ASSOCIATION RULE MINING FOR RISK ASSESSMENT IN EPIDEMIOLOGY

A Dissertation Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Giulia Toti

August 2016

ASSOCIATION RULE MINING FOR RISK ASSESSMENT IN EPIDEMIOLOGY

Giulia Toti

APPROVED:

Ricardo Vilalta, Chairman
Dept. of Computer Science

Edgar Gabriel
Dept. of Computer Science

Peggy Lindner
Honors College

Daniel Price
Honors College

Nikolaos Tsekos
Dept. of Computer Science

Dean, College of Natural Sciences and Mathematics

Acknowledgements

I would like to thank Dr. Peggy Lindner and Dr. Dan Price for their unconditional support, both academically and personally. Being a member of the DASH laboratory was for me like being part of a family. I would also like to acknowledge the hard work of Carol Upchurch, who helped completing the experimental part of the project.

I would like to thank my advisor, Dr. Ricardo Vilalta, for the good advices shared with me through the years, and for being not only very knowledgeable in his field, but also an exceptional and compassionate individual.

A special thanks goes to Alexander Urban for helping with the editing of this dissertation.

A final thanks goes to my family, for the sacrifices they had to make to let me come to the United States of America to pursue my career.

ASSOCIATION RULE MINING FOR RISK ASSESSMENT IN EPIDEMIOLOGY

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Giulia Toti

August 2016

Abstract

In epidemiology, a risk assessment measures the association between exposures and a health outcome. Risk characterization has traditionally been performed using statistical methods such as logistic regression, but such methods are not effective when working with highly correlated variables and when trying to assess synergic actions between exposures.

These limitations become evident in studies related to asthma, a common chronic that affects 25 million people in the US. The prevalence of asthma is growing and research is struggling to find the reason. Many factors have been associated with causing and triggering asthma, but their interactions, as well as which one is the most responsible for the spreading of asthma, are still unclear. Outdoor air pollution is on the list of possible causes and triggers. Characterizing the connection between asthma and air pollution is not an easy task, because of high collinearity between pollutant agents, possible synergic actions, and difficulty in controlling the exposure. The research community is currently encouraging the use of multi-pollutant models to yield better results.

In this dissertation we propose: (i) a modified Apriori association rule mining method for identification of connections between exposures and risk variations, and (ii) a novel genetic algorithm (GA) designed to mine risk-based quantitative association rules. Both methods were tested on a group of synthetic datasets, and on real data collection about pediatric asthma cases and pollution levels in Houston. The results on the synthetic datasets show the advantages of applying our methods to augment traditional logistic regression, and help determining the best metrics to include in the GA fitness function (*odds ratio, length, repetition* and *redundancy*).

Tests on clinical data suggest the existence of a correlation between asthma and outdoor air pollutants, both alone and as a mixture. The genetic algorithm improves the results of the Apriori-based method by recognizing what appear to be the most dangerous levels of exposure.

Future work will help to improve aspects of the GA such as population initialization or rule selection. To date, the proposed methods represent a significant step in the direction of risk assessment based on association rule mining in epidemiological studies.

Contents

1 Introduction **1**

1.1 Data mining 1

1.1.1 Association Rule Mining 3

1.2 Asthma 4

1.2.1 Symptoms and causes 4

1.2.2 Epidemiology of Asthma 5

1.2.3 Effects of air quality on health 7

1.3 Association rule mining for risk assessment 9

2 Background **11**

2.1 Association Rule Mining - State of the art 11

2.1.1 Definitions 11

2.1.2 The Apriori algorithm 14

2.1.3 Quantitative ARM 16

2.1.4 A genetic approach to QARM 21

2.2 Epidemiology and clinical studies 25

2.2.1 Measures of health outcome 27

2.2.2 Clinical studies 31

2.3 Previous studies on asthma mechanisms and the influence of pollution 36

2.4 Association Rule Mining in Clinical Studies 42

3 Algorithms **46**

3.1 Method I: Apriori-OR 46

3.2 Method II: GA-OR 50

3.2.1 GA fitness function 57

4 Methods **62**

4.1 Data 62

4.1.1 Synthetic datasets 62

4.1.2 TEDAS-TCEQ data 66

4.2 Experiments 71

4.2.1 Apriori-OR assessment of Dataset 5 71

4.2.2 Apriori-OR assessment of TEDAS-TCEQ dataset 72

4.2.3 Fine tuning of GA fitness function 74

4.2.4 GA-OR assessment of TEDAS-TCEQ dataset 76

5 Results **78**

5.1 Apriori-OR: results on Dataset 5 78

5.2 Apriori-OR: results on TEDAS-TCEQ dataset 84

5.3 Fine tuning of GA fitness function: results 88

5.4 GA-OR: results on TEDAS-TCEQ dataset 91

6 Conclusion **96**

6.1 Summary of Contributions 96

6.2 Future Work 100

Bibliography **102**

List of Figures

1.1 Asthma health care encounters per 100 persons with asthma: United States, 2001-2009. Source: CDC/NCHS, National Ambulatory Medical Care Survey, National Hospital Ambulatory Medical Care Survey, National Hospital Discharge Survey, and National Health Interview Survey 6

2.1 Graphical representation of the Apriori algorithm searching for rules of interest in the database $\mathbf{D}$, composed of 5 transactions over the set of items $I = \{A, B, C, D, E\}$. In this example, $minsupp = 2$. If the support of a node is less than $minsupp$, no further items are added to that itemset, which becomes a terminal node. In this figure, nodes with insufficient support are shaded and marked with a dashed border. The only valid frequent itemsets in $\mathbf{D}$ are $\{A, B, C, E, AC, AE, BE, CE, ACE\}$. 15

2.2 Bipartition of two numerical features (Age and Income) from a sample database, using the two more common strategies (equal-width and equal-depth). 19

2.3 Chromosome structure used to represent quantitative rules in the original implementation proposed by Mata *et al.* [52] (a), and later on by Alataş and Akin [6], who added to the encoding a new bit to allow for the search of negated rules (b). 23

2.4 Original map of Soho drawn by John Snow. The black areas indicate presence of cholera. Drawn and lithographed by Charles Cheffins. . . 26

3.1 Example of database suitable for mining using our proposed ARM method for risk assessment. All columns must be logical, indicating the presence of the exposure or health outcome. When combination of exposures are evaluated, subjects partially exposed, are not included in the computation. 48

3.2 Schematic representation of OR confidence interval of different rules. Rule $X \rightarrow Y$ is the parent. By adding other exposures to the parent rule, we obtaine the new rules $Xz \rightarrow Y$ and $Xw \rightarrow Y$. Because only the confidence interval of $Xw \rightarrow Y$ does not overlap with the parent rule, only this new association is statistically different. $Xw \rightarrow Y$ brings new relevant information, while $Xz \rightarrow Y$ should be pruned. . . 50

3.3 Flowchart illustrating every passage of the proposed genetic algorithm for mining of risk-related quantitative rules. 52

3.4 Illustration of single point crossover and uniform crossover operators. When the single point crossover is used, a crossover point is selected at random along the string and the segments before and after the point are swapped between the parents to create two new children. In the case of uniform crossover, each bit has a chance P_{cu} to come from one parent or the other. 56

3.5 Rules are penalized if they replicate rules with a higher rank. The penalty scores decrease linearly from 1 (highest ranked rule) to 0 (lowest ranked rule). In this example, Rule #3 receives a penalty of -1 for replicating the highest ranked rule. Rule #5 receives a penalty of $-1 + (-0.5) = -1.5$, because it is identical to Rule #1 and #3. . . . 60

3.6 This flowcharts explains the steps to follow to compute and adjust the fitness scores of a population of rules. 61

4.1 Representation of dataset expansion following case-crossover study design. The assumption made is that a subject who visits the emergency department on a given day did not visit again 1 or 2 weeks before and after the event. A similar approach has been used by Raun *et al.* [66] 67

4.2 Colormap of correlation between daily pollutant distributions. Darker values indicate high correlation, while light values indicate no correlation (independence). 69

4.3 Representation of distribution of sensors of the TCEQ network over the Houston area. Some sensors are located beyond the map boundaries, and they are too far from any registered patient to be of interest, therefore were not included in this map. 70

5.1 Descriptive statistics of the synthetic dataset, calculated by R using the command `summary`. This brief reports includes minimum and maximum value, median and mean, and first and third quartile for each column of the dataset. 79

5.2 Estimates of the coefficients obtained using the `glm` function on the synthetic case-control study, together with their respective standard error, z value and associated p value. Every variable except *gender* is marked with the symbol "***", indicating strong correlation between the variable and the outcome. Notice that the feature *exercise*, being categorical, had to be handled as two separate variables. 80

5.3 OR of the different exposures as calculated using logistic regression, with associated 95% confidence interval. 81

5.4 Frequency with which different pollutants at different day lags appear in the final group of 27 rules. 86

5.5 Average number of rules found at each iteration using training sets of different size. When basic Apriori search is used, thousands of associations are reported. The other lines of the chart represent the effect of adding additional filters (in sequence). When all filters are used, less than 100 rules need to be validated. 88

List of Tables

3.1 Summary of the parameters used for mining and post-pruning rules in the modified ARM algorithm, for rules of the form $X \rightarrow Y$. 51

4.1 Summary of every rule embedded in each different dataset. The tuples (E_k, T_k) indicate the exposures and the thresholds necessary to cause an impact on the odds of experiencing the health outcome. Exposures with linear impact have no definite thresholds and are marked as $(E_k, -)$. 66

4.2 Summary of the distribution of the six pollutants under analysis over the Houston area from January 1^{st} 2002 to December 31^{st} 2012. All measures are in ppb (parts per billion), with the exception of $PM_{2.5}$. . 68

4.3 Thresholds above which a subject is considered to have been exposed to a particular pollutant, compared with most recent EPA standards for 1-hour average value regulation (with the exception of $PM_{2.5}$, for which only a 24-hour average limit has been established). NO is not currently regulated [1]. 73

5.1 odds and OR for 20 year-old female patients, with median blood pressure, no diabetes and not exercising, smoking a different number of cigarettes per day, computed using equation 5.1. The OR is calculated using 0 cigarettes as reference. 81

5.2 16 rules generated by the ARM algorithm for risk assessment when used to mine Dataset 5. 83

5.3 Set of 10 rules with highest frequency across training sets. 84

5.4 Set of 10 rules with highest support across training sets. 85

5.5 Scores obtained by the different objective metrics during the iterative process used to determine which of them should be included in the fitness function. The scores represent penalties assigned when the algorithm was not able to find the rules embedded in the dataset, therefore lowest scores indicate better performance. Occasionally, the lowest score in an iteration was not selected as winner, and they are marked by †. The reasons of these choices are explained in the paragraph. Once a metric is selected, it is finalized as part of the fitness function and it does not need to receive further scores. Because the metrics *repetition* and *redundancy* are adjustment to the fitness score, they have not been tested singularly in the first iteration. 90

5.6 Set of 11 rules selected by the proposed genetic algorithm and satisfying the conditions of significant OR, $supp \geq 0.001$ and occurring in at least two final populations. 92

5.7 OR, 95% CI and p-value associated with the rule $\{\text{day1\_PM}\} \rightarrow case$ when different binning thresholds are used. The first two rows appeared in the final population produced by the genetic algorithm and show a significant risk in the odds of having an asthma attack. The last rule was obtained by binning the dataset using the resulting average threshold, and then computing the required statistics. The rule with the average threshold is no longer significant. 93

Chapter 1

Introduction

1.1 Data mining

The term "data mining" entered the popular vocabulary in the '90s, although the idea behind it was already gaining popularity in the machine learning and computer science community in the '80s. The first conference on the topic was held in 1989 under the name *KDD - Knowledge Discovery in Databases*, coined by the organizer, Gregory I. Piatetsky-Shapiro. Data mining is a well-established area of computer science and it interfaces with other disciplines such as artificial intelligence, machine learning, and statistics [16].

