Mineração de dados para a descoberta de associações em microcefalia em adultos

Juliana M. Teixeira[1], **Rogerio Salvini**[1], **Paula V. Nunes**[2]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 131 – 74.001-970 – Goiânia – GO – Brazil

[2]Instituto de Psiquiatria – Universidade de São Paulo (USP)

{julianamelo,rogeriosalvini}@inf.ufg.br, paula@formato.com.br

Abstract. *Despite the vast knowledge about microcephaly in children and newborns, little is known about its consequences in adults. Microcephaly is defined as a congenital malformation characterized by inadequate brain development and a smaller than normal head circumference. The present study aims to look for associations between microcephaly in adults and sociodemographic, clinical, family history, and other diseases. A database of 2,300 cases collected by the bank of brains of the USP Medical School was used in this study. Results using association rules showed relationships involving microcephaly, age, dementia and hypertension. This work aims to contribute to increase knowledge of microcephaly in adults and assist specialists and researchers in the field.*

Resumo. *Apesar do vasto conhecimento sobre a microcefalia em crianças e recém-nascidos, é pouco ainda o que se sabe a respeito das suas consequências em adultos. A microcefalia é definida como uma malformação congênita caracterizada pelo desenvolvimento do cérebro de maneira inadequada, apresentando o perímetro cefálico menor que o normal. O presente estudo tem como objetivo buscar associações entre a microcefalia em adultos e fatores sociodemográficos, clínicos, histórico familiar, e outras doenças. Uma base de dados com 2.300 casos coletados pelo Banco de Cérebros da Faculdade de Medicina da USP foi utilizada neste estudo. Resultados usando regras de associação mostraram relações envolvendo a microcefalia, idade, demência e hipertensão. Este trabalho visa a contribuir para o aumento do conhecimento da microcefalia em adultos e auxiliar especialistas e pesquisadores na área.*

1. Introdução

A microcefalia é uma malformação congênita que tem acometido diversos recém-nascidos no Brasil e é caracterizada pelo desenvolvimento do cérebro de maneira não adequada, onde se apresenta perímetro cefálico menor que o normal. O perímetro cefálico é um parâmetro antropométrico altamente correlacionado com o tamanho cerebral [Weaver and Christian 1980]. O crescimento cerebral, diferentemente de outras partes do corpo, tem aumento rápido durante o primeiro ano de vida, completando 83,6% de seu desenvolvimento [Graham 1967]. Embora não seja comum a medição do perímetro cefálico após o primeiro ano de vida, sabe-se que ele aumenta suavemente até os dezoito anos [Eichorn and Bayley 1962, Nelhaus 1968]. Após o terceiro ano, apesar de lento, o

crescimento reflete 25% do volume cerebral [Roche and Mukherjee 1986]. A fim de facilitar o rastreio, um dos critérios para diagnóstico de microcefalia em recém-nascidos já definido pela Organização Mundial de Saúde é de um perímetro cefálico igual ou inferior a 32 centímetros.

As manifestações de microcefalia podem ser leves ou graves, dependendo do grau de comprometimento no desenvolvimento do cérebro. Entre suas causas estão problemas genéticos e fatores externos, por exemplo, exposição do feto à radiação, uso de drogas, consumo de álcool, desnutrição intrauterina e doenças como rubéola [Rowland 2002]. Em casos leves, a criança com microcefalia pode desenvolver-se normalmente, mas na maioria dos casos há um comprometimento no desenvolvimento do cérebro, trazendo complicações neurológicas, como retardos mentais graves e em situações extremas até mesmo o óbito.

Os estudos a respeito de microcefalia em sua maioria são direcionados a recém-nascidos, já que é a fase em que a maior parte dos casos de microcefalia é identificada. Porém, é importante que a patologia em adultos também seja estudada, já que o comportamento da doença na fase adulta pode apresentar diferenças se comparada a idades menores. No trabalho de Mantovani e Nunes [Mantovani and Nunes 2017] foi realizado um estudo para a definição de um padrão brasileiro para a microcefalia em adultos. Além disso, os autores utilizaram métodos estatísticos para investigar a possível relação da doença com a depressão. Os estudos da microcefalia em adultos ainda são incipientes e são necessárias mais informações sobre os parâmetros brasileiros, sua prevalência e fatores que possam estar relacionados, possibilitando a contribuição com a área médica e o auxílio em pesquisas. Desta forma, o presente trabalho tem como objetivo buscar associações entre a microcefalia e fatores clínicos em adultos. Para isto, foi usada uma abordagem não supervisionada de mineração de dados baseada em regras de associação que, até onde é de nosso conhecimento, ainda não havia sido reportada na literatura para este problema. O algoritmo *Apriori* [Agrawal et al. 1993] foi escolhido para gerar as regras de associação, que são de fácil interpretação para a análise de profissionais especialistas na área.

