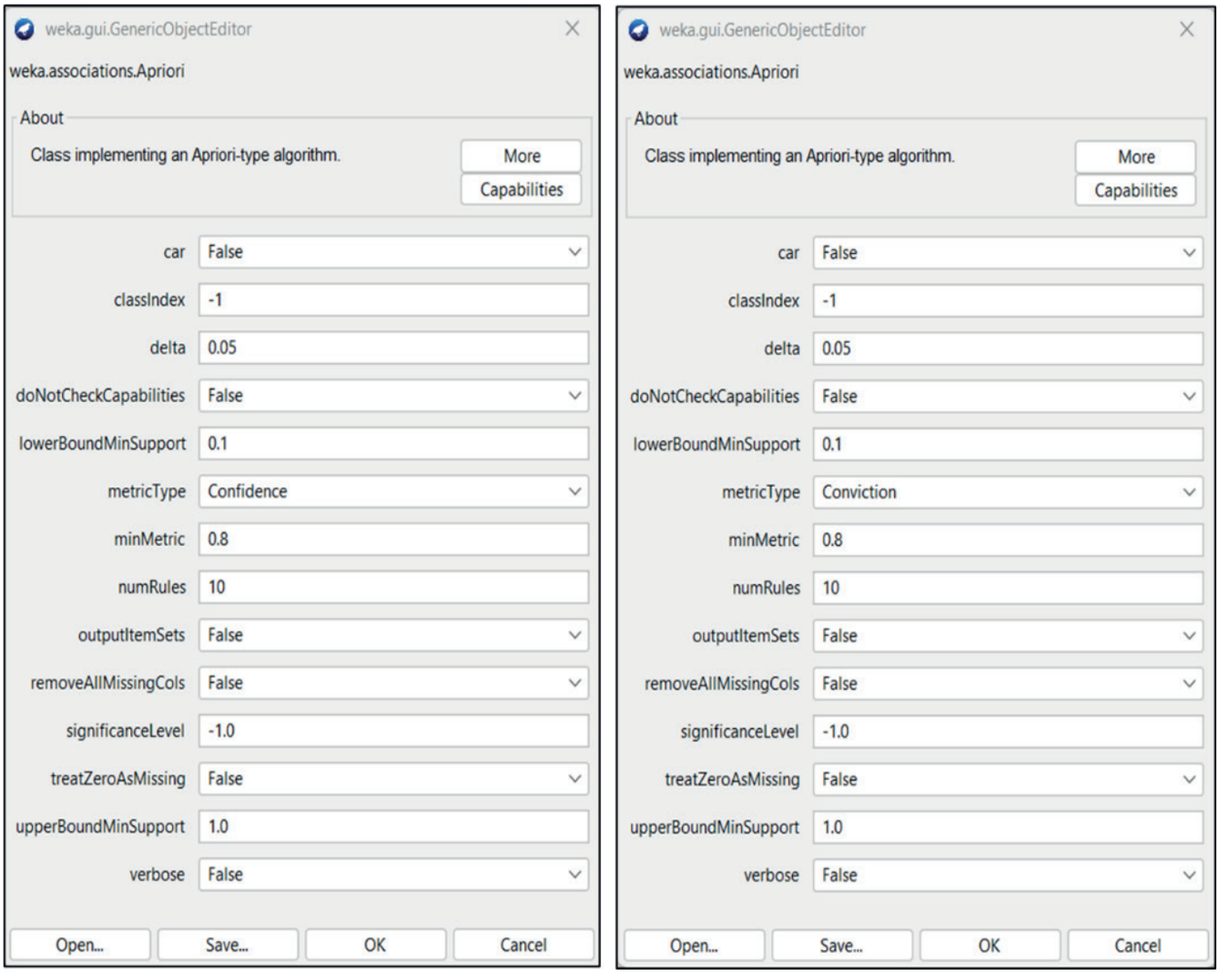
Mineração de Dados

Inicialmente, foi realizada uma série de experimentos com o intuito de identificar os resultados mais promissores a serem classificados dentre os atributos constituintes da base utilizada. Após tais experimentos, o algoritmo *Apriori* mostrou-se oportuno. Trata-se de uma técnica de associação exploratória de aprendizado não supervisionado proposto por Agrawal e Srikant[14]. A associação gera regras que descrevem os padrões mais relevantes em dados nominais. As regras são compostas por precedentes e consequentes, ou seja, a regra contém no precedente um subconjunto de atributos e seus valores e no consequente um subconjunto de atributos que decorrem do precedente (A => B)[15-16].