The following terms are used interchangeably: data mining, knowledge discovery, data archeology, information mining and information discovery. They all refer to the process of extracting interesting and novel information from large amounts of

data. This information can be extracted in different ways and serve different purposes, therefore it is common to find in the literature descriptions of data mining techniques organized by tools and purposes. Although not universally accepted, a clear classification of different techniques is proposed by [85]:

1. **Class Description:** the goal of class descriptions is the collection and portrayal of similar items in a database, with the purpose of gaining information about the class and understanding what differentiates it from other classes.

2. **Association:** association refers to the extrapolation of relationships between items in a dataset, ofter described in the form of *rules*.

3. **Classification:** during classification, a training set of labeled data is analyzed by a learning algorithm, which produces a model to separate each class. The model can be used to classify new unlabeled items.

4. **Prediction:** the goal of prediction is to assign a value to future or missing data, based on the knowledge extracted from a related database.

5. **Clustering:** by clustering it is possible to group together items with similar characteristics. It is a popular form of unsupervised learning, which means previous knowledge of labeled samples is not necessary.

6. **Time-Series Analysis:** analysis of large datasets organized in the form of a temporal sequence, in order to extrapolate sequential patterns, deviation and future trends.

In this dissertation, we will focus on the task known as *association*. This task is also often referred to as *association rule mining* or *itemset mining*.

1.1.1 Association Rule Mining

Association Rule Mining (ARM) was introduced in 1993 by Agrawal *et al.* [5]. In this paper, the authors sought interesting relationships between the items sold in a store, given a large database of customer transactions. The algorithm would present the result in the form of *rules*. For example, a possible rule could be $\{fries, hamburgers\} \rightarrow \{beer\}$, which indicates that when customers presented at checkout with fries and hamburgers, they were very likely to also purchase beer. For this particular application, the rules output by the algorithm would help the manager in organizing the items in the store, by putting products frequently sold together close to each other.

The algorithm showed strong potential and it was quickly expanded to other areas of research. Today it is possible to find a description of basic association rule mining algorithms in every introductory data mining book. The original paper from Agrawal *et al.* currently counts more than 15,000 downloads and 2,700 citations on Google Scholar.

How is a rule of interest defined? The following criteria are often used to identify interesting associations between a multitude of available item sets:

- **Support:** the support of the rule $(X \rightarrow Y)$ is a ratio between how many times the item set $\{X, Y\}$ appears in the database and the total number of

transactions. Normally, a set is required to have a minimum support before it can be considered a rule.

- **Confidence:** the confidence of the rule $(X \rightarrow Y)$ indicates the probability of item Y knowing that item X is present. In traditional theory of probability, it would be noted as $p(Y|X)$.

- **Lift:** also referred to as *interest*. For the rule $(X \rightarrow Y)$, it is the ratio of the support of $\{X, Y\}$ and the product of the supports of $\{X\}$ and $\{Y\}$ separately. If the ratio is close to 1, it indicates that the items X, Y are appearing together by mere accident and do not constitute an interesting rule.

Other criteria have been proposed in the literature. More details about rule selection and itemset mining algorithms will be discussed in the *Background* section of this dissertation (2.1).

1.2 Asthma

1.2.1 Symptoms and causes

Asthma is a largely diffused chronic respiratory disease. It is a result of a chronic inflammation of the smooth muscles surrounding the airways. The contractibility of those muscles is increased and this results in occasional narrowing of the airways. During these episodes, called acute asthma exacerbations or, more commonly, asthma attacks, patients experience symptoms such as wheezing, coughing, chest tightness

and shortness of breath.

The causes of asthma's emergence are not completely understood, but there is a broad agreement on the importance of some genetic and environmental factors, possibly in combination [55, 50]. Environmental factors believed to have a role in asthma development and exacerbation are allergens (i.e., pollen, animal hair), indoor air pollution (i.e., mold, dust), outdoor air pollution (i.e., ozone, traffic smog), and unhealthy working conditions [35, 82]. Patients can experience their first asthma attack at any age.

1.2.2 Epidemiology of Asthma

Although there is variability in how asthma is diagnosed and reported in different countries, asthma affects anywhere from 1% to 18% of the population (depending on the country) with differing levels of symptom intensity and frequency [24]. Recent estimates of the number of affected people worldwide are around 300 million. 250,000 people die every year as a consequence of an acute asthma event [82]. Developed countries are more affected than developing countries [24]. In 2007, the Center for Disease Control (CDC) published an issue of *Vital Signs* describing the impact of asthma in the United States [14]. Here we summarize some of its findings for the U.S. as a whole:

- asthma prevalence is rising. In 2001, 1 out of 14 people (about 20 million, or 7% of the population) had asthma in 2001. In 2009, the ratio went up to 1 in 12 (25 million, or 8%)

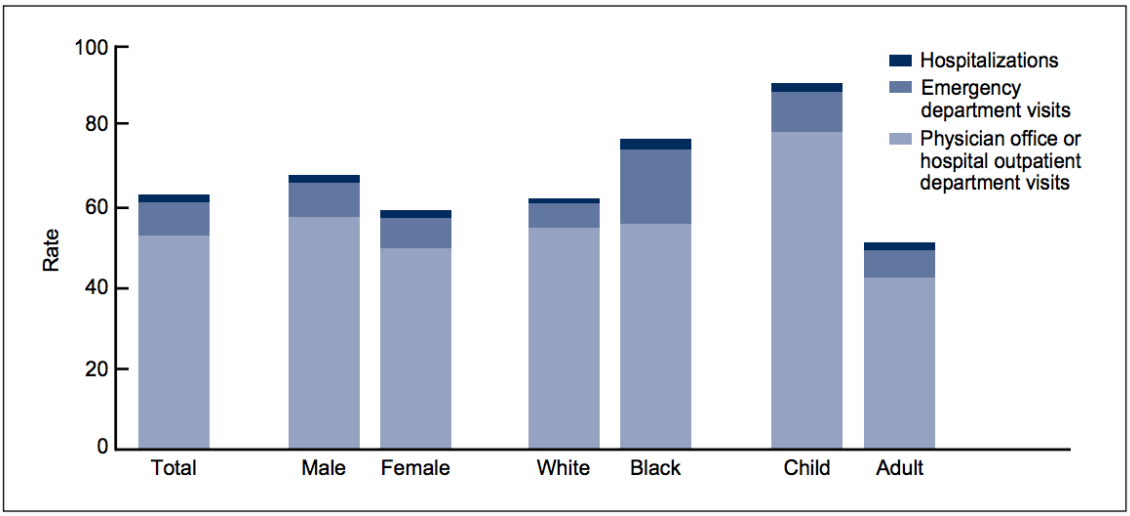

Figure 1.1: Asthma health care encounters per 100 persons with asthma: United States, 2001-2009. Source: CDC/NCHS, National Ambulatory Medical Care Survey, National Hospital Ambulatory Medical Care Survey, National Hospital Discharge Survey, and National Health Interview Survey

- 185 children and 3,262 adults died in 2007 as a consequence of an acute asthma event
- prevalence of asthma in children is higher than in adults (10% versus 8%) and they experience more acute events (57% of children with asthma had an attack in 2008 versus 51% of the total adult population)
- 50.1 billion dollars were spent in 2007 to cover the cost of medical expenses related to asthma events, with an increase of 1.5 billion compared to 2002

Although there is no cure for asthma, its symptoms can be contained with better management of the condition. Hospitals across the country focus their efforts on providing adequate care and knowledge to patients. Particular attention is given to avoiding common known triggers. Among the known triggers cited above, this

dissertation will focus on acute asthma events caused by exposure to outdoor air pollution. Many research groups across the globe are investigating the possible connection between outdoor pollution and asthma prevalence [66, 88]. In the *Background* section of this dissertation (2.3), we will illustrate in greater detail how pollution can affect patients with asthma, what studies validate this hypothesis, and the current difficulties encountered in extracting and using this information.

1.2.3 Effects of air quality on health

The simple act of breathing exposes people to contact with different chemicals. Some of them are known to be harmful (such as carbon monoxide) others are thought to be harmless, and others have not yet been assessed. Evaluating the effects of chemicals present in the atmosphere is a challenging work in progress. In 1996, the Environmental Protection Agency (EPA) launched a project called National-Scale Air Toxic Assessment (NATA). The goal of the project is to characterize the effects of all the 187 chemicals listed in the Clean Air Act. To date, 134 of the 177 analyzed chemicals have shown a cancer and/or non-cancer dose-response value. For more information, refer to [3].

A famous example of known air pollution hazard is its proven correlation with cardiovascular diseases, such as heart attack and stroke. In 2004, the American Hearth Association (AHA) released a statement about the effects of exposure to fine particular matter (diameter smaller than 2.5 μm, or $PM_{2.5}$) [12]. They declared that short-term exposure to the pollutant (a few hours to a few weeks) was linked to an

increased risk of triggering cardiovascular disease-related events, fatal and nonfatal. Long term exposure (a few years) produced an even higher risk and resulted in a shorter life expectancy. A few years of reduced exposure to $PM_{2.5}$ would result in a decreased risk for cardiovascular mortality. It was possible to reach these conclusions only after several studies had been conducted across the globe for many years before the AHA publication. Afterwards, researchers started focusing on other pollutants with similar suspected behaviors, such as ozone.

The assessment of potential harmfulness of chemicals is not an easy task. First of all, it would be unethical to conduct studies in which people are exposed to substances suspected to be harmful. Therefore, only observational studies are eligible for this task. The control of the exposure during the study poses another problem. To make a valid comparison, detailed information between subject exposures during the study are needed, but not always easy to obtain because of subject mobility and the scarcity of pollution sensors. And lastly, pollutants are always present in the atmosphere as mixtures, which makes it difficult to evaluate the effects of each pollutant separately. The possibility for some pollutants to have a synergic action (harmful in combination but not separately) should also be accounted for.

As we mentioned in section 1.2.1, it is reasonable to suspect a correlation between asthma prevalence and exacerbation and the presence of certain pollutants in the air we breathe. More details about previous studies on asthma and air pollution will be discussed in the *Background* section of this dissertation (2.3).

1.3 Association rule mining for risk assessment

So far, we have introduced the concept of data mining with particular attention to mining for association rules, and we have discussed the impact of asthma and air pollution on our society. These two apparently unrelated topics come together to form the core of this dissertation: association rule mining for risk assessment.

In this document, we present an improved method for association rule mining dedicated to the assessment of risk in clinical trials, and a novel implementation of a Genetic Algorithm (GA) for risk assessment. Association rule mining has several characteristics that make it an interesting choice for the study of risk in clinical trials:

- as with every data mining algorithm, ARM does not rely on the hypothesis formulation and testing as it happens in traditional statistic methods; the knowledge is extracted directly from the data
- rules are readily understandable even for people outside the data mining domain
- there is a vast literature on the topic which creates a solid foundation (more details in Section 2.1 and 2.4)

The novel implementation of the association rule mining method could help in understanding the effects of air pollutants on asthma incidence. The concept could be extended to assess risk variation in other clinical studies. The Genetic Algorithm for risk assessment is a further improvement upon this method, because it allows the

researcher to handle features with continuous values and it automatically assesses the threshold of exposure associated with the most significant risk change.

This dissertation is organized as follows: in the next chapter (*Background*) we will expand upon the different implementation of rule mining algorithms, their limitations and current research trends (2.1). Then we will discuss introductory concepts in public health and clinical studies (2.2) and review previous studies on the effects of pollution on asthma (2.3). The *Background* section concludes with an overview of previous application of ARM mining in the clinical domain. The following chapter (*Algorithms*) describes the implementation of the proposed methods for risk assessment using association rule mining techniques. In the *Method* section we explain the data and the experiments used to validate the proposed methods. Results and discussions of these experiments are reported in the following chapter, *Results*. Finally, in the last chapter, we propose a summary of the contributions of this work and the plans for future improvements.