O restante do artigo é estruturado como a seguir. Na seção 2 é apresentado o método de produção de regras de associação e o algoritmo Apriori. Na seção 3 é apresentada a base de dados utilizada neste trabalho e o pré-processamento feito nos dados antes da aplicação do algoritmo. Na seção 4 são apresentados os resultados obtidos com as regras de associação mais relevantes selecionadas. Na seção 5 é apresentada uma discussão sobre as associações encontradas. Por fim, na seção 6 são apresentadas algumas conclusões e trabalhos futuros.

2. Regras de Associação

A produção de regras de associação é um método de Mineração de Dados com abordagem não supervisionada, isto é, quando não há uma variável dependente a ser prevista, cujo objetivo é encontrar relacionamentos ou padrões frequentes entre conjuntos de exemplos.

Formalmente, considerando $I = \{i_1, ..., i_n\}$ um conjunto de n atributos categóricos chamados de itens e $T = \{t_1, ..., t_m\}$ o conjunto de m transações definido como a base de dados, cada transação em T tem um identificador único e contém um subconjunto de itens de I. Uma regra de associação possui o formato $A \rightarrow B$ com o significado lógico de se A então B, onde A e B são conjuntos disjuntos não vazios, A é

dito antecedente da regra e B consequente, e ambos A e B estão contidos em I.

Para a geração de regras de associação, duas medidas são utilizadas: suporte e confiança. O suporte (Equação 1) de uma regra $A \rightarrow B$ é definido como a porcentagem das transações que possuem os itens de A e B em relação à quantidade total de transações na base de dados, e a confiança (Equação 2) de uma regra de associação $A \rightarrow B$ é definida como a probabilidade B ocorrer dado que A ocorreu.

$$Suporte(A \rightarrow B) = \frac{Ocorrências\ de\ A\ juntamente\ de\ B}{Total\ de\ transações\ na\ base\ de\ dados} \tag{1}$$

$$Confiança(A \rightarrow B) = \frac{Ocorrências\ de\ A\ juntamente\ de\ B}{Ocorrências\ de\ A} \tag{2}$$

Logo, são estabelecidos valores mínimos de suporte e confiança para a geração de regras de associação, onde as regras resultantes do processamento possuem valores de suporte e confiança que satisfaçam tais limites.

Outra medida importante é o *lift* (Equação 3), que estabelece o relacionamento entre o antecedente e o consequente de uma regra de associação. Se o valor do *lift* de uma regra for igual a 1, isso implica que não há relação entre antecedente e consequente. Se o mesmo for maior que 1, significa que se o antecedente ocorrer, é muito provável que o consequente também ocorra, e de forma análoga, se for menor que 1, caso o antecedente ocorra, é pouco provável que o consequente ocorra. Este valor permite avaliar as associações mais fortes que aparecem nas regras.

$$Lift(A \rightarrow B) = \frac{Ocorrências\ de\ A\ juntamente\ de\ B}{Ocorrências\ de\ A\ \times Ocorrências\ de\ B} \tag{3}$$

2.1. Algoritmo Apriori

Para a produção das regras de associação foi utilizado o algoritmo Apriori, introduzido por Agrawal, Imielinski e Swami [Agrawal et al. 1993].

Seu primeiro passo consiste em selecionar *itemsets* frequentes, definidos pelos conjuntos com suporte maior ou igual ao limite estabelecido. Já o segundo e último passo consiste em definir o conjunto de regras confiáveis, que também acontece de acordo com o limite de confiança dado.

Considerando o conjunto total de transações de uma base de dados como na Tabela 1, o algoritmo funciona selecionando conjuntos formados por itens, onde inicia com apenas 1 item, usa-o para formar um conjunto com 2 itens de acordo com o suporte e a combinação dos elementos, e continua a sequência até k, onde $k + 1$ resulta em um conjunto vazio, determinando o fim da primeira parte do algoritmo.

