COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 122 (2015) 175–181

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Using association rule mining to identify risk factors for early childhood caries

Vladimir Ivančević[a,*], Ivan Tušek[b], Jasmina Tušek[c], Marko Knežević[a], Salaheddin Elheshk[a], Ivan Luković[a]

[a] University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21 000 Novi Sad, Serbia
[b] University of Novi Sad, Medical Faculty, Hajduk Veljkova 3, 21 000 Novi Sad, Serbia
[c] Private Dental Office Palmadent, Mažuranićeva 4, 21 131 Petrovaradin, Serbia

ARTICLE INFO

Article history:
Received 9 February 2015
Received in revised form
19 June 2015
Accepted 20 July 2015

Keywords:
Early childhood caries
Risk factor
Association rule mining
Objective measure of interestingness
Data mining

ABSTRACT

Background and objective: Early childhood caries (ECC) is a potentially severe disease affecting children all over the world. The available findings are mostly based on a logistic regression model, but data mining, in particular association rule mining, could be used to extract more information from the same data set.

Methods: ECC data was collected in a cross-sectional analytical study of the 10% sample of preschool children in the South Bačka area (Vojvodina, Serbia). Association rules were extracted from the data by association rule mining. Risk factors were extracted from the highly ranked association rules.

Results: Discovered dominant risk factors include male gender, frequent breastfeeding (with other risk factors), high birth order, language, and low body weight at birth. Low health awareness of parents was significantly associated to ECC only in male children.

Conclusions: The discovered risk factors are mostly confirmed by the literature, which corroborates the value of the methods.

1. Introduction

Early childhood caries (ECC) is a special form of caries of the primary dentition that affects the teeth after eruption, has a rapid progression, and later leads to a number of symptoms and complications. The American Academy of Pediatric Dentists (AAPD) defined ECC as "the presence of one or more decayed (noncavitated or cavitated lesions), missing (due to caries), or filled tooth surfaces in any primary tooth in a child under the age of six" [1]. Available data, often grouped into different age ranges, indicate dramatic variation in prevalence across different parts of the world, e.g., from 1% to 12% in Australia, England, Sweden, and Finland [2,3], 19.3% in USA [4], 28.4% in northeastern Brazil [5], 29.1% in central Trinidad [6], 44% in southwestern India [7], 56.2% in Poland [8], from 59% to 94% in northern Philippines [9], and 98.9% in Canada's northern Manitoba [10].

Although ECC is widely spread in Serbia, it has not been a frequent research topic. According to the data used in the present study, ECC prevalence in preschool children in the South Bačka area (SBA) of the Province of Vojvodina, Serbia, was 30.5%. This lies in the moderate category in relation to the low prevalence reported in UK, Sweden, and Finland [5] and the high prevalence in the Middle-East [11], some African countries [12], Hispano-Americans [13], and in some regions of Australia [14]. In 1998, Vulović and Carević [15] reported that the ECC prevalence in three-year-old children in SBA was

* Corresponding author. Tel.: +381 21 485 2441; fax: +381 21 6350727.
E-mail address: dragoman@uns.ac.rs (V. Ivančević).

http://dx.doi.org/10.1016/j.cmpb.2015.07.008
0169-2607/

22.07%. It appears that ECC in Serbia may be on increase, which could be linked to the rapid decrease in living standards, therapeutic approach to ECC treatment, as well as to specific demographic, psychosocial and behavioral characteristics of the environment. For these reasons, researchers need ECC models that correspond to the current epidemiological situation in the affected area.

In a literature review on caries models [16], logistic regression (LR) analysis appears to be the most prominent, while the only study that featured data mining (DM) relied on classification and regression tree (CART) analysis [17]. In a comparison between LR and two DM techniques, namely CART and artificial neural networks (ANNs), Gansky [18] concluded that Knowledge Discovery and Data Mining (KDD) methods may provide advantages over traditional statistical methods in dental data. In more recent caries studies that employed DM, predictive models in Japan have included C 5.0 [19], ANNs [19], and CART [20], of which the last one has been applied in the USA as well [21]. According to the cited studies and to the best of our knowledge, DM-based approaches to ECC examination have been published for the USA and Japan (primarily concerning decision trees and ANNs), but rarely, if ever, for Serbia or Europe.