Chapter 2

Background

2.1 Association Rule Mining - State of the art

In Section 1.1 we introduced the meaning of data mining and its purposes, with particular emphasis on Association Rule Mining (ARM). In this section, we will discuss in greater detail the implementation, strengths and current limitations of ARM.

2.1.1 Definitions

Let $I = \{i_1, i_2, ..., i_m\}$ be a generic set of items. A subset of items $X \subseteq I$ is called an *itemset*. An itemset is normally associated with its cardinality, or size, k. Let $T = \{t_1, t_2, ..., t_n\}$ be the set of transaction identifiers or *tids*. Then, it is possible to define tuples of the form $\langle t, X \rangle$ called *transactions*. Often, the transaction $\langle t, X \rangle$ is

referred to using its identifier t.

In order to mine information from a database, transactions are represented through a binary matrix $\mathbf{D} \subseteq T \times I$. The columns of $\mathbf{D}$ represent the set T of all possible items. The rows of $\mathbf{D}$ represent all the transaction identifiers available, normally sorted in lexicographical order. For example, if $\mathbf{D}$ is the database representing the sales of a grocery store, the columns of $\mathbf{D}$ would represent all the items available for sale (bread, milk, eggs...), and its rows would represent the different customers. $\mathbf{D}_{i,j}$ is true (1) if the customer i purchased the item j, and false (0) otherwise.

Association Rule Mining is the process through which meaningful associations between items are extracted from $\mathbf{D}$. For example, in the grocery store scenario, it would be interesting to know that a customer who purchases bread will oftentimes buy milk as well. Rules are expressed through the notation

$$X \rightarrow Y \tag{2.1}$$

where X and Y are itemsets of I of cardinality $1 \leq k \leq m$ and $X \cap Y = \emptyset$.

As we mentioned in Section 1.1.1, not every possible combination of itemsets (X, Y) forms an interesting rule. Different criteria have been defined to differentiate meaningful rules from the rest. The most common, introduced by Agrawal in his first ARM formulation [5], form the **support-confidence framework**. The support-confidence framework requires that selected rules satisfy at least two criteria: minimum support and minimum confidence.

The **support** of an itemset X represents how often the itemset appears in the

database, or how many transactions in $\mathbf{D}$ contain X:

$$supp(X) = |\{t \quad | \quad \langle t, \mathbf{i}(t)\rangle \in \mathbf{D} \quad \text{and} \quad X \subseteq \mathbf{i}(t)\}| \tag{2.2}$$

An ARM algorithm looks first for *frequent* itemsets to form rules of interest. A rule $X \rightarrow Y$ is a good candidate if $supp(X \cup Y) = supp(XY) \geq minsup$, where $minsup$ is a user defined threshold.

After determining whether a rule has sufficient support, the algorithm calculates its **confidence**. The confidence of a rule $X \rightarrow Y$ measures the chance of finding the itemset Y in a transaction, knowing that the itemset X is in the transaction. Therefore, the confidence is nothing but a conditional probability:

$$conf(X \rightarrow Y) = P(Y|X) = \frac{P(X \wedge Y)}{P(X)} = \frac{supp(XY)}{supp(X)} \tag{2.3}$$

If a rule has sufficient support ($minsupp$) and sufficient confidence ($minconf$), it is an interesting rule in the support-confidence framework.

In 1991, G. Piatetsky-Shapiro [63] argued that minimum support and confidence are not enough to ensure that a rule is meaningful. According to probability theory, two independent events X and Y can happen simultaneously with a chance $p(X)p(Y)$. Therefore, if

$$p(X \wedge Y) \approx p(X)p(Y) \tag{2.4}$$

then

$$supp(X \rightarrow Y) \approx supp(X) \times supp(Y) \tag{2.5}$$

If Equation 2.5 is verified, then it is not true that X implies Y, but rather that the two itemsets are happening together by chance. A new metric is needed to filter and remove this kind of rules from the set of meaningful rules. The concept of **lift** (also called *interest*) provides a possible solution:

$$lift(X, Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)} \tag{2.6}$$

If the ratio computed through Equation 2.6 is too close to 1, the rule is not interesting. A threshold *minlif* can be introduced to ensure that the interest of the rule is sufficiently distant from 1:

$$\left|\frac{supp(X \cup Y)}{supp(X)supp(Y)} - 1\right| \geq minlift \tag{2.7}$$

Other metrics have been introduced to measure the importance of rules in a database, but they are not within the scope of this dissertation.

2.1.2 The Apriori algorithm

Knowing what kind of rules should be extracted from the binary database of transactions, it is possible to formulate an algorithm to search for them automatically. One obvious solution would be to consider every possible itemset in I and individually compute their support and confidence. Because it needs to evaluate all the $2^{|I|}$ itemsets in I, this algorithm (called brute-force) is effective but highly inefficient.

The support metric has an interesting property that allows for the exclusion of some of the itemsets from the rule search without having to test their support. In

Database **D**	
t	i(t)
1	ACE
2	AD
3	ACE
4	BCE
5	BE

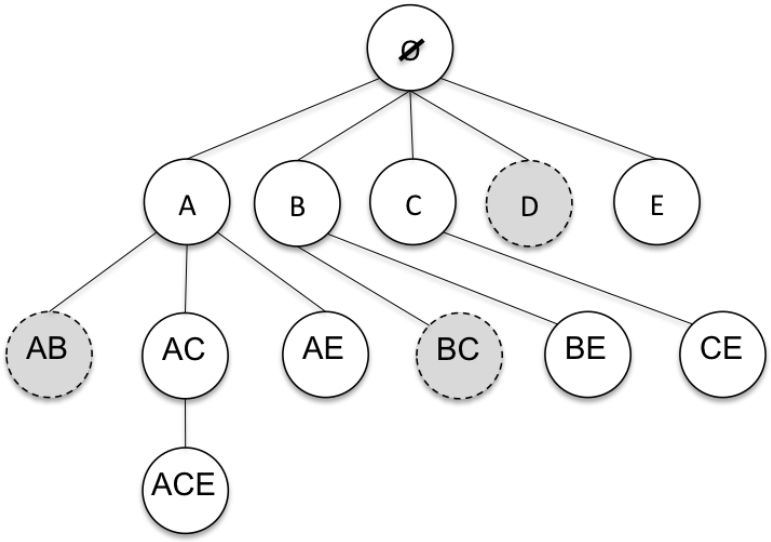

Figure 2.1: Graphical representation of the Apriori algorithm searching for rules of interest in the database **D**, composed of 5 transactions over the set of items $I = \{A, B, C, D, E\}$. In this example, $minsupp = 2$. If the support of a node is less than *minsupp*, no further items are added to that itemset, which becomes a terminal node. In this figure, nodes with insufficient support are shaded and marked with a dashed border. The only valid frequent itemsets in **D** are $\{A, B, C, E, AC, AE, BE, CE, ACE\}$.

fact, if $X, Y \subseteq I$ and $X \subseteq Y$, then the support of Y can not be greater than the support of X ($supp(X) \geq supp(Y)$). This means that if an itemset X is not frequent, than every superset of X can not be frequent and can be excluded from the rule search. We say that the support has a **down-closure property**. Notice that the same property is not true for confidence. If $X \rightarrow Y$ and $W \rightarrow Y$ are rules of **D** and $X \subset W$, the confidence of $W \rightarrow Y$ could be greater than that of $X \rightarrow Y$.

The support down-closure property is the foundation of a popular algorithm for association rule mining, known as the Apriori algorithm. The Apriori algorithm was proposed in 1993 by Agrawal *et al.* [5]. The algorithm performs a tree-like search

starting from the base leaf (empty itemset) and progressively adding new items to create new nodes. For each node, the support of the itemset is computed. If it is greater than *minsup*, the search can continue and new items are added to create the next nodes. If the support of a node is too small, supersets on that branch are not evaluated and the search moves to other branches. This process is easier to understand by looking at a graphic representation (Figure 2.1).

Many other algorithms for ARM of binary databases are available in the literature, but they will not be discussed in this dissertation.

2.1.3 Quantitative ARM

Association rule mining techniques have a relatively short history and are currently under development. Some of the most interesting features under study include:

- **Beyond the support-confidence framework.** In some fields, researchers might be interested in rules that do not appear often in the database. Think, for example, about a medical study on association between certain symptoms and a rare disease. The combination of symptoms of interest will likely be infrequent across the entire patient database, making it at greater risk of being overlooked when searching itemsets with a high support. The data mining community is working on algorithms that are not bounded by high frequency, but are still computationally efficient. This is particularly challenging because metrics with down-closure property are hard to find. An example of mining for weak patterns was proposed by Liu *et al.*, who called these rules "reliable

exceptions" [47].

- **Negative association rules.** Negative associations can also provide valuable insights. For example, searching the grocery store database, one could find that people who buy milk almost never buy beer. The frequency and the confidence of this pattern make it interesting, but basic ARM algorithms, such as Apriori, do not look through the negated form of the database. Methodologies for negative association rule mining have been proposed in [47, 69, 85].

- **From correlation to causality.** Associations discovered through rule mining are useful to understand correlations, but not causality. For example, the rule $X \rightarrow Y$ indicates that Y has a high probability to happen when X is present, but this does not mean that X is the *cause* of Y. Further information is needed to make this claim. The topic of causality in ARM is extensively discussed by Zhang and Zhang in the fourth and fifth chapter of their book, "Association Rule Mining" [85].

- **Quantitative ARM (QARM).** A major limitation of basic algorithms for rule mining such as Apriori is the inability to handle non-boolean features. If the user is interested in mining frequent rules from continuous data, she will have to group them in intervals before running the algorithm. For example, a feature such as the income of a person will have to be divided in appropriate bins (i.e., low, average and high income) to be able to mine rules. This problem often results in errors and loss of information. Researchers have tried to overcome this difficulty by developing algorithms for *quantitative association rule*

mining - algorithms and methods capable of handling continuous variables. A good overview of the issue and of different proposed solutions was given by Adhikary and Roy in [21].

The genetic algorithm for risk assessment proposed later in this dissertation was developed to allow for the inclusion of numerical features in the analysis. But why is this capability so important? Why not let the user decide on the appropriate binning to use?

Some scenarios allow for a degree of user discretion when selecting thresholds for the grouping of numerical features. This can be appropriate when a threshold has been established and accepted, such as the level that indicates high cholesterol or blood pressure. In different situations, arbitrarily deciding upon a threshold can result in information loss. Important associations risk appearing weaker or not at all as a result of features not being grouped correctly.