Na Tabela 1 é indicada uma base de dados com 7 itens e 6 transações, onde o 1 significa o pertencimento de um item à uma transação, e o 0 o não pertencimento.

Considerando o suporte de 50% e a confiança de 75%, o algoritmo inicia criando o *1-itemset* (Tabela 2) e calculando seus respectivos suportes.

Tabela 1. Base de dados de transações

Transações	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
1	1	0	1	1	0	1	1
2	1	1	1	1	0	1	0
3	0	0	1	1	1	0	0
4	1	0	0	1	0	1	0
5	1	1	0	1	1	1	0
6	0	0	1	1	1	1	1

Tabela 2. 1-itemset

Itemset	Suporte
item 1	66,7%
item 2	33,3%
item 3	83,3%
item 4	100,0%
item 5	50,0%
item 6	83,3%
item 7	33,3%

Tabela 3. 2-itemset

Itemset	Suporte
item 1, item 3	50,0%
item 1, item 4	66,7%
item 1, item 5	16,7%
item 1, item 6	66,7%
item 3, item 4	83,3%
item 3, item 5	50,0%
item 3, item 6	66,7%
item 4, item 5	50,0%
item 4, item 6	83,3%
item 5, item 6	33,3%

Tabela 4. 3-itemset

Itemset	Suporte
item 1, item 3, item 4	50,0%
item 1, item 3, item 6	50,0%
item 1, item 4, item 6	66,7%
item 3, item 4, item 5	50,0%
item 3, item 4, item 6	66,7%

Os *itemsets* de item 2 e item 7 são então excluídos por não atingirem o suporte mínimo. Com isso é construído então o *2-itemset* (Tabela 3).

Os *itemsets* com os itens 1,5 e 5,6 são também excluídos por não atingirem o suporte mínimo. O *3-itemset* é então construído (Tabela 4).

Para a produção do *4-itemset* é importante notar que para uma combinação de itens em qualquer *n-itemset* frequente, devem ser considerados todos os subconjuntos de *n-1-itemset* envolvidos para essa construção, já que todos devem ter suporte maior que o limite para serem então utilizados na produção de um itemset maior. Por isso, apenas um itemset é considerado na produção do *4-itemset*.

Como o *5-itemset* é vazio, a primeira parte do algoritmo é finalizada.

Tabela 5. 4-itemset

Itemset	Suporte
item 1, item 3, item 4, item 6	50,0%

Tabela 6. Conjuntos representativos

Conjunto Representativo	Itemset	Suporte
1	item 1, item 3, item 4, item 6	50,0%
2	item 1, item 4, item 6	66,7%
3	item 3, item 4, item 5	50,0%
4	item 3, item 4, item 6	66,7%
5	item 3, item 4	83,3%
6	item 4, item 6	83,3%

Tabela 7. Regras conjunto representativo 1

Regra	Confiança
item 3, item 4, item 6 $\rightarrow$ item 1	75,0%
item 1, item 4, item 6 $\rightarrow$ item 3	75,0%
item 1, item 3, item 6 $\rightarrow$ item 4	100,0%
item 1, item 3, item 4 $\rightarrow$ item 6	100,0%

Os conjuntos representativos são então selecionados, iniciando com o *itemset* maior e que possua suporte maior ou igual ao limite. Então os próximos são selecionados se não são subconjuntos de conjuntos representativos maiores e tem suporte maior ou igual ao limite, ou se são subconjuntos de conjuntos representativos mas possuem suportes maiores que os mesmos. Os conjuntos representativos do exemplo é representado pela Tabela 6.

A partir desses conjuntos, a segunda parte do algoritmo é iniciada, onde é considerada a confiança para a produção das regras. Em cada conjunto representativo são dadas as possibilidades de antecedentes e consequentes, e então considerada a confiança de cada possibilidade para então selecionar as regras. Como geralmente as regras são restringidas a terem apenas um item como consequente, o algoritmo considera apenas as possibilidades que atendam essa restrição.

Em relação ao conjunto representativo 1, há 4 possibilidades de regras (Tabela 7).

Como todas as regras respeitam ao limite de confiança, todas são selecionadas. Assim, da mesma forma são feitas as possibilidades de todos os conjuntos representativos (Tabela 8).

Dessa forma, as regras resultantes do algoritmo são aquelas que atingiram pelo menos a confiança mínima, e o algoritmo chega ao fim.