For ECC in SBA, various individual risk factors were identified in a study that relied on LR [22]. The collection of the relevant data sample about ECC was a challenge because such data are scarce in Serbia. In an attempt to extract additional knowledge that could not be provided by the previously employed statistical methods, we are now performing DM, namely association rule mining (ARM), on the same data. ARM, which is a group of DM techniques used to detect associations in data, may be utilized in an ECC study to detect relationships between ECC and potential risk factors.

When employing LR, it may be difficult to understand impact of an individual risk factor or interplay between multiple risk factors. Researchers typically need to formulate a hypothesis for each risk factor combination before doing formal evaluation, which may become practically infeasible even for a moderately sized set of variables. On the other hand, in DM, many patterns may be extracted in a single run, but many resulting formats are of low readability. ARM may be used to avoid these problems because it provides:

- numerous readable patterns (rules) that describe interaction between variables;
- more straightforward interpretation than for the LR coefficients; and
- numerous interpretable measures of rule interestingness, which facilitate identification of relevant rules and rule comprehension.

Nonetheless, ARM may produce an enormous number of rules, but this may be alleviated by rule pruning and ranking. Therefore, in this exploratory study, we first performed ARM to generate ECC-related association rules (ARs) and then, with the goal of uncovering ECC risk factors and their interaction, we pruned, ranked, and inspected the highly ranked association rules. More information about the used data set and ARM methods is provided in Section 2. The uncovered risk factors, as well as the comparison of our findings to those from other studies and regions, are presented in Section 3.

2. Material and methods

2.1. Data Set

2.1.1. Data collection

The aim of the data collection was to investigate potential risk factors and the prevalence of ECC, as well as the degree of severity (white spot lesion, type 1-5) among different social and ethnic groups of preschool children in SBA, Republic of Serbia.

The survey was a cross-sectional analytical study of the 10% sample of preschool children in SBA, aged 13–71 months of both sexes, and different ethnicities, social status, and human environmental (urban, rural) groups. The presence or absence of ECC was recorded depending on the presence of no-cavity caries (white spot lesions), or cavity caries. All primary teeth were examined and caries was recorded using WHO recognized indices DMFT and DMFS, which are calculated as the total number of decayed (D), missing due to caries (M), and filled (F) teeth (T) and surfaces (S), respectively [23]. The diagnosis and the clinical form of ECC were defined by dental check-up according to the Wyne's modified criteria [22].

Epidemiological data of the different social and ethnic groups, as well social status, habits, attitudes, behavior and health knowledge, were obtained by the interview of the parents of the examined children through a series of closed questions (see Supplementary materials 1). Unerupted or congenitally missing teeth were excluded from the DMFT and DMFS scores.

Ethical considerations: All identifiable personal information was adequately disguised in the data in order to preserve the anonymity of the individuals involved. This study was approved by the Committee of Human Research of the Medical Faculty of Novi Sad, University of Novi Sad.

2.1.2. Data overview

In the collected data set, there are 341 individual records collected at different locations in SBA (see Supplementary materials 2). Besides the binary variable denoting ECC presence/absence, there are 35 categorical variables carefully selected by the domain experts as potential risk factors (Table 1).

2.2. Association rule overview

An association rule (AR) in this study is of the form $A \Rightarrow B$, where A (the left-hand-side or the antecedent) and B (the right-hand-side or the consequent) are disjoint sets of attribute-value pairs (AVPs), e.g.,

$$attribute\,A = value\,A1,\ attribute\,B = value\,B1 \Rightarrow attribute\,C = value\,C1.$$

The rule is valid for a case from the data set if all the AVPs from the rule hold true for the particular case. Support (SUPP) of a rule denotes the percentage of cases from the whole data