Different binning strategies have been proposed to avoid this issue. Srikant and Agrawal [73] originally proposed a partitioning approach, where the quantitative variable is first divided in intervals, then a boolean variable is used to map each sample to the appropriate interval. This method resembles the use of dummy variables in logistic regression. The authors were aware that this strategy creates a dilemma known as the *min-support problem*, which can be summarized as follows: if many intervals are used, the support of single intervals may be so low that not enough frequent rules are generated. If few intervals are used, they may lack granularity and result in significant information loss. Furthermore, different partition strategies can

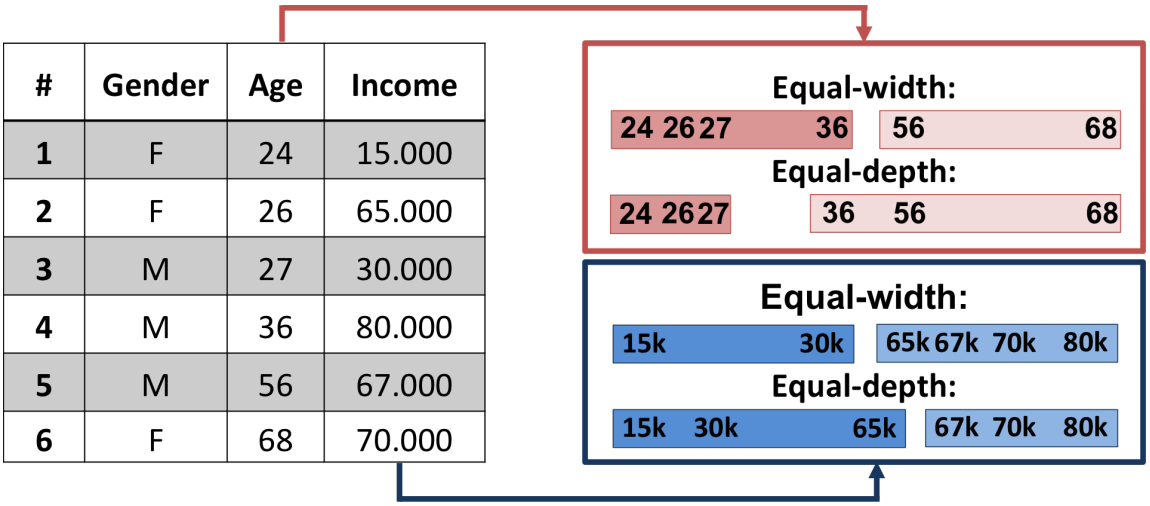

Figure 2.2: Bipartition of two numerical features (Age and Income) from a sample database, using the two more common strategies (equal-width and equal-depth).

be used, with different results and problematics. The most common are *equal-width*, in which the intervals span over ranges of the same size, and *equal-depth*, in which the intervals contain the same number of samples. Using an equal-depth partition strategy ensures that all intervals include the same number of samples, but this may result in less meaningful bins. For an example, observe Figure 2.2: the division of income into intervals based on width resulted in the value 65k being included in the first group, despite being more similar to the higher values in the second group. On the other hand, if an equal width strategy is used, we may end up with empty or almost empty intervals, which will be less likely to produce frequent rules.

In later years, a clustering approach was proposed to produce meaningful intervals from the data at hand. The clustering approach uses a density-based strategy that is somewhat similar to the equal-depth partitioning, but with the capability of scaling to higher dimensions thus producing intervals of interest based on more than one

feature. A partition based on density increases the likelihood of finding frequent and applicable rules, while filtering out infrequent scenarios. Problematics of this approach include the necessity of having a user defined number of clusters, and the difficulty of clustering skewed data. QARM techniques that implement a clustering approach can be found in [19, 45, 54, 78, 84].

Other QARM techniques utilize information extracted from the data distribution in an attempt to generate the best possible intervals, relying on traditional statistical measures such as mean or standard deviation. The weaknesses of this approach seem to outweigh its strengths: the partition is often limited to single features or bipartition. Furthermore, the computational cost of these methods can be significant, and they may produce uninteresting rules if a measure of deviation from the mean is not used. Implementations of this techniques are described in [8, 34].

Occasionally, it may not be convenient or even necessary to define precise binning thresholds. Qualitative terms such as *high blood pressure* or *new customer* may be more appropriate depending on the analysis or discipline. For this reason, a "fuzzy" approach to QARM was introduced [86]. In fuzzy QARM, it is no longer necessary to bin continuous values using sharp thresholds. In particular, [87] proposes a method to automatically optimize the fuzzy sets and their partition points based on the original quantitative data. Downsides of this method include high computational costs and, on occasion, difficulty in defining appropriate fuzzy intervals.

Another major technique for quantitative association rule mining employs an evolutionary approach, particularly with the use of genetic algorithms. Genetic Algorithms (GA) are a family of biologically-inspired methods dedicated to solving

problems of various natures through search and optimization. Using GA to solve the rule mining task brings several advantages: first, a population of rules is generated and progressively optimized instead of being searched for through a sequence of database scans, thus being computationally more advantageous. Second, it is no longer necessary to define parameters such as support and confidence (although the specification of other parameters to guide the genetic search may be necessary). Furthermore, because GA allow for multi-objective optimization, rule selection can be based on several different criteria at once, making it possible and convenient to define appropriate selection criteria for the problem under analysis. For these reasons, we believe Genetic Algorithms to be the most promising approach to solve our problem rule mining for risk assessment.

The following section offers a more detailed explanation of Genetic Algorithms and their use in QARM. More information about itemset mining in clinical application can be found toward the end of this chapter (Section 2.4).

2.1.4 A genetic approach to QARM

Genetic algorithms are biology-inspired heuristic methods designed for search and optimization purposes. They were first proposed in 1975 by John H. Holland [31]. This first implementation, called Simple Genetic Algorithm (SGA), begins with the generation of a population of binary strings (*chromosomes*) that encode possible solutions to the problem under study. The population is evaluated through a *fitness function* - the function to be optimized. In the next step, the population is used to

generate new strings. Each new string is generated as follows: first, two strings from the original population are selected. The selection is random but proportional to the fitness of the string, so that better strings generate more offspring. The selected chromosomes are subjected to single-point *crossover*, where portions of the two strings before a selected cut point are swapped, creating two new chromosomes. Finally, the two new chromosomes go through a process of *mutation*, where random bits (*genes*) of each string are changed from 0 to 1 or vice versa. The probability of mutation is normally low, but it helps to preserve variety in the population and avoid local minima. When enough new strings have been generated, they join the population of chromosomes. The entire population, comprised of parents and children, is evaluated using the fitness function. Only the best half is preserved and survives to the next iteration, where it is used to generate new offspring. The process is repeated until the population converges toward a solution or the maximum number of generations is reached. An excellent introduction to genetic algorithms is offered by Srinivas and Patnaik in [74]. A scheme of the Simple Genetic Algorithm, also from [74], is visible in the box Algorithm 1.

```
Simple Genetic Algorithm ();
initialize population;
evaluate population;
while termination condition = false do
    select solutions for next population;
    perform crossover and mutation;
    evaluate population;
end
```

Algorithm 1: Simple Genetic Algorithm structure, as illustrated in [74]

After this first implementation was proposed, many improvements to GA were

a)

Gene 1			Gene 2				Gene n		
A_1	L_1	U_1	A_2	L_2	U_2	...	A_n	L_n	U_n

b)

Gene 1				Gene 2					Gene n			
A_1	N_1	L_1	U_1	A_2	N_2	L_2	U_2	...	A_n	N_n	L_n	U_n

Figure 2.3: Chromosome structure used to represent quantitative rules in the original implementation proposed by Mata *et al.* [52] (a), and later on by Alataş and Akin [6], who added to the encoding a new bit to allow for the search of negated rules (b).

researched and published, including better population initialization, better crossover methods, different selection strategies, and creation of new fitness functions to find solutions to a vast range of problems. Covering this literature is beyond the scope of this dissertation. We will instead focus on how GA integrate with quantitative ARM.

To our knowledge, the first evolutionary approach to QARM was proposed in two consecutive papers by Mata *et al.* [51, 52]. The authors first introduced the idea of encoding itemsets in the form of chromosomes that could be manipulated by an evolutionary algorithm. The structure of each chromosome is visible in Figure 2.3 (a). In this representation, A_k is the index of one of the n attributes of the database under analysis. The itemset size can span from 2 to n (the itemset including all attributes). L_k and U_k are the upper and lower bounds of the interval covered in this itemset (naturally, they must be within the range of A_k).

The fitness function used to evaluate the quality of an itemset was the following:

$$f(i) = covered - (marked * \omega) - (amplitude * \psi) + (nAtr * \mu) \tag{2.8}$$

Covered is equivalent to the support of the itemset. But this measure alone is not sufficient to identify frequent itemsets when quantitative variables are involved. If no other measure is used to control the intervals, the algorithm would set the lower and upper bounds of the numerical variables to cover their whole range: this would produce the highest values for support, because all instances can now be included in the set. Other metrics are needed to ensure that the chosen intervals are meaningful - for instance, the authors decided to use *amplitude* to penalize large intervals. The fitness function also includes *nAtr*, which promotes itemsets with a larger number of attributes, and *marked*, which penalizes records already covered by other itemsets in the population (without it, the whole population would converge toward the best overall itemset, ignoring other relevant discoveries). ω, ψ and μ are weights that the user can use to regulate the impact of the different metrics.

After this first implementation, many steps were taken to improve the quality of itemsets and rules mined using genetic algorithms. We list some noteworthy examples: Yan *et al.* [83], who used GA to eliminate the need of a user specified minimum support; Qodmanan *et al.* [64], who also eliminated minimum confidence; [6] and [49] introduced the possibility to mine for rules with negated attributes. Other papers have validated the effectiveness of these methods [7].

Building on the foundations provided by this vast literature, we are now able to propose a Genetic Algorithm for QARM designed to find those rules that result in the largest changes in risk. This method employs some of the existing features for rules generation and selection while offering a newly formulated fitness function that rewards desirable rules characteristics, such as being associated with relevant risk

variations. The details of this algorithm are explained in Chapter 3.

2.2 Epidemiology and clinical studies

Epidemiology is the founding science of public health. Epidemiological studies analyze the distribution of a health outcome in a population (as opposed to medicine and biology, which target individuals) in relation to different risk factors.

Before epidemiology was born, people did not have a scientific explanation for sickness and the outburst of epidemics. Large parts of the population believed that diseases were the result of miasmas, "bad air" or witchcraft, although various literates and scientists through history have been seeking a more logical explanation [38]. John Snow, an English physician of the 19^{th} century, is considered the father of modern epidemiology [80]. Snow is known for identifying the source of a cholera outbreak that struck the London neighborhood of Soho in August 1854. Although Snow could not have known what was causing the cholera (it would be another 7 years before Louis Pasteur's discovery of germ theory), through his investigation he was able to connect the infected patients to a water pump on Broad Street. During his study, he produced a detailed map of cholera cases in the area and showed how the pump was at the center of the affected territory (Figure 2.4). The pump was shut down, contributing to the end of the epidemic [33].

What is most remarkable about this story is how Snow was able to come to his (correct) conclusion through extended investigation and use of statistics, even though he could not physically prove the danger of the water through chemical analysis.