3. Materiais e Métodos

3.1. Base de Dados

A base de dados utilizada neste trabalho foi concedida pelo Grupo de Estudos do Envelhecimento do Cérebro da Faculdade de Medicina da USP, e contém 2.300 casos co-

Tabela 8. Regras dos conjuntos representativos

Regra	Confiança
item 4, item 6 → item 1	80,0%
item 1, item 6 → item 4	100,0%
item 1, item 4 → item 6	100,0%
item 4, item 5 → item 3	100,0%
item 3, item 5 → item 4	100,0%
item 3, item 4 → item 5	60,0%
item 4, item 6 → item 3	80,0%
item 3, item 6 → item 4	100,0%
item 3, item 4 → item 6	80,0%
item 4 → item 3	83,3%
item 3 → item 4	100,0%
item 6 → item 4	100,0%
item 4 → item 6	83,3%

letados desde 2004 pelo Banco de Cérebros, os quais foram doados com o consentimento dos familiares, que também forneceram dados clínicos. Ela é composta por dados socio-demográficos, uso de medicamentos, história clínica, fatores de risco cardiovasculares, história familiar de transtornos psiquiátricos, diagnóstico de demência, e perímetro cefálico, medida que foi utilizada para definir os casos com e sem microcefalia, conforme [Mantovani and Nunes 2017]. No total foram identificados 119 casos de microcefalia, o que corresponde a 5,2% da base de dados. Na Tabela 9 são descritos todos os atributos presentes na base de dados.

3.2. Seleção e pré-processamento dos dados

Devido ao caráter qualitativo dos itens das regras de associação, os dados da base de dados que eram categóricos foram mantidos, e os que eram numéricos foram modificados para representar então uma categoria. Todas as modificações na base de dados citadas neste trabalho foram realizadas em linguagem R pela plataforma RStudio na versão 3.4.3.

Primeiramente, os nomes dos atributos foram padronizados na língua inglesa com o intuito de facilitar futuras contribuições de âmbito internacional.

Os atributos que medem a escala de demência (*CDR*) e de perda de cognição (*iqcode*) foram categorizados de acordo com suas definições, onde o *CDR* foi renomeado como *dementia*, e o *iqcode* como *iqcode_cat*, definindo um valor maior ou igual a 3,4 como perda de cognição.

Os dados a respeito do peso estavam distribuídos em função do valor do peso, valor do Índice de Massa Corpórea (IMC) e categorização do IMC (*weight*, *BMI* e *BMI_pattern*), estabelecendo redundância de informações. Desta maneira, foram defi-

nidos apenas os atributos de sobrepeso e subpeso (*BMI_overweight* e *BMI_underweight*), que consistem nas informações mais relevantes para este estudo. O mesmo ocorre com os atributos altura e padrão de altura (*height* e *height_pattern*), onde apenas o segundo é considerado.

A noção de nível socioeconômico foi mantida, sendo modificada em *lowsociallevel_bin* para indicar apenas se este é baixo ou não.

A idade foi estratificada convenientemente para maior ou igual a 70 anos e menor que 70 anos pela suspeita durante o estudo de que a demência se manifesta em idades relativamente mais avançadas, como pode ser visto nas seções de Resultados e Discussão.

Em relação à etnia, para uma melhor distribuição entre os seus valores, uma vez que brancos constituiam 66,5% dos casos na base de dados, foi criada uma nova variável (*race_white_notwhite*) separando apenas a etnia branca da não branca.

O atributo de perímetro cefálico (HC) foi removido já que ele provia a mesma indicação do atributo *microcephaly*.

Além disso, os atributos contínuos referentes a volume e peso do cérebro (*brainvolume* e *brainweight*) foram tratados de forma parecida. Como na base de dados o valor mínimo para *brainvolume* era de 500 ml e o maior de 1750 ml, o atributo foi discretizado em 5 intervalos de 250 ml, iniciando pelo menor valor e terminando com o maior. Equitativamente, o atributo *brainweight* em que seu menor valor era de 720 g e o maior de 2045 g, foi discretizado em 7 intervalos de 200 g, iniciando pelo valor 700 g e terminando com 2100 g, com o intuito de uniformizar a divisão dos intervalos.

Por fim, os outros atributos foram mantidos já que estavam em formato satisfatório para a aplicação do algoritmo Apriori.

A descrição dos atributos, após o processamento descrito, e que foram usados como entrada pelo Apriori para a geração das regras de associação, é mostrada na Tabela 10.