Avaliação

Após o processo de MD, as regras geradas pelo algoritmo *Apriori* foram avaliadas a fim de apresentar os melhores resultados obtidos. As definições de parâmetros do método *Apriori* foram ajustadas na ferramenta *Weka* conforme as execuções foram realizadas e um único conjunto de configurações foi utilizado como modelo final. De modo geral, foram utilizadas as configurações padrão, havendo ajuste apenas no parâmetro "*minMetric*=0.8", que corresponde ao valor mínimo para confiança aceitável nesta pesquisa. Normalmente o valor mínimo mais aceitável em trabalhos para regras de associação é de 70.0% de confiança. A Figura 1 ilustra as configurações aplicadas.

Figura 1 - Hiperparâmetros do *Apriori* no *Weka*: "*Confidence*" (esquerda) e "*Conviction*" (direita)



O parâmetro "*metricType*" trata-se da especificação da medida de interesse que irá determinar a validade da regra. O conjunto de resultados minerados é ordenado de acordo com essa medida. Para esta pesquisa, foram realizadas duas análises de associação, uma tendo como parâmetro "*Confidence*" e outra com o parâmetro "*Conviction*".

No *Apriori*, o cálculo das probabilidades está fortemente ligado ao conceito de "suporte", que indica o quão frequentemente os itens ocorrem juntos em uma mesma regra. Ou seja, o suporte é dado pela quantidade de ocorrências de um determinado item em relação ao total de itens do conjunto. A métrica "*Confidence*" consiste na probabilidade condicional de um consequente ocorrer em relação a um determinado antecedente. Em outras palavras, é a probabilidade de um antecedente e um consequente aparecerem na mesma transação, independente da ordem[18]. Traduzindo-a para uma equação, a confiança é dada pela divisão do suporte do item consequente e antecedente – na mesma transação – dividido pelo suporte do consequente. O resultado da equação é um número que varia entre 0 e 1 e indica que, quanto mais elevado, maior a probabilidade e, por consequência, a confiança da correlação desses itens nas transações do conjunto. A primeira análise de associação realizada, que utilizou esta métrica, encontrou duas regras que atenderam o mínimo de 0.8 de confiança, conforme ilustra o Quadro 3.

Quadro 3 - Regras de associação encontradas pelo *Apriori* com métrica Confiança

Regra	Associação
1	QUIMICO=Radiação solar 191 ==> DIAG_ESP=Outras Neoplasias Malignas da Pele e doenças relacionadas 187 <conf:(0.98)> lift:(2.71) lev:(0.17) [118] conv:(24.42)
2	QUIMICO=Radiação não ionizante e Agrotóxico 154 ==> ID_OCUPA_N=Produtor agrícola polivalente 124 <conf:(0.81)> lift:(3.03) lev:(0.12) [83] conv:(3.65)

A regra 1 indica que no ano de 2019, 191 trabalhadores que estiveram expostos à "Radiação solar" apresentaram o diagnóstico de "Outras Neoplasias Malignas da Pele e doenças relacionadas" em 187 instâncias, com confiança (*conf*) de 0.98. Já na segunda regra, observou-se que, dos 154 registros de trabalhadores que estiveram expostos a "Radiação não ionizante e Agrotóxico", 124 são produtores agrícolas polivalentes, correspondendo a *conf* de 0.81.

Uma medida de interesse importante e utilizada para definir a validade de uma regra de associação é o *lift*. O *lift* de uma regra de associação indica o quanto mais frequente torna-se B, quando A ocorre[19]. Ou seja, na primeira regra os trabalhadores que estiveram expostos a "Radiação solar" tiveram 2.71 vezes a frequência de desenvolverem diagnóstico de "Outras Neoplasias Malignas da Pele e doenças relacionadas". Na segunda regra, trabalhadores que estiveram expostos a "Radiação não ionizante e Agrotóxico" tiveram 3.03 vezes a frequência de serem "Produtor agrícola polivalente". Nas regras apresentadas neste estudo, o *lift* teve um valor aproximado de 3.00 vezes, o que indicou que os itens A e B (antecedente e consequente) ocorrem mais frequentemente juntos nas duas regras, ou seja, são positivamente dependentes. Outra medida interessante para avaliar o índice de dependências é o *leverage* (*lev*). O *lev* indica a diferença entre o suporte real e o suporte esperado de uma regra de associação, variando de -0.25 e 0.25. Quanto maior o valor apresentado, mais interessante se torna para análise[20]. Ao trazer a análise para o estudo, na primeira regra o *lev* foi de 0.17 e na segunda regra foi de 0.12, ou seja, ambas as regras são positivamente dependentes. Os valores em parênteses expressam a quantidade de vezes que essa dependência entre A e B ocorreu.