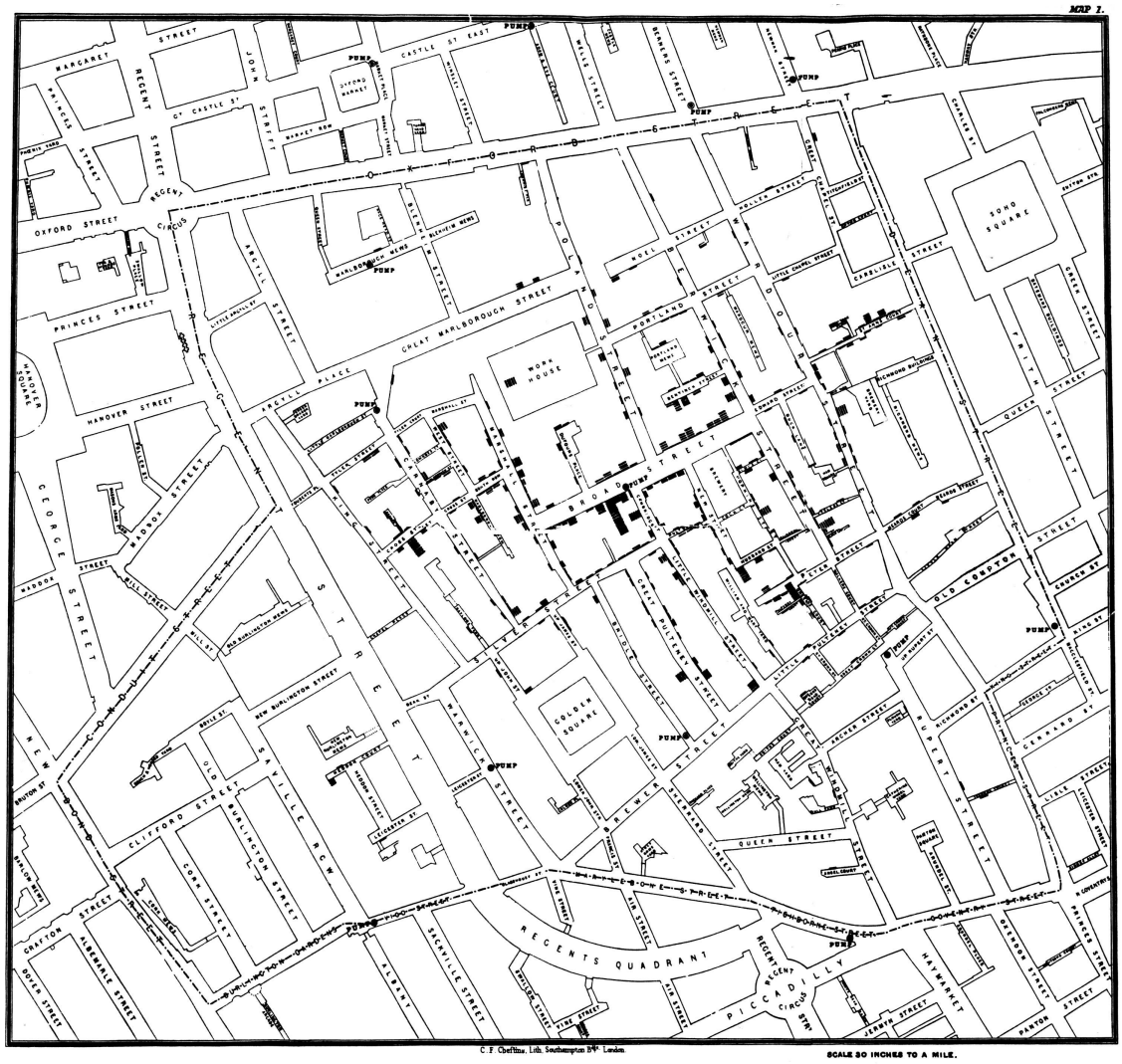

Figure 2.4: Original map of Soho drawn by John Snow. The black areas indicate presence of cholera. Drawn and lithographed by Charles Cheffins.

Progress in science has since given us new tools for the study of public health. Today, epidemiology examines on a macro scale the relationships between people's health and their environment, diet, lifestyle, and genetic makeup/constitution.

2.2.1 Measures of health outcome

In this section we define the parameters most frequently used in epidemiology to measure health outcomes and associations with risk factors.

Here, a list of the parameters used to quantify health outcomes is presented:

- **Prevalence:** proportion of people presenting the outcome of interest. For example, the prevalence of asthma in the United States is 8% of the total country population (25 million out of around 300 million). When talking about prevalence, it is important that the denominator (total population) is clearly defined.

- **Risk:** ratio between number of people presenting the outcome and number of total people at risk. For example, consider a hypothetical case of food poisoning at a restaurant. Through investigation it is found that, out of 100 people who ate at the restaurant, 50 had chicken (population at risk) and 40 experienced food poisoning (population with outcome). The risk of having food poisoning when eating chicken at this restaurant is 40/50, or 80%. Notice that risk always applies at the population level, and not at the level of the individual.

- **Rate:** outcomes of interest per person-time. For example, consider a study

on migraines in which 200 subjects are followed for 3 years. During the study period, 450 episodes of migraines are reported. Therefore, the rate for migraines in this study is 450/(200*3) = 0.75 migraine episodes per person per year. Rate, as opposed to risk, is not a proportion. It is useful as it adds the temporal information.

- **Odds:** ratio of the chances of two mutually exclusive outcomes. If p is the probability of having a certain outcome, then

$$odds = \frac{p}{(1-p)} \tag{2.9}$$

  For example, if the probability of having a male child is 0.51, or 51%, the odds of expecting a boy are 1.04.

It is also important to differentiate *incident* from *prevalent* cases. Incident cases are new cases. The status of these individuals changed from *without outcome* to *with outcome* over the defined period of time. For example, saying "50,000 people contract HIV every year in the US" is a measure of incidence. Prevalent cases are the total number of affected individuals in the defined population. The point in time when the outcome occurred is not considered. "1,144,500 people in the US have been diagnosed with HIV by the end of 2010" is an estimate of prevalence [15].

Often it is interesting to compare these measures among different populations, particularly when the two populations differ for some exposure. To do so, the parameters are combined in ratios. The most frequently used are the **risk ratio (RR)** (also called relative risk) and the **odds ratio (OR)**. In the fictional food poisoning study above, we estimated that the risk of food poisoning after eating chicken was

0.8. Assume that out of the 50 people who did not eat chicken, 5 had food poisoning. The risk in this case is 0.1. The risk ratio between the two populations is $0.8/0.1 = 8$. In other words, people who ate chicken were 8 times more likely to feel sick than people who did not eat chicken.

The odds ratio is computed in a similar way, although the meaning is slightly different. It measures the association between an exposure and an outcome by comparing the *odds* of having the outcome with and without the exposure. But an odds ratio of 8 does not necessarily mean that an exposed individual is 8 times more likely to have the outcome. The concept of odds ratio is harder to grasp for non-statisticians, and it is often the cause of confusion in the medical community. Odds ratios are normally greater than risk ratios, which may lead researchers into publishing the most impressive figure without understanding its meaning. In 2001, a study conducted over 151 papers estimated that 26% of them misinterpreted odds ratio as risk ratio [30].

The value of odds ratios and risk ratios can be interpreted in the following way:

- greater than 1: there is a positive correlation between the exposure and the outcome of interest
- equal to 1: exposure and outcome are not correlated
- less than 1: there is a negative correlation between the exposure and the outcome of interest

In clinical papers, measures of association are often presented with their confidence interval, a range that serves as the lower and upper bound for the parameter. The confidence interval is affected by certain aspects of the study methodology, such as sample size and population variability. Confidence intervals can change in level of confidence, the most frequently used being the 95% confidence level. In this case, if the parameter were calculated multiple times using different samples, 95% of the interval estimates would be expected to include the population parameter. Different applications may call for larger or smaller confidence levels (i.e., 50% or 99%).

The confidence interval helps determine whether the result of the study is statistically significant. We said before that we talk about increased risk if the risk ratio is greater than 1. What happens if the confidence interval encompasses 1?

In this scenario, no definitive claims can be made about the effect of the exposure on the outcome of interest. The result is *not statistically significant*. The same goes for the odds ratio. Odds ratios and risk ratios are statistically significant only when their confidence interval does not include 1. If the study has a non statistically significant result, researchers can try to change part of the setup to improve upon it (i.e., select a larger population sample). Note that in the field of epidemiology results that are not statistically significant can still have some clinical importance.

2.2.2 Clinical studies

Different studies have been designed by epidemiologists to understand and quantify the connection between particular factors and a physiological outcome. In this section, we will offer an overview of some basic concepts and of the most frequently used study designs.

The first distinction that can be made between studies is whether they are *experimental* or *observational*. Experimental studies imply that researchers have discretion over which group of subjects will be exposed to a particular variable. One of the more common experimental study methods - the **randomized control trial** - is often used to study the effects of new drugs. Here, a sample of subjects with a particular condition is selected before being divided into two groups: one of them treated with the drug under study, and the other treated with either another drug or a placebo. This allows for the assessment of differences between the treatments. There are many additional precautions that must be taken in experimental studies to remove bias. In a *blind study*, patients are left in the dark as to what drug they are receiving to prevent any bias in their responses. In a *double-blind study*, the researchers are also unaware of which patients received which drugs to further remove the possibility of bias in treatment. Researchers must also try to account for and remove any **confounders**, which are extraneous variables that might correlate with both the dependent and independent variables. An example of a confounder within an experiment would be if the control group and experimental group were not properly randomized (along gender, age, or any number of other variables) to make sure that the only significant difference between them is the treatment being tested.

Yet another strategy, known as a **clinical crossover trial**, involves the switching of treatment assigned to the groups, to further validate the effects of the exposure.

Experimental studies can be conducted only under the principle of *equipoise*, which is the genuine uncertainty about the benefits or harms of the exposure. It would be unethical to expose subjects to a substance suspected to be toxic, or to treat patients with a drug that appears to be less effective than another. When the principle of equipoise does not hold, one must opt for an observational control study, in which the exposure is not imposed by the researcher, but is a pre-existing condition of the subject (i.e., smokers versus non-smokers, or people living in a metropolis versus the countryside).

One of the most commonly used observational studies is called **cohort study**. In a cohort study, researchers select two groups of subjects based on the exposure. Then the two groups are followed for a certain period of time, during which the occurrence of the outcome of interest is observed. Consider for example a cohort study to evaluate the correlation between lung cancer and smoking. First, eligible subjects would be divided into two groups based on whether they smoke or not. Then the subjects would be observed for a follow-up period and the occurrence of lung cancer cases in the two groups would be recorded. At the end of the follow-up period, researchers can determine if there is a statistically significant difference between the number of lung cancer cases in the two groups and make claims on the effect of the exposure on the outcome, such as "smoking increases the risk of lung cancer by a factor x".

In this example, the subjects have been involved in the research before the occurrence of the outcome. This format is called *perspective* cohort study. Alternatively, researchers can opt for a *retrospective* cohort study, which traces back in time events that have already taken place. In the lung cancer example, we could select subjects that have been smoking for a certain number of years and people who did not smoke in the same period, and then evaluate the presence of lung cancer in the two groups. Retrospective cohort studies are normally quicker and less costly than prospective cohort studies, but it can be more difficult to control for confounders. Also, there is more uncertainty on the real exposure in the two groups, because it is based on the memory and accuracy of the subjects.

Generally, cohort studies are a great choice to estimate risk variation due to some exposure. The possible disadvantages are cost, difficulty in following up with the subjects for the necessary time, and increased difficulty in avoiding confounders.

Another popular observational study is the **case-control study** design. In a case-control study, the researcher identifies a group of individuals who manifest the outcome of interest (*cases*) and another group in whom the event did not occur (*controls*). Then, he or she will study the history of exposure in the two groups and look for significant differences. Since the selection of the subjects is done after the outcome of interest occurred, case-control studies are always retrospective [37]. The case-control study design offers several advantages: it fits the study of rare conditions, it allows for the simultaneous study of numerous potential causes, and because the outcome has already occurred it is relatively fast to conduct. Once the condition to study has been defined, the selection of cases is straight-forward. Controls are

selected to match some of the cases characteristics (i.e. age, sex, occupation...) and therefore are at similar risk to developing the condition. Case-control studies can be matched, if the ratio cases-controls is 1:1, or unmatched. A researcher interested in studying the connection between smoke and lung cancer through a case-control study would select a group of individuals with lung cancer (cases), a group of healthy subjects with similar characteristics and then evaluate the incidence of smoking among the two groups.

After the data has been collected, it is possible to summarize them in a *contingency table*, as seen below:

	Exposed	Not Exposed
Cases	a	b
Controls	c	d

Where:

- a is the number of subjects who present the outcome and have been exposed to the cause under study (i.e., smokers with lung cancer)
- b is the number of subjects who present the outcome but have not been exposed (i.e., non-smokers with lung cancer)
- c is the number of subjects who do not present the outcome of interest but have been exposed (i.e., healthy smokers)
- d is the number of subjects who do not present the outcome of interest and have not been exposed (i.e., healthy non-smokers)

From the contingency table it is possible to compute the odds ratio between the exposed and unexposed group through the formula:

$$\text{OR} = \frac{ad}{bc} \tag{2.10}$$

Note that it is not possible to compute a risk ratio in a case-control study, because the sample of subjects is not selected to be representative of the real distribution of the disease in the population. Although under certain conditions the odds ratio value is very close to the risk ratio value for the whole population [67].

