Artificial Intelligence in Medicine 74 (2016) 44–52

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining

Giulia Toti[a,*], Ricardo Vilalta[a], Peggy Lindner[d], Barry Lefer[b,c], Charles Macias[e], Daniel Price[d]

[a] Department of Computer Science, University of Houston, Philip Guthrie Hoffman Hall, 3551 Cullen Blvd., Room 501, Houston, TX 77204-3010, USA
[b] Department of Earth and Atmospheric Sciences, University of Houston, Science & Research Building 1, 3507 Cullen Blvd, Room 312, Houston, TX 77204-5007, USA
[c] Now at: Earth Sciences Division, NASA Headquarters, 300 E St SW, Washington, DC 20546, USA
[d] Honors College, University of Houston, M.D Anderson Library, 4333 University Dr, Room 212, Houston, TX 77204-2001, USA
[e] Department of Pediatrics, Baylor College of Medicine/Texas Children's Hospital, One Baylor Plaza, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form
22 November 2016
Accepted 23 November 2016

Keyword:
Association rule mining
Rule redundancy
Risk assessment
Multiple exposures
Pediatric asthma
Outdoor pollution

ABSTRACT

Objectives: Traditional studies on effects of outdoor pollution on asthma have been criticized for questionable statistical validity and inefficacy in exploring the effects of multiple air pollutants, alone and in combination. Association rule mining (ARM), a method easily interpretable and suitable for the analysis of the effects of multiple exposures, could be of use, but the traditional interest metrics of support and confidence need to be substituted with metrics that focus on risk variations caused by different exposures.
Methods: We present an ARM-based methodology that produces rules associated with relevant odds ratios and limits the number of final rules even at very low support levels (0.5%), thanks to post-pruning criteria that limit rule redundancy and control for statistical significance. The methodology has been applied to a case-crossover study to explore the effects of multiple air pollutants on risk of asthma in pediatric subjects.
Results: We identified 27 rules with interesting odds ratio among more than 10,000 having the required support. The only rule including only one chemical is exposure to ozone on the previous day of the reported asthma attack (OR = 1.14). 26 combinatory rules highlight the limitations of air quality policies based on single pollutant thresholds and suggest that exposure to mixtures of chemicals is more harmful, with odds ratio as high as 1.54 (associated with the combination day0 SO_2, day0 NO, day0 NO_2, day1 PM).
Conclusions: The proposed method can be used to analyze risk variations caused by single and multiple exposures. The method is reliable and requires fewer assumptions on the data than parametric approaches. Rules including more than one pollutant highlight interactions that deserve further investigation, while helping to limit the search field.

1. Background and objectives

The adverse impact of air pollution on health is well established, and estimates of total mortality and of connections to individual disease suggest that even relatively low levels may impose substantial health burdens [1]. Exposure to pollutants has been linked to numerous health outcomes, including both the inception and triggering of asthma in children. Asthma is a chronic respiratory disease that is responsible for thousands of deaths every year in the U.S. and affects 1 in every 12 people [2]. Children are the most vulnerable to asthma because of their developing, narrower airways. They also inhale more air per pound of body weight than adults, which causes them to inhale, in proportion, a higher quantity of pollutants [3]. Available literature tends to agree on the existence of a positive correlation between asthma and outdoor pollutants, concerning both incidence and cases of exacerbation [4–7]. The assessment of the impact of the single chemicals present in the air still poses a challenge, due to the difficulty in controlling and separating the exposures [8]. The six criteria pollutants (carbon monoxide, lead, nitrogen oxides, ground-level ozone, particulate matter, and sulfur oxides) have been extensively monitored and individually regulated since the Clean Air Act of 1970. Currently, the US Environmental Protection Agency (EPA) issues air quality

* Corresponding author at: Honors College, M.D. Anderson Library, 4333 University Dr, Room 212, Houston, TX, 77204–2001, USA.
E-mail address: giulia.toti@kcl.ac.uk (G. Toti).

http://dx.doi.org/10.1016/j.artmed.2016.11.003
0933-3657/

warnings based on predicted high levels of any one of the pollutants [9]. Unfortunately, these warnings fail to account for potential combinatory effects of the chemicals, which may result in decreased effectiveness and unclear action plans for asthmatics.

There are multiple reasons for favoring the analysis of the effects of mixtures over single pollutants. First of all, because single chemicals have different effects on the respiratory and circulatory systems [5], it is reasonable to suspect that the action of a pollutant could be more harmful if another chemical has already weakened a sensitive part of the organism. This synergic action has already been observed in studies on combinatory effects of pollutants and aeroallergens [50,51]. Second, The chemicals in the atmosphere manifest high correlation between each other, with seasonal weather patterns, and with other seasonal or meteorological causes, such as humidity or pollen. An analysis with standard statistical approaches, such as logistic regression, would be inadequate to identify the effects of single pollutants on the risk of asthma exacerbation, because of well-known problematics in handling highly correlated variables. Furthermore, controlling for single pollutants, as for current EPA standard, may not be particularly useful if chemicals are always present in the atmosphere as a mixture. The environmental health community, including regulators, epidemiologists, and health practitioners, encourages the development of new paradigms to explore the diverse contributions of multiple air pollutants to health outcomes [10].