A métrica "*Conviction*" é outra forma de medir a associação, mas de uma maneira diferente da anterior. Ela compara o quão provável é de um item consequente ocorrer sem que haja um determinado antecedente. Traduzido para uma equação, a convicção é dada pela subtração de 1 do suporte do antecedente, dividindo pela subtração de 1 da confiança do consequente e antecedente, ou seja, quanto mais acima de 1, mais forte será a relação entre A e B. Nas regras 1 e 2 ocorreram bom índice de relação entre A e B, de 24.42 e de 3.65, respectivamente.

Já a convicção que resulta em 0 ocorre em casos de confiança de 100 de uma correlação, sendo assim, são métricas complementares de compreensão de uma associação. A segunda análise de associação realizada, que utilizou essa métrica, identificou quatro regras, conforme ilustra o Quadro 4.

Quadro 4 - Regras de associação encontradas pelo algoritmo *Apriori* com a métrica Convicção

Regra	Associação
1	QUIMICO=Radiação solar 194 ==> DIAG_ESP=Outras Neoplasias Malignas da Pele e doenças relacionadas 187 conf:(0.96) lift:(3) lev:(0.16) [124] < conv:(16.46)>
2	QUIMICO=Radiação não ionizante e Agrotóxico 154 ==> ID_OCUPA_N=Produtor agrícola polivalente 124 conf:(0.81) lift:(3.45) lev:(0.11) [88] < conv:(3.81)>
3	DIAG_ESP=Outras Neoplasias Malignas da Pele e doenças relacionadas 253 ==> QUIMICO=Radiação solar 187 conf:(0.74) lift:(3) lev:(0.16) [124] < conv:(2.85)>
4	ID_OCUPA_N=Produtor agrícola polivalente 184 ==> QUIMICO=Radiação não ionizante e Agrotóxico 124 conf:(0.67) lift:(3.45) lev:(0.11) [88] < conv:(2.43)>

Ao aplicar as mesmas interpretações dos hiperparâmetros do Quadro 3, foi possível observar que as 4 regras do Quadro 4 apresentaram similaridades com relação a presença de "Radiação solar", "Outras Neoplasias Malignas da Pele e doenças relacionadas", "Radiação não ionizante e Agrotóxico" e "Produtor agrícola polivalente". A primeira regra do Quadro 4 apresentou confiança notavelmente alta de 0.96, indicando os casos em que houve exposição à "Radiação solar" e obtiveram o diagnóstico de "Outras Neoplasias Malignas da Pele e doenças relacionadas". O valor de *lift* de 3.0 norteou uma forte associação e a convicção de 16.46 reforçou a confiabilidade da associação. A segunda regra apresentou confiança de 0.81, o que destaca a exposição simultânea à "Radiação não ionizante e Agrotóxicos" com a ocupação "Produtor agrícola polivalente". O *lift* de 3.45 indica uma associação substancial, e a convicção de 3.81 reforça a validade da relação. A terceira regra apresenta confiança de 0.74, o que indica o diagnóstico de "Outras Neoplasias Malignas da Pele e doenças relacionadas" por consequência a exposição à "Radiação solar". O valor de *lift* sugeriu uma associação significativa, e a convicção de 2.85 destacou a confiabilidade dessa associação. Por fim, na quarta regra a confiança de 0.67 indicou os casos em que a ocupação "Produtor agrícola polivalente" esteve exposta à "Radiação não ionizante e Agrotóxicos". O valor de *lift* e a convicção reforçaram a robustez dessa associação.

DISCUSSÃO

A tarefa de associação escolhida permite a geração de regras que apontam a ocorrência de respostas nos registros da base de dados. A análise dessas medidas, permite identificar quais atributos têm influência na ocorrência de outro atributo. A análise desses dados pode prover hipóteses de intervenções sobre determinados aspectos na saúde e vida dos trabalhadores.

Estudos que usam técnicas de MD buscam encontrar conhecimento em diferentes áreas, o que também acontece na saúde com os mais variados temas. Para Santos[17], após uma pesquisa realizada na base E-Saúde – principal base de dados eletrônica médica sobre prontuários –, os problemas mais comuns encontrados foram: dificuldade de trabalhar os dados; a indisponibilidade de download da base completa; conjunto de dados incompleto e indisponibilidade de formato aberto. Além disso, o autor constatou que a maioria das bases publicadas correspondem aos temas Administração Pública e Infraestrutura Urbana, com poucas dando destaque à área da saúde.