Another variant for observational studies is the **cross-sectional study** design. It is often the least expensive and time-consuming option available for researchers interested in epidemiological studies. In this type of study, all the data are collected at a defined time. Often, it is not the researcher interested in the disease assessment to perform the collection, but they use data collected for other purposes or during routine operations (i.e., data collected from patients visits to the emergency room). The use of routine data allows to reach large portions of the population at minimal cost. This makes for another advantage in comparison with case-control studied: cross-sectional studies allow to estimate the risk ratio. On the downside, using routine data means that the information was not collected to answer the specific question. This can leave out important material, needed for example to better handle confounders. Another issue with cross-sectional studies is known as ecological fallacy. This happens when the data is not available at individual level, which is very common in censuses provided by large institutions. Working with information related to groups instead of individuals can lead to error and imprecisions in the results, such

as incorrect mean estimates, or masking of significant correlations.

The last kind of observational study we are covering in this dissertation is the **case-crossover study** design [48]. This design is suited for assessing the effects of transient exposures. The study is self-matched, which means that each subject acts as his/her own control. Differences in the outcome of interest are estimated between periods when the subject was exposed versus periods when he/she was not. An example will help understanding this design: consider a study on the risk of cellphone use while driving. The study compares the outcome (accident) across periods when the subject was driving while using the phone versus other time windows where he or she was driving without using the phone. The study is conducted on different subjects, and then a global conclusion is proposed. The case-crossover study design is particularly effective in eliminating confounders. On the downside, it is highly dependent on how the control and exposure windows are defined.

2.3 Previous studies on asthma mechanisms and the influence of pollution

In Section 1.2.1 we introduced basic concepts of asthma pathophysiology and talked about the impact of the disease on the population. In this section, we will provide more details on current asthma studies and related problematics. Then we will focus our attention on studies related to the effects of pollution on asthma. We will start from studies conducted using statistical methods largely used and accepted by the

epidemiology community (i.e., logistic regression). Then we will present more recent approaches based on data mining and machine learning techniques.

Different studies on asthma epidemiology can focus on different aspects of the disease's impact on the population. To understand this literature overview, the reader should know the difference between an acute asthma event and chronic asthma, and be familiar with the concepts of incidence and prevalence. For more information on these definitions, refer to Section 1.2.1 for asthma symptoms and Section 2.2.1 for health outcome measures.

Because of its impact and increasing prevalence, asthma is the target of numerous research studies worldwide. Questions range from drugs effectiveness and possible improvements, leading causes, to future trends in the disease prevalence. The large number of studies produced all over the globe may have contributed to one of the key problematics of asthma assessment: the inconsistent definition and diagnosis of the disease.

In 2014, a study by Sá-Sousa *et al.* [68] showed that the different definition of asthma used by various studies can significantly change the prevalence estimates. Regarding U.S. studies, the paper identified 7 different definitions of the disease, resulting in prevalence estimates ranging from 1.1% to 17.2%. This inconsistency can generate some confusion while studying the literature, and even be an obstacle in the improvement of care provided by physicians. In the Houston area, the Texas Emergency Department Asthma Surveillance (TEDAS) is an example of how the medical community is trying to respond to the problem. This network was established to collect data related to emergency department (ED) visits for pediatric

asthma. It is based at the Texas Children's Hospital, with the collaboration of three partners (Lyndon B. Johnson General Hospital, Ben Taub General Hospital and the University of Texas Medical Branch in Galveston). The primary goal of TEDAS is to collect better information on patients visiting ED, in order to improve strategies for chronic asthma management. The TEDAS database includes information related to more than 20,000 ED visits from 01/01/02, reaching roughly 11,000 children in the Houston-Galveston area. The network managers declared that, by using the collected data to improve physicians' training, they improved the consistency between diagnosis and even reduced the number of emergency room visits [4]. The content of this database has been used in the study on asthma and pollution discussed in this dissertation.

The first step in understanding what causes asthma is possibly determining when most of the acute asthma events happen. Studies conducted in different areas of the globe (i.e., Finland [29], New Zeland [36], Singapore [18]) looked for peaks in annual asthma incidence trends. Studies conducted in Finland and New Zealand reported a peak for hospitalization among younger age groups (less than 15 years) in the early winter months, coincident with the return to school. The Finnish group also reports a peak in May for very young patients (0-4 years), coincident with the allergy season. On the other end, Chew *et al.* could not find statistically significant correlation between asthma incidence and season. Another study by Han *et al.* [28] employed a cross-sectional design to find correlations between pediatric asthma and season and determine the probable cause. Of the 1725 Taiwanese children with asthma involved in the study, 187 were found to have perennial asthma, 590 had more attacks in

winter, 629 in spring, and 255 in summer-fall. The author suggests some possible triggers for the different seasons, for example cockroaches in summer-fall and mold in winter and spring. Year-round contributing factors included younger age, parental smoking during pregnancy, air pollution. The findings of this paper suggest that different subjects are triggered by different exposures. This kind of study offers clues to identify suspected asthma triggers, but they have not been conclusive. Differences such as climate, floral population, and pollution sources make asthma incidence a highly local issue.

The next papers that will be described focus on the impact of one particular trigger: ambient air pollution. We have already talked about known negative effects of ambient air pollution, particularly $PM_{2.5}$, on cardiovascular diseases [12]. In 2013, Sram *et al.* dedicated a study on impact of air pollution on children. They found positive correlations between exposure to pollutants (PM_{10} and c-PAHs, an aromatic hydrocarbon) during gestation and in early childhood and negative health outcomes, particularly bronchitis and asthma [72].

An excellent overview on the asthma and outdoor pollution problematic is offered by Guarnieri and Balmes in [25]. A summary of the most important contributions of this paper is the following:

- Specific pollutants have been proved to cause airway inflammation and hyper-responsiveness, and oxidative stress, which are characteristic of asthma. The list of suspects comprises ozone, nitrogen dioxide and particulate matter of various size.

- Short-term exposure to these pollutants is likely to increase the risk of asthma exacerbation.
- Asthma exacerbation could be reduced by implementing local pollution warning systems.
- Since pollutants are always present in the atmosphere as a mixture, the risk contribution of single pollutants is unknown.
- Results of epidemiological studies are affected by imprecise assessment of the exposure and of the diagnosis of asthma itself.
- Gene-environment interaction seems to have a primarily important role in the onset of new asthma cases, but its mechanisms are unclear.
- Risk modifiers to consider when studying asthma epidemiology include gender, diet, ethnicity and socio-economical status.

Another study from Patel *et al.*, conducted on Dominican and African American children of 0-5 years of age, found an increased odds ratio for asthma and wheezing symptoms when the subjects were living near stationary sources of air pollution, such as highways [62]. In Texas, notable associations between asthma onset and the pollutants listed above was found in the El Paso area [88], in the Harris County [81], and more specifically in the Houston area [66].

Although pointing in the right direction and validating air pollution as probable cause of asthma increasing prevalence and exacerbation, the studies presented so far have limitations, especially when it comes to modeling and explaining the effects of

pollutants as a mixture. In Europe, the Scientific Committee on Health and Environmental Risks is already encouraging the introduction of multi-pollutant models to improve the study of indoor air pollution effects [70] and the idea is spreading to other areas of research on air quality. Exhaustive reviews on recent progresses in the area of multi-pollutant modeling are already available [11, 32, 75]. The paper by Billionnet *et al.* is particularly effective in delineating the limitation of traditional statistical methods, such as the impossibility to consider synergic effects of pollutants. We talk about synergic action when two or more exposures have a greater combined effect than the sum of each effect individually. Furthermore, pollutants often present high collinearity (high variables dependency), which affects the results of logistic regression models. When the variables included in a logistic regression model are largely collinear, it is impossible to estimate each individual regression coefficient with confidence, and therefore it is impossible to determine the real impact of the individual predictors.

The methods cited by Billionnet *et al.* include Bayesian approaches, CART (Classification and Regression Trees), K-means, PCA (Principal Component Analysis), Logic Regression, and more. A good example of CART for exposure characterization is offered by Gass *et al.* [23]. This paper proposes a modified regression tree to evaluate the effects of four pollutants (CO, NO_2, O_3 and $PM_{2.5}$) on the number of emergency department visits related to asthma in children in Atlanta, Georgia. The method incorporates a division of pollutants in quartiles, for simplification and easier understanding. The lowest quartile is held as reference for the computation of the risk ratio. The conclusions presented suggest that not a single pollutant is

the cause for increase asthma onset, but it is the result of generally higher levels of pollution. Although the recursive partitioning implemented by CART is valid in assessing hierarchy and non-linearity in the pollutant effects, it may fail in detecting synergic actions.

[60] presents a Bayesian approach for the assessment of multiple-source health effects. The method is tested on a database of $PM_{2.5}$ levels in Phoenix, Arizona, and cardiovascular mortality (but not asthma). The authors declare that their results are validated by previous studies. Bayesian approaches are a good choice when dealing with measurement error and collinearity, but they normally rely on some *a priori* knowledge which is not always available.

The knowledge acquired through machine learning and data mining techniques is not only useful in understanding exposures effects, but it can also be used in the context of a warning system to predict the insurgence of unfavorable air conditions and avoid patients exposure to triggers. Lee *et al.* made an attempt in this direction, presenting two data mining methods, one based on decision trees, one on association rules, aimed at predicting asthma attacks [40]. Predictors include bio-signals from patients and environmental data. Results are promising, although the need to include biological data in the feature set poses an obstacle to the method implementation.

2.4 Association Rule Mining in Clinical Studies

Data mining in general, and association rule mining in particular, have been attracting the interest of the medical domain in the recent years. Every year, medical

applications of machine learning and data mining are presented at major conferences such as KDD or ICDM. Even dedicated journals made their appearance (i.e., Artificial Intelligence in Medicine). In 1999, N. Lavrač published an overview on data mining in medicine [39], founded on the argument that the exponential growth of medical data was making impossible the manual analysis done so far. Since new tools for medical data collection and storage were available, new tools for their analysis should be used as well. The author groups data mining techniques in three categories: pattern-recognition methods (i.e., k-nearest neighbor), artificial neural networks, and inductive learning of rules, which encompasses association rule mining (although ARM is not directly mentioned in the paper, in favor of older algorithms for decision rules production such as CN2 [20] and ID3 [65]). In 2008, Bellazzi *et al.* proposed their guidelines for data mining in clinical medicine [9]. Again, ARM is not cited explicitly, but a large portion of the paper is dedicated to decision trees and decision rules.

Association rule mining for clinical application has a relatively short but rich history. The first application we know of was the study of Brossette *et al.* on association between hospital infections and public health surveillance [13]. Other publications include studies on chronic hepatitis, septic shock, heart disease, association deficit disorder, cancer prevention, response to drugs and general lifestyle risk behaviors [57, 59, 58, 41, 76, 56, 17, 61].

Using association rule mining to find patterns in medical data has obvious advantages, the most prominent being the readily interpretability of the resulting rules,

even from people outside the data mining domain, However, traditional ARM algorithms face different issues when applied to clinical databases. First of all, the support-confidence framework poses a serious limitation: often, interesting patterns related to diseases are not frequent, and they would not be selected as interesting rules. Reducing (or eliminating) the minimum support brings two additional challenges: increase of computational cost, and selection of most interesting rules. Lowering the minimum support and confidence means to inevitably output a higher number of rules, sometimes too many to screen them manually. A solution is required to select the most relevant rules. Some algorithms have been proposed to avoid mining redundant rules and mining a limited number of rules. However, this solutions are still not optimal. Mostly, they are affected by a data-coverage problem, resulting in all the selected rules being related to a particular portion of the database, while the remaining portions are ignored.

In this dissertation, we propose: (i) a new method that uses a combination of Apriori and rule post-processing to mine interesting risk-based rules from clinical databases; and (ii), a new genetic algorithm for mining of risk-based quantitative rules. The first method has the advantage of having an acceptable computational cost and being able to sort the output rules by a parameter that is relevant to medical practitioner. Previous attempts in this direction have been done by Li *et al.* [43]. This group has the merit of proving that risk patterns have a down-closing property that can be used to guide the rule-searching process. The second method improves upon the first by eliminating the support-confidence framework and the need to discretize quantitative variables. We believe that this algorithm could be

particularly helpful in the legislative process, because its capability of automatically determining the most dangerous exposure thresholds.