3.3. Geração das regras de associação

Neste trabalho o algoritmo Apriori foi executado utilizando a plataforma RStudio e o pacote Rsenal, mais especificamente as bibliotecas arules [Hashler et al. 2005] e arulesViz [Hashler 2017], onde são dispostas as regras de associação e algumas visualizações para as mesmas.

A definição dos valores limiares de suporte e confiança foi feita considerando que a porcentagem de casos de microcefalia na base de dados é de aproximadamente 5%. Portanto, após a verificação de cada passo feito, principalmente no pré-processamento, muitas atividades, como testes de novos valores de suporte e confiança e formas de seleção de dados, foram reavaliadas e executadas novamente com alterações ponderadas para melhores resultados, e o valor mínimo estabelecido de suporte foi 1% e de confiança 30%.

4. Resultados

A execução do algoritmo Apriori resultou em 34.355 regras de associação. Foram então separadas todas as regras com a ocorrência de microcefalia para o estudo de suas relações.

Tabela 9. Descrição dos atributos originais da base de dados

Atributo	Descrição	Tipo de Dado	Valores Possíveis
CDR	escala que mede intensidade de demência	categórico	0 - nada, 0.5 - em risco, 1 - demência leve, 2 - demência moderada, 3 - demência grave
iqcode	escala que mede piora cognitiva nos últimos 10 anos	contínuo	3,0 a 3,3 - sem perda, a partir de 3,4 - perda gradativa
age	valor numérico para idade	contínuo	de 28 a 106
gender	classificação de gênero	categórico binário	male, female
race	categorias raciais	categórico	branco, pardo, negro, asiático
educationlevel_cod	categorias de escolaridade	categórico	analfabeto, fundamental I, fundamental II ou mais
schooling	anos de estudos	contínuo	de 0 a 25
height	valor numérico para altura	contínuo	de 88 a 198
height_pattern	categorização da altura	discreto	1 - ¡ 150 cm, 2 - de 150 cm a 169 cm, 3 - de 160 cm a 169 cm, 4 - de 170 cm a 179 cm, 5 - $\geq$ 180
weight	valor numérico para peso	contínuo	de 23 kg a 189 kg
BMI	índice de massa corpórea	contínuo	de 10 a 58
BMI_pattern	classificação de BMI	categórico	0 - abaixo do peso, 1 - normal, 2 - obesidade
cognition	baseada em CDR	categórico	0 - normal, 1 - em risco, 2 - demência
social_level_cod	categorias de classe socioeconômica	categórico	1 - classe A ou B, 2 - classe C, 3 - classe D ou E
brainvolume	valor numérico para volume do cérebro	contínuo	de 500 ml a 1750 ml
brainweight	valor numérico para peso do cérebro	contínuo	de 720 g a 2045 g
HC	circunferência da cabeça	contínuo	de 45 cm a 66 cm
microcephaly	presença ou ausência de microcefalia	categórico binário	yes, no
HTN	presença ou ausência de hipertensão arterial	categórico binário	yes, no
DM	presença ou ausência de diabetes mellitus	categórico binário	yes, no
CAD	presença ou ausência de doença arterial coronariana	categórico binário	yes, no
CHF	presença ou ausência de insuficiência cardíaca digestiva	categórico binário	yes, no
DLP	presença ou ausência de dislipidemia (colesterol alto)	categórico binário	yes, no
stroke	presença ou ausência de AVC	categórico binário	yes, no
arrhythmia	presença ou ausência de arritmia cardíaca	categórico binário	yes, no
alcoholism	presença ou ausência de etilismo	categórico binário	yes, no
smoking	presença ou ausência de tabagismo	categórico binário	yes, no

Tabela 10. Descrição dos atributos utilizados para a geração das regras de associação