De Oliveira e Costa[18], revela que trabalhadores da agropecuária da região norte do Brasil obtém a categoria ocupacional mais afetada, principalmente o sexo masculino. Embora metade das notificações não apresentasse esta informação, o resultado indica a exposição de diversas categorias ocupacionais aos agrotóxicos, utilizados, principalmente, com finalidade herbicida e inseticida.

Este estudo revelou que o registro indevido e a falta de completude de campos prejudicam a qualidade dos estudos e a vigilância da população exposta aos agrotóxicos. Considerando essas doenças e sua frequente ligação com trabalhadores do meio rural, pode-se presumir que essa ocupação de trabalho tem menos acesso aos recursos de saúde, seja por questões de infraestrutura ou seja por questões da informalidade. Essa suposição para os dados brasileiros também encontra paralelos em outros países, mesmo os desenvolvidos, como abordado por Jakob et al.[19], onde, inclusive, o melhor uso dos sistemas de registro destes casos é um desafio na Europa.

As regras de associação podem ser particularmente úteis para estudar na escolha por tratamentos, nas pesquisas de sintomas e na identificação das características de doenças. Por exemplo, Silva e Nascimento[20], Lima Júnior et al.[21], e Mota e Barros[22] em suas pesquisas em três cidades diferentes do Brasil: Boa Vista/RR (Roraima), Parintins/AM (Amazônia) e Recife/PE (Pernambuco), respectivamente, confirmam que a idade, cor/raça, etilismo e tabagismo são fatores que influenciam no câncer de próstata. Usando tais tipos de informações, a área da saúde pode fornecer melhores recomendações para maximizar o tratamento dos pacientes ou até mesmo, minimizar a gravidade de certas doenças.

Apesar de complexo, os resultados demonstraram que é possível a aplicação deste processo, permitindo a identificação de padrões nos dados, que podem ser avaliados quantitativamente, através de métricas, e qualitativamente, através do julgamento do significado dos padrões identificados. De maneira geral, configura-se um estudo experimental de modelagem inicial para a introdução da mineração de dados na compreensão das correlações existentes na exposição de indivíduos à agrotóxicos.

CONCLUSÃO

Este estudo analisou a base de dados pública do DATASUS buscando a identificação de padrões que podem auxiliar na tomada de decisão preventiva de gestores administrativos/equipes assistenciais a partir das informações obtidas por meio de técnicas de MD. Os resultados apresentados podem incentivar as organizações a se aproximarem de seus objetivos a favor da saúde, qualidade de vida e segurança do trabalhador e a tomarem decisões mais inteligentes, dando suporte aos gestores das Unidades de Saúde em todas as esferas, desde programas de prevenção de novos casos até protocolos de tratamentos. O processo de DCBD aplicado mostrou bom desempenho de regras de associação, o que pode estimular a inclusão deste recurso em organizações de diversos segmentos e para diversas finalidades. Sugere-se, como trabalhos futuros, o estudo da mesma base em diferentes anos para observar possíveis tendências, explorar outros atributos não abordados no presente estudo, aplicar outros algoritmos (quando o caso) e até mesmo utilizar outros hiperparâmetros de métricas.

REFERÊNCIAS

1. Zambolim CM, Oliveira TP, Hoffmann AN, Vilela CEB, Neves D, Anjos FR, et al. Perfil das intoxicações exógenas em um hospital universitário. *Revista de Medicina de Minas Gerais* 2008;18(1):5-10.
2. Brasil. Ministério da Saúde. *Sistemas de informação em saúde.* 2021 [Citado 2024 jan 31]. Disponível em: https://www.gov.br/saude/pt-br/composicao/svs/vigilancia-de-doencas-cronicas-nao-transmissiveis/sistemas-de-informacao-em-saude.
3. Brasil. Ministério da Saúde. Portaria 130, de 12 de fev. de 1999. Institui e formaliza a distribuição de competências dos órgãos do Ministério da Saúde em relação ao Sistema Nacional de Informações em Saúde. – Brasília (DF): *Diário Oficial da União*; 1999 [Citado 2022 ago 17]. Disponível em: http://www.portalsinan.saude.gov.br/images/documentos/ Legislacoes/Portaria_130_12_02_1999.pdf.
4. DATASUS. *Departamento de Informática do SUS.* [Citado 2024 jan 31]. Disponível em: https://datasus.saude.gov.br/sobre-o-datasus/.
5. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. *Advances in Knowledge Discovery and Data Mining.* California: AAAI Press/The MIT Press; 1996.
6. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI magazine*, v.17, n.3; 1996. p.37-37.
7. Lakshmi BN, Raghunandhan GH. A conceptual overview of data mining. In: *2011 National Conference on Innovations in Emerging Technology.* IEEE; 2011. p.27-32.