In the following chapters we will present the details of both of the risk-based association rule mining algorithms, how they can be applied to improve our knowledge on air pollution-asthma relationship, and what are the advantages of these methods in comparison with traditional Statistics.

Chapter 3

Algorithms

3.1 Method I: Apriori-OR

In this chapter we will provide the details of two methods that use association rule mining techniques for risk assessment in epidemiological studies. Both our methods are designed to estimate odds ratio from data collected in case-control studies.

Our first design brings together the Apriori implementation of ARM and a group of post-processing criteria to filter valid and interesting rules. To be suitable for rule mining, a dataset must include a column indicating wether the subject is a case or a control, and a series of columns, one for each exposure suspected of being related to the health outcome (Figure 3.1). All columns must be in logical form, indicating the presence or absence of the exposure or health outcome. Let us call $E = \{\exp_1, ..., \exp_n\}$ the set of all exposures in the dataset, and C the particular set

of size 1 including only the item *case*. Given $X \subseteq E$, we can easily compute OR and 95% CI of the rule $X \rightarrow C$. The odds of developing the health outcome for the exposed population are:

$$odds(X \rightarrow C) = \frac{p(C, X)}{p(\neg C, X)} = \frac{supp(CX)}{supp(\neg CX)} \tag{3.1}$$

Where $\neg C$ and $\neg X$ represent the group of controls (not cases) and not exposed subjects, respectively. When the impact of multiple exposures at the same time is evaluated ($|X| > 1$), only subjects who have not been exposed to any feature in X are included in the non-exposed population (See figure 3.1). This strategy was chosen because it offers a more clear definition of exposed and non-exposed group. A paper by Toti *et al.* provides more information on the effects of using different definitions of non-exposed populations in ARM [79].

The odds ratio can then be written as:

$$\mathrm{OR}(X \rightarrow C) = \frac{odds(X \rightarrow C)}{odds(\neg X \rightarrow C)} = \frac{supp(CX)/supp(\neg CX)}{supp(C\neg X)/supp(\neg C\neg X)} = \frac{supp(CX)supp(\neg C\neg X)}{supp(C\neg X)supp(\neg CX)} \tag{3.2}$$

Finally, we can compute standard error (SE) and 95% confidence interval:

$$\ln(\mathrm{SE}) = \sqrt{\frac{1}{supp(CX)} + \frac{1}{supp(C\neg X)} + \frac{1}{supp(\neg CX)} + \frac{1}{supp(\neg C\neg X)}} \tag{3.3}$$

$$\ln(\mathrm{CI}_{95\%}) = [\ln(\mathrm{OR}) - 1.96\ln(\mathrm{SE}), \ln(\mathrm{OR}) + 1.96\ln(\mathrm{SE})] \tag{3.4}$$

$$\mathrm{CI}_{95\%} = e^{\ln(\mathrm{CI}_{95\%})} \tag{3.5}$$

Using ARM for an epidemiological study is equivalent to evaluating a total of 2^{m} contingency tables such as the one visible in Figure 3.1, where m is the number

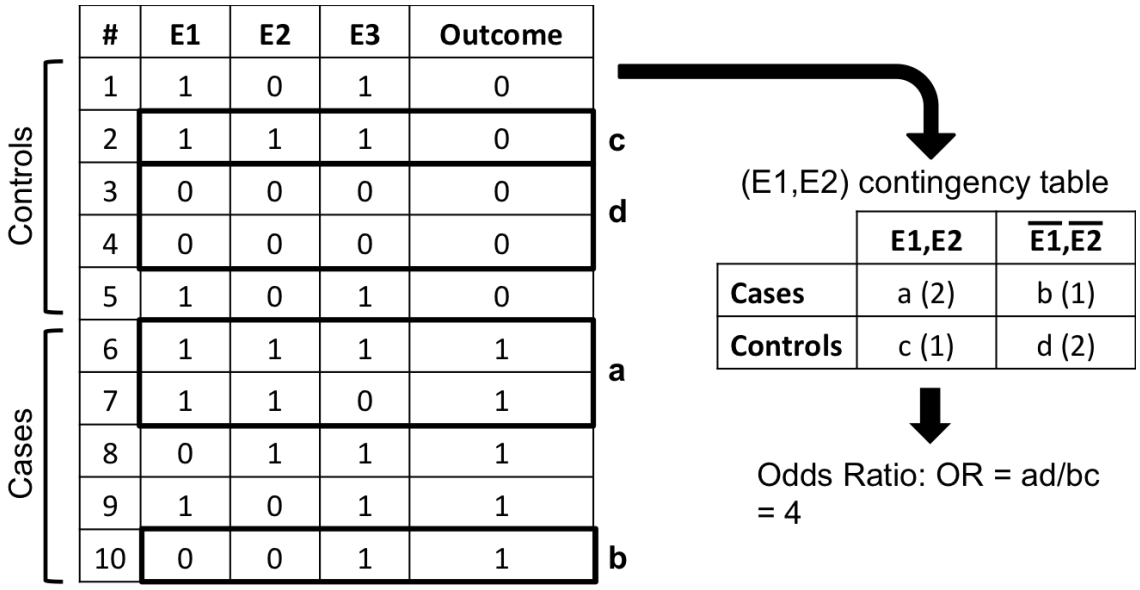

Figure 3.1: Example of database suitable for mining using our proposed ARM method for risk assessment. All columns must be logical, indicating the presence of the exposure or health outcome. When combination of exposures are evaluated, subjects partially exposed, are not included in the computation.

of exposures under analysis. This is the first advantage in using ARM for studies of this type: the algorithm produces a complete analysis of all possible associations in the available database. Naturally, the number of contingency tables increases exponentially with the number of risk factors. Even a small number of exposures (e.g., 20), can produce an incredibly high number of contingency tables (1,048,576, in our example). Obviously, it would be impossible to evaluate by hand the significance of every contingency table, so the evaluation needs to be automated.

Associations are first mined using relatively low values for minimum support (slightly above 0%), because for this kind of study it is not of primary importance for an association to appear in the database a high number of times, but rather we are interested in associations that produce significant changes in the odds of having

the health outcome under investigation. A minimum confidence is not required.

Because mining with very low support normally results in a very high number of associations, further steps are necessary to filter the unimportant ones. The post-processing criteria included in this method are:

- A rule is pruned if its 95% CI for the OR crosses the value of 1 (with some tolerance, if desired), because in this case the effect of the exposure(s) on the health outcome is either irrelevant or ambiguous.

- A rule is pruned if it is redundant, that is, another simpler rule exists that carries analogous information. In our case, we require no overlapping of the 95% CI of the rule with all of its parents, as proposed by [44]. A representation of how the criterion categorizes redundant rules is visible in Figure 3.2.

- If the p-value of the rule is $\geq 5\%$ the rule is considered non-statistically significant and pruned.

- We also control each rule for lift. This parameter helps identify casual associations. If the lift of a rule is close to 1, the association is happening with the same incidence as a random choice model and should not be interpreted as a true interaction. We imposed for all rules lift ≤ 0.95 or lift ≥ 1.05.

The procedure has been implemented using RStudio (R version 3.1.2 "Pumpkin Helmet"). The database is first mined using the Apriori function from the package *arules*, available on the CRAN website [27, 26]. Ad-hoc functions have been designed to compute odds ratio [77], non-redundancy based on [44], and chi-squared statistics

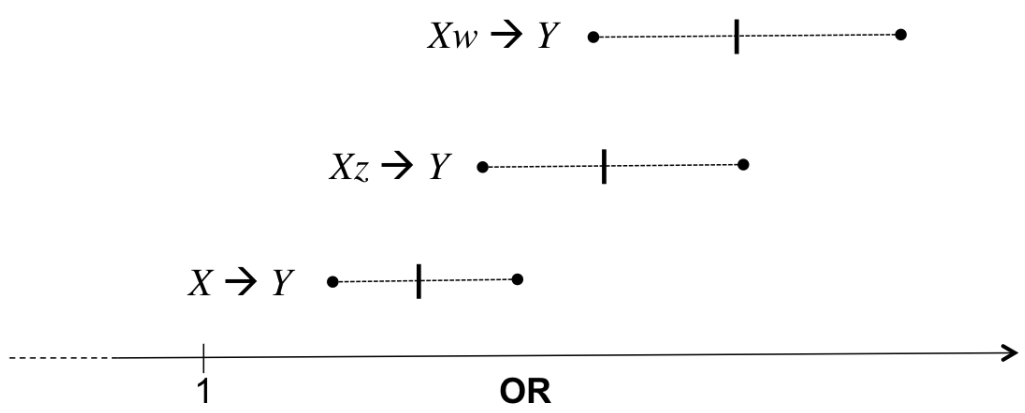

Figure 3.2: Schematic representation of OR confidence interval of different rules. Rule $X \rightarrow Y$ is the parent. By adding other exposures to the parent rule, we obtaine the new rules $Xz \rightarrow Y$ and $Xw \rightarrow Y$. Because only the confidence interval of $Xw \rightarrow Y$ does not overlap with the parent rule, only this new association is statistically different. $Xw \rightarrow Y$ brings new relevant information, while $Xz \rightarrow Y$ should be pruned.

[46]. Table 3.1 presents a summary of the equations associated with each criterion used for mining and post-pruning.

3.2 Method II: GA-OR

The method described in the previous section is a first step in the direction of extraction of OR-based rules from epidemiological data, but it is afflicted by one major inconvenience: all the columns in the database to be mined must be logical. In Section 2.1.3 we have extensively discussed why this can result in poor approximation and information loss. Often, exposures included in the kind of study we are targeting are numerical (i.e., age, blood pressure, quantity of chemicals/drugs) and their discretization is far from obvious. Actually, finding that critical threshold could be the

Table 3.1: Summary of the parameters used for mining and post-pruning rules in the modified ARM algorithm, for rules of the form $X \rightarrow Y$.

Parameter	**Equation**
Support ($supp$)	$P(X \wedge C) = supp(X \wedge C)$
Confidence ($conf$)	$P(X\|C) = \frac{supp(X \wedge C)}{supp(X)}$
Lift ($lift$)	$\frac{supp(X \wedge C)}{supp(X) \cdot supp(C)}$
Odds Ratio (OR)	$\frac{supp(CX)supp(\neg C \neg X)}{supp(C \neg X)supp(\neg CX)}$
Chi-squared measure (χ^2)	$\sum_{event} \frac{(supp(event) - E(supp(event)))^2}{E(supp(event))}$ where $event = X \rightarrow C, \neg X \rightarrow \neg C, \neg X \rightarrow C, X \rightarrow \neg C$
Redundancy	95% $\text{CI}_{X \rightarrow C} \cap$ 95% $\text{CI}_{X \wedge x \rightarrow C} = \emptyset$ where $x \in E$ and $x \cap X = \emptyset$

goal of the study: at what age do we become more susceptible to developing a particular condition? In what quantity does a certain chemical start showing significant impact on the human body?