Atributo	Descrição	Tipo de Dado	Valores Possíveis
dementia_bin	presença ou ausência de demência	categórico binário	yes, no
iqcode_cat	perda ou não de cognição	categórico binário	yes, no
age_cat	idade estratificada	categórico binário	$<$70 anos , $\geq$ 70 anos
gender	classificação de gênero	categórico binário	male, female
race_white_notwhite	raça branca ou não branca	categórico binário	yes, no
educationlevel	nível de escolaridade	categórico	illiterate, 1-4 years, 5 years or more
height_pattern	categorização da altura	categórico	small, medium, mid-tall, tall, very tall
BMI_underweight	presença ou ausência de subpeso	categórico binário	yes, no
BMI_overweight	presença ou ausência de sobrepeso	categórico binário	yes, no
brainvolume	discretização para o volume do cérebro	discreto	de 500 ml a 1750 ml com intervalos de 250 ml
brainweight	discretização para o peso do cérebro	discreto	de 700 g a 2100 g com intervalos de 200 g
lowsociallevel_bin	se é ou não de um baixo nível social	categórico binário	yes, no
microcephaly	presença ou ausência de microcefalia	categórico binário	yes, no
HTN	presença ou ausência de hipertensão arterial	categórico binário	yes, no
DM	presença ou ausência de diabetes mellitus	categórico binário	yes, no
CAD	presença ou ausência de doença arterial coronariana	categórico binário	yes, no
CHF	presença ou ausência de insuficiência cardíaca digestiva	categórico binário	yes, no
DLP	presença ou ausência de dislipidemia (colesterol alto)	categórico binário	yes, no
stroke	presença ou ausência de AVC	categórico binário	yes, no
arrhythmia	presença ou ausência de arritmia cardíaca	categórico binário	yes, no
alcoholism	presença ou ausência de etilismo	categórico binário	yes, no
smoking	presença ou ausência de tabagismo	categórico binário	yes, no

Tabela 11. Regras de associação selecionadas

Id	Regra	Suporte	Confiança	Lift
1	microcephaly=yes, dementia_bin=yes $\rightarrow$ iqcode_cat=cognition lost	1,3%	100,0%	5,6
2	iqcode_cat=cognition lost, microcephaly=yes $\rightarrow$ dementia_bin=yes	1,3%	100,0%	5,6
3	microcephaly=yes, dementia_bin=yes $\rightarrow$ age_cat= $\geq$ 70	1,16%	87,9%	1,63
4	microcephaly=yes, age_cat= $\geq$ 70 $\rightarrow$ dementia_bin=yes	1,2%	42,0%	2,42
5	microcephaly=yes, dementia_bin=no $\rightarrow$ age_cat= $<$70	1,83%	54,1%	1,17
6	microcephaly=yes, dementia_bin=no $\rightarrow$ age_cat= $\geq$70	1,6%	45,9%	0,85
7	microcephaly=yes $\rightarrow$ HTN=yes	2,2%	45,4%	0,8
8	microcephaly=yes $\rightarrow$ HTN=no	1,8%	37,0%	1,2
9	microcephaly=no $\rightarrow$ HTN=yes	56,9%	59,8%	1,0
10	microcephaly=no $\rightarrow$ HTN=no	30,4%	32,0%	1,0

Para a seleção das regras mais relevantes, foram considerados os valores de confiança e *lift*, já que o suporte se apresenta baixo devido à pouca presença de casos de microcefalia na base de dados, além da análise de variáveis de interesse feita pela médica especialista coautora deste trabalho. Regras com *lift* maior ou menor que 1,0 foram consideradas de maior interesse, já que assim pode-se estabelecer o tipo de relação como descrito na seção 2.2.

As regras de associação selecionadas são apresentadas na Tabela 11.

5. Discussão

As regras 1 e 2 apresentadas na Tabela 11 reforçam a relação de que ao possuir microcefalia e demência, há a perda de cognição captada pela escala IQCODE, assim como a presença de microcefalia e perda de cognição tem como consequência a presença de demência.

De acordo com resultados obtidos relacionando microcefalia com certos níveis de retardo mental, é possível reforçá-los com as regras de 3 a 6, que são complementares ao mostrar que a microcefalia está relacionada à demência. As regras 3 e 4 declaram que caso haja microcefalia e demência, é muito possível que a idade seja maior ou igual a 70 anos, e que a ocorrência de microcefalia juntamente com uma idade maior ou igual a 70 anos resulta em demência. Como elas possuem o valor *lift* maior que 1, é conhecido que para cada regra se o antecedente ocorrer é muito provável que o consequente ocorra. Portanto, foi concluído que muito certamente em idades menores a demência não tenha ocorrido devido ao tempo de vida, o que é constatado com as regras 5 e 6, já que apontam que a presença de microcefalia juntamente com a ausência de demência ocorre possivelmente em idades menores de 70 anos e dificilmente em idades maiores ou iguais a 70 anos.