8. Holloway, Jack; et al. *Evaluating the performance of a predictive modeling approach to identifying members at high-risk of hospitalization.* 2019 [Citado 2024 jan 31]. Disponível em: https://doi.org/10.1080/13696998.2019.1666854
9. Kajungu DK, Selemani M, Masanja I, et al. Using classification tree modelling to investigate drug prescription practices at health facilities in rural Tanzania. *Malar J* 11, 311. 2012 [Citado 2024 jan 31]. Disponível em: https://doi.org/10.1186/1475-2875-11-311
10. Parsaye K. *Intelligent databases: object-oriented, deductive and hypermedia technologies.* New York: John Wiley; 1989.
11. *Machine Learning Project at the University of Waikato in New Zealand.* [Citado 2024 jan 31]. Disponível em: https://www.cs.waikato.ac.nz/ml/index.html.
12. Ministério da Saúde. *Câncer relacionado ao Trabalho.* 2022 [Citado 2024 jan 31]. Disponível em: https://www.gov.br/saude/pt-br/composicao/svsa/saude-do-trabalhador/vigilancia-em-saude-do-trabalhador-vigisat/doencas-e-agravos-relacionados-ao-trabalho/cancer-relacionado-ao-trabalho.
13. Costa CN, Coutinho JV, Magalhães LH, Arbex MA. Descoberta de conhecimento em bases de dados. *FESJ Revista Eletrônica.* 2018 [Citado 2024 jan 31]. Disponível em: https://www.fsd.edu.br/wp-content/uploads/2019/12/artigo9.pdf.
14. Agrawal R, Srikant R. *Fast Algorithms for Mining Association Rules.* Proc. 20th International Conference on Very Large Data Bases, VLDB; 1994. p.478-499.
15. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques* 3. ed. Morgan Kaufmann Publishers Inc, San Francisco, USA; 2011.
16. Witten IH, Frank E, Hall MA. *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers Inc, 3.ed, San Francisco CA US; 2011.
17. Santos WH, et al. *Estudo da base de dados abertos E-Saúde da prefeitura de Curitiba usando técnicas de mineração de dados.* Dissertação de Mestrado. Universidade Tecnológica Federal do Paraná. 2018.
18. De Oliveira Silva SL, Costa EA. Intoxicações por agrotóxicos no estado do Tocantins: 2010–2014. *Vigilância Sanitária em Debate: Sociedade, Ciência & Tecnologia*, v. 6, n. 4; 2018. p.13-22.
19. Jakob MC, Santa D, Holte KA, Sikkeland IJ, Hilt B, Lundqvist P. Occupational health and safety in agriculture – a brief report on organization, legislation and support in selected European countries. *Ann Agric Environ Med.*, 28(3); 2021. p.452-457 [Citado 2024 jan 31]. Disponível em: https://doi.org/10.26444/aaem/140197
20. Silva JS, Nascimento LP. *Fatores Culturais Associados à não Adesão aos Exames Preventivos de Câncer de Próstata em Parintins* [Trabalho de Conclusão, Universidade do Estado do Amazonas]. 2017 [Citado 2024 jan 31]. Disponível em: http://repositorioinstitucional.uea.edu.br//handle/riuea/759.
21. Lima Júnior MM, Reis LO, Ferreira U, Cardoso UO, Barbieri RB, Mendonça GB, et al. Unraveling Brazilian Indian Population Prostate Good Health: Clinical, Anthropometric and Genetic Features. *International braz j urol*, 41(2). 2015 [Citado 2024 jan 31]. Disponível em: https://doi.org/10.1590/S1677-5538.IBJU.2015.02.23.
22. Mota TR, Barros DPO. Perfil dos pacientes com câncer de próstata em hospital de referência no estado de Pernambuco. *Revista brasileira de análises clínicas*, 50(4), 334-338; 2018 [Citado 2024 jan 31]. Disponível em: https://doi.org/10.21877/2448-3877.201900766.