Relative strengths and weaknesses of different methodological approaches must be assessed for any new multi-pollutant paradigm to develop, although it is unlikely that a single approach will fit all the needs of the community of researchers. Several methods have been explored, and while concerns have been expressed about the effectiveness of traditional epidemiological approaches [11], interest is rising around techniques from the field of data mining and knowledge discovery. In [12] the authors provided a review of various statistical methods used to evaluate the effects of combinations of chemicals. Logistic regression, the gold standard in epidemiological studies, is described as struggling with assumptions of linearity with the logarithm of the odds, pollutants collinearity, difficulty of accounting for multiple interactions between pollutants, and measurement and sampling error [13–18]. Hierarchical Bayesian methods are cited as showing promise, since the approach allows for averaging over different models and retains the interpretability of traditional regression [19,20]. Other recent data mining approaches using sparse principal component analysis [21] and smoothing algorithms for the regressions may be able to provide an overall sense of which environmental variables are most responsible for asthma incidence and acute episodes. However, the existing approaches to the multi-pollutant problems remain poorly suited for isolating clinical and policy relevant effects of multiple pollutants [12]. Classification and Regression Trees (CART) are another viable option, and an attempt to identify and characterize harmful mixtures of exposures with CART has been proposed by Gass et al. [42]. The authors use a Poisson generalized linear model to iteratively determine the exposure with the smallest p-value, and select it as split node. The main problem with this approach lays in the hierarchical structure typical of a classification tree, which could potentially miss synergic actions between chemicals. CART would not recognize a dangerous mixture unless at least one of the included chemicals is identified to be harmful independently (nodes in a classification tree are selected in order, on bases such as changes in entropy in the resulting subgroups). The authors also state that this method is susceptible to collinearity, which could result in the selection of the exposure with the smallest measurement error, rather than the real causal exposure. The wrong selection of a node would affect all subsequent branches of the tree. Finally, approaches that allow for treating air pollution as a homogenized mixture, especially in the context of indoor exposures, have been developed [22], although skeptics can point to the growing risk of ecological fallacies, and of data massaging based on researcher bias, that emerge when the individual researchers are categorizing groupings of interest that rely on interpretation at the same time that they are preparing and analyzing the data.

In response to this context, our research team chose to investigate a variant of an association rule mining (ARM) algorithm. Compared to other methods, ARM has excellent interpretability, even for people who do not have data mining expertise. For this reason, ARM has found several applications in the medical domain, including research on chronic hepatitis, septic shock, heart disease, association deficit disorder, cancer prevention, and more [23–27].

Beside interpretability, other reasons make ARM a valid candidate for shedding more light on the difficult problem of correlation between asthma and outdoor pollution. First, the rules produced by this algorithm are capable of summarizing the impact of several factors in combination in a non-hierarchical fashion. In other words, the risk produced by a mixture of chemicals is not modeled as the effect of a pollutant of interest worsened by the simultaneous presence of other substances, but rather as the total effect of the exposures. Such combinatory effects include both joint effects between pollutants (two or more chemicals acting simultaneously on the body to increase the overall chance of an asthma attack) and interactions (synergic action of two or more chemicals that causes more harm than the combined exposure to the single elements). Second, our modified version of association rule mining retains all original risk factors and allows for the analysis of their possible combinations, shifting the work of establishing statistical significance toward understanding which ones of the proposed rules are meaningful (previous research on ARM has produced a considerable variety of useful metrics that establish statistical significance [43]). Finally, this approach limits unverifiable researcher presuppositions, such as hypothesis of cause-effect through time. This is an important improvement considering recent attacks on the possibility of meeting those conditions, even under the most carefully constructed randomized control trials. A mathematical method that does not rely on implicitly linear constructions of cause and effect should be preferred [28].

When applying well-known ARM algorithms, like *Apriori* [44], to extract new knowledge from data, one often has to face the limitations posed by the *support-confidence framework*. A detailed description of the meaning of support and confidence is offered in the Method section (2). For now, it is sufficient to know that this implies looking for frequent and strong associations in the data. In epidemiological studies, however, relevant associations may not be particularly strong or frequent (in a typical population, subjects with the condition of interest are outnumbered by healthy subjects). If we want to employ ARM to extract knowledge from data in this scenario, we must allow it to search for less frequent associations, and then use other criteria to identify relevant ones in what will possibly be a very long list.

In our studies, we developed a new methodology for rules selection from epidemiological data, based on traditional association rule mining. In this paper, we present the resulting configuration, comprising an *Apriori* ARM implementation [44,45], two criteria to eliminate non-statistically significant and redundant rules, and a training-testing strategy necessary to mine more interesting associations while limiting noise. We then discuss the results obtained by applying the proposed methodology on a large dataset on pediatric asthma cases and pollution levels in the greater Houston area. We find the results easy to interpret on a rule-by-rule basis, and the statistical techniques for building a set of rules into a model is appropriately insulated from problems of massaging categories to fit a model or unjustifiably smoothing underlying non-linear relations. Emerging patterns in the final set of rules could be further explored with other techniques, considerably reducing the