Já as regras de 7 a 10 propõem um baixo relacionamento entre microcefalia e hipertensão. Como pode ser visto, não ter microcefalia não é associado a ter ou não pressão sanguínea alta, devido ao valor de *lift* igual a 1, porém há uma baixa associação considerando que ao se ter microcefalia é relativamente possível que não se tenha hipertensão, assim como é difícil que se tenha pressão alta.

Além disso, também podemos destacar como resultado deste estudo, que as demais regras de associação produzidas pelo Apriori não apresentaram relações relevantes entre a microcefalia e outros fatores presentes na base de dados utilizada no presente estudo.

6. Conclusão

A microcefalia é uma malformação congênita que pode ser identificada pela medição do perímetro cefálico, com casos leves a graves, determinados de acordo com o nível de comprometimento no desenvolvimento do cérebro e sua causa.

Neste trabalho foi utilizada a base de dados com 2300 casos coletados desde 2004 pelo Banco de Cérebros do Grupo de Estudos do Envelhecimento do Cérebro da Faculdade de Medicina da USP, composto por variáveis clínicas e alterações de perímetro cefálico, possuindo cerca de 5% casos de microcefalia.

Com o objetivo de descobrir novas informações relacionadas à microcefalia em adultos, foi utilizada uma abordagem não-supervisionada através do método de mineração

de dados para produção de regras de associação, mais especificamente com o algoritmo Apriori. Sendo assim, os dados foram pré-processados em conjunto com a médica especialista parceira do projeto, com a discretização e seleção dos atributos para a aplicação do algoritmo.

Os resultados foram filtrados de acordo com a variável de interesse, estudados, visualizados e então selecionados de acordo com as informações mais significativas adquiridas, respeitando o processo de Descoberta de Conhecimento em Base de Dados (Knowledge Discovery in Databases - KDD) [Fayyad et al. 1996].

As regras confirmam que, juntamente à associação da idade categorizada em maior ou igual a 70 anos e menor que 70 anos, há a relação entre microcefalia e demência, o que caracteriza um tempo maior de vida para a ocorrência de demência. Além disso, mostram também a relação entre a ausência de hipertensão e a presença de microcefalia, onde é possível que pessoas com microcefalia estejam menos propensas a terem hipertensão. Há ainda a indicação que outros fatores presentes na base de dados, não teriam associações relevantes com a microcefalia.

Este trabalho é um dos poucos encontrados na literatura voltado para o estudo da microcefalia em adultos. Além disso, ele adiciona uma abordagem alternativa, usando uma análise multifatorial, para encontar associações entre a microcefalia em adultos e outras variáveis clínicas. Dentre as limitações do estudo está a baixa porcentagem de casos com microcefalia na base de dados, totalizando 5,2% da mesma, o que fez necessária a adaptação da interpretação dos resultados do algoritmo Apriori, assim como da seleção das regras.

Para estudos futuros a base de dados será expandida com novas variáveis que incluirão dados com medições clínicas, mais precisos e confiáveis do que aqueles fornecidos por familiares, e que possibilitão a descoberta de mais associações relevantes com a microcefalia em adultos.

Referências

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, pages 207 – 216, Washington DC, United States.

Eichorn, D. H. and Bayley, N. (1962). Growth in head circumference from birth through young adulthood. *Child Dev.*, 33:257–271.

Fayyad, U., Piatestsky-Skapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*, pages 37 – 54.

Graham, C. G. (1967). Effect of infantile malnutrition in growth. *Fed Proc.*, 26:139–143.

Hashler, M. (2017). Arulesviz: Visualizing association rules with r. *R Journal*, 9(2):163 – 175.

Hashler, M., Grün, B., and Hornik, K. (2005). Arules – a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1 – 25.

Mantovani, R. O. and Nunes, P. V. (2017). Estudo de microcefalia em adultos: definição de um padrão brasileiro e sua relação com depressão maior. In *XXXV Congresso Brasileiro de Psiquiatria.*, São Paulo, Brasil.

Nelhaus, G. (1968). Head circumference from birth to eighteen years. practical international and interracial graphs. *Pediatrics vol*, 41:106 – 114.

Roche, A. F. and Mukherjee, D. (1986). Head circumference growth patterns: Birth to 18 years. *Hum Biol.*, 58:893 – 906.

Rowland, L. (2002). *Merritt: Tratado de Neurologia*. Rio de Janeiro: Guanabara Koogan.

Weaver, D. D. and Christian, J. C. (1980). Familial variation of head size and adjustment for parental head circumference. *J Pediatr.*, 96:990 – 994.