The literature on quantitative association rule mining (QARM, Section 2.1.3) is extensive, but no algorithm has been designed, to our knowledge, to tackle the problem of mining epidemiological databases to provide an overview of risk changes associated with various exposure, while automatically estimating the most critical exposure threshold for continuous variables. The genetic approach to QARM, however, offers a possible starting point. GA are extremely flexible and able to optimize multiple metrics simultaneously. Such a powerful optimization tool will be able to perform the required task, as long as it is correctly formulated as an optimization problem. In this case, the optimal solution is the set of rules associated with the

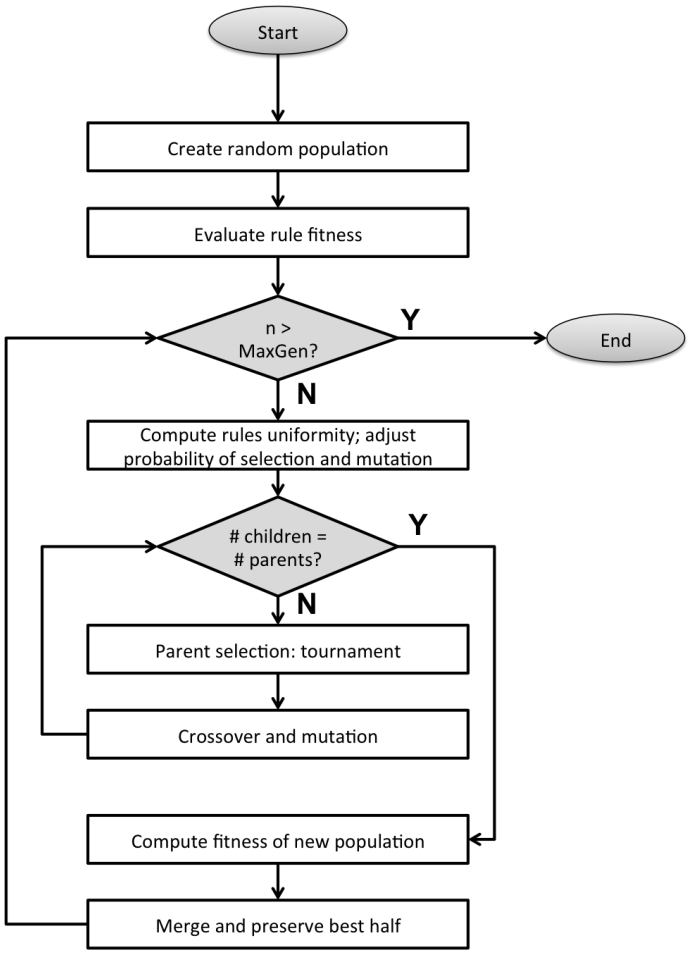

Figure 3.3: Flowchart illustrating every passage of the proposed genetic algorithm for mining of risk-related quantitative rules.

most relevant changes in odds of manifesting a health outcome.

Genetic algorithms include multiple operations and functions (crossover, selection, mutation...), but the core of the method is the *fitness function*. One could think of the fitness function as the question that the GA is trying to answer through heuristic search. Naturally, if the question is poorly formulated, no acceptable answer can be found.

Figure 3.3 illustrates all the steps included in the proposed algorithm. It begins with the random initialization of a population of chromosomes. These chromosomes encode different rules of the form $X \rightarrow case$, where X is a set of one or more exposures. To be manipulated by the GA, the rules where organized in the following string structure:

E_1	T_2	E_2	T_1	...	...	E_N	T_N

Each tuple $\{E, T\}$ represents an exposure and the threshold above which subjects should be considered exposed. Logical features can still be used, and their threshold is set to *true* by default. Every E_k is a logical variable and if $E_k = true$ the exposure is included in the rule. T_k is a continuous variable that can assume any value in within the range of E_k. For example, given a dataset including 4 exposures, to encode the rule $\{E_1, E_2, E_4\} \rightarrow case$, with 5.0, 2.5 and 6.5 as threshold above which a subject should be considered exposed, we would use the array:

E_1	T_1	E_2	T_2	E_3	T_3	E_4	T_4
1	5.0	1	2.5	0	7.8	1	6.5

The user can exert some control on the initial population by specifying desired population size (normally between 30 and 100) and the probability of each exposure to be included in a rule (P_{in}). For example, if the database includes 10 exposures and $P_{in} = 0.5$, we can expect the average rule length of the initial population to be around 5. P_{in} should be used to control the initial rule length, because too-large exposure combinations are probably irrelevant and very infrequent, but very short rules could

also be not relevant and reduce the population variety, making the discovery of more interesting rules more difficult. Initial threshold are selected uniformly at random within each exposure range, with the exception of logical exposure, that if included are always set to *true*. After this first population is created, the algorithm assesses the fitness of each rule. Because of its importance, the details of the fitness function will be discussed extensively in the following section.

The algorithm then proceeds to estimate the uniformity of the population, using the following equation:

$$Unif = \frac{|\sum_{k=1}^{N}[(2\sum_{i=1}^{pop\_size} E_{i,k}) - pop\_size]|}{pop\_size} \tag{3.6}$$

Where N is the number of exposures and pop_size is the number of rules in the population. This equation returns a value between 0 and 1, where 0 indicates total dissimilarity, and 1 indicates that the population includes only one rule repeated pop_size times. The uniformity of the population is used to adjust the probability of selection (P_t) and mutation (P_m), within the range specified by the user. If the population is becoming more uniform, P_t and P_m increase. As a consequence, less fit rules will have a higher probability to be selected for reproduction and more genes in the children population will mutate, re-introducing some variability in the gene pool and preventing GA from getting stuck around non-optimal solutions. P_t and P_m are adjusted within user specified ranges using a linear equation:

$$P = P_{LOW} - (P_{HIGH} - P_{LOW}) \cdot Unif \tag{3.7}$$

Suggested ranges are [0.9, 0.6] for P_t and [0.001, 0.01] for P_m.

The generation of the new population can now begin. It is an iterative process repeated pop_size times, because, in this implementation, the children population size should match the one of their parents. At each iteration, two parents are randomly selected using a tournament selection strategy. In a k-ary tournament selection, k strings are selected uniformly at random from the population to compete against each other. The string with the highest fitness has higher chances to win, although less fit strings can also emerge as winners. Allowing weaker strings to be selected preserves genetic variety and helps avoiding local minima. The best rule has a probability P_t to win the tournament. The second best has a probability to win equal to $P_t(1-P_t)$. The third can win with a probability of $P_t(1-P_t)^2$, and so on. The user can decide how many strings participate in the tournament. The two parents are selected in two separate tournaments.

Once the two parents are selected, they can mix their genes through *crossover* operation with a probability P_c. We have mentioned how the idea of a single-point crossover operator was already introduced in Holland's original paper [31]. Many improvements have been proposed since then. For our algorithm, we chose to use a *uniform* crossover operator. When a uniform crossover operator is used, each gene is swapped between the parents with a probability P_{cu}. The advantage of this strategy over swapping segments of genes is that each gene can be exchanged with the same probability, while single-point crossover is biased by the gene's position. The differences between single-point and uniform crossover are illustrated in Figure 3.4. There is also a probability $1-P_c$ that the two parents do not mix their genes, in which case the two children are copies of their parents. Note that, in the proposed

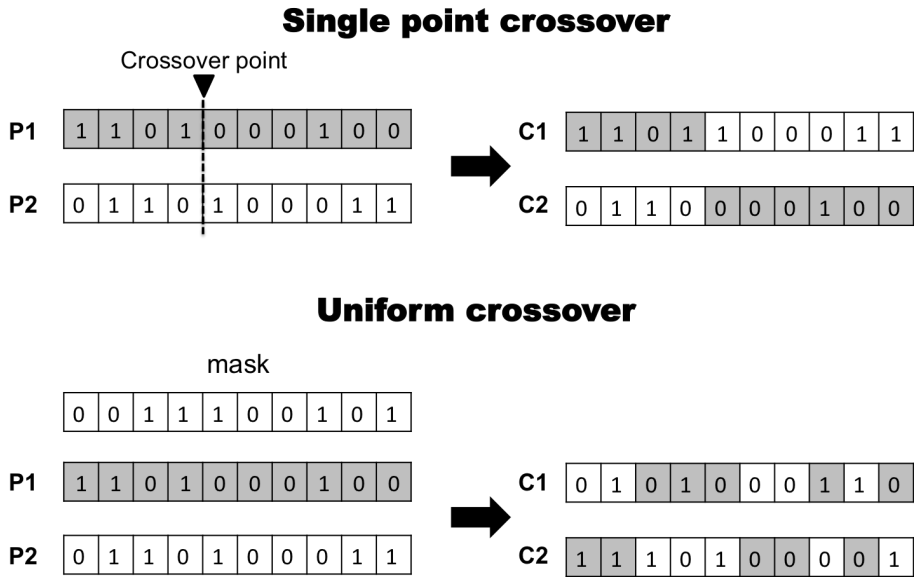

Figure 3.4: Illustration of single point crossover and uniform crossover operators. When the single point crossover is used, a crossover point is selected at random along the string and the segments before and after the point are swapped between the parents to create two new children. In the case of uniform crossover, each bit has a chance P_{cu} to come from one parent or the other.

implementation, the crossover operator acts only on the binary genes E_k, which define the presence or absence of an exposure in a rule. When two genes $E_{k,p1}$ and $E_{k,p2}$ are swapped, their associated thresholds also change to become the arithmetic average of the two parents ($T_{k,c1} = T_{k,c2} = (T_{k,p1} + T_{k,p2})/2$).

The chromosomes generated through crossover will then undergo under a process of *mutation*, during which anyone of their binary genes E_k can be swapped from 1 to 0 or vice versa with a probability P_m. The mutation rate is normally low, but it helps exploring more solutions and avoiding getting stuck in local minima. Imagine, for example, that all solutions in the population are converging toward strings where the first gene is 1. By using crossover only, we lose the chance to explore strings

having 0 as first gene. The mutation operator preserves this possibility. If a gene is swapped, its associated threshold also mutates, using the equation:

$$T_{k,new} = T_k \pm 10 \cdot P_m \cdot T_{1/2} \tag{3.8}$$

Where $T_{1/2}$ is the median value of the distribution of the exposure E_k. Whether the threshold is reduced or increased ($\pm$) is decided randomly by the mutation operator.

After mutation, the two children are added to the new population (and kept separate from the parent pool). The operation is repeated until the size of the new population is equal to *pop_size*. The fitness of each children is then computed and the two populations are merged to form an overall rank. The *pop_size* best chromosomes are selected and survive to become parents in the next iteration. The process is repeated until the number of generations reaches the limit set by the user. The *pop_size* chromosomes surviving the last iteration are presented as resulting mined rules.

3.2.1 GA fitness function

The fitness function proposed in this dissertation is what allows the algorithm to find rules with the required characteristics, such as relevant odds ratio and sufficient frequency. In order to design the optimal fitness function, we started by thinking about what qualities a rule should have to be rewarded, and what characteristics should instead be penalized. Each of these objective functions is a possible candidate to be included in the final fitness function. By formulating this as a multi-objective

optimization problem, we prevent the algorithm from focusing only on one desirable characteristic (i.e. significant OR), and we encourage it to find balanced solutions among multiple criteria. We imposed for each objective function to be limited to the range [0,1], so that the weight of each metric would be comparable to the others. The final list of candidates includes:

- **Odds ratio fitness:** this metric is of course the first on the list, because the algorithm is designed to look for rules that produce interesting changes in odds ratio between exposed and unexposed population. The contribution to this metric to the fitness function is

$$OR_{fitness} = \begin{cases} 0, & \text{if } 1 \subseteq 95\% \text{ CI} \\ OR - 1, & \text{if } OR < 1 \\ 2\left[\frac{1}{1+e^{1-OR}} - 1/2\right], & \text{otherwise} \end{cases}$$

  This formulation assigns 0 fitness points if the 95% CI of the rule includes 1, because this class of rules is ultimately empty of interesting associations between exposures and outcome. If 1 is not included in the confidence interval, the rule gains a higher fitness score the further the OR is from 1. The proposed formulation used to compute the score when OR $>$ 1 limits the maximum possible score to 1, for a fair comparison of positive and negative associations.

- **Support:** the support of the rule is added and contributes to its fitness score, because we wish to reward frequent over unfrequent associations.

- **Confidence:** the traditional formulation of confidence of a rule is also added to the global fitness score, to reward stronger associations.

- **Length:** for better understandability, we favor shorter rules over longer ones. Each rule receives therefore a penalty depending on its length. The penalty is normalized by the largest possible rule size ($length_{penalty} = length/N$, where N is the total number of exposures).

- **Extremity:** this metric is meant to prevent the threshold of a continuous exposure to get too far from the median value. It was designed in response to other metrics that have a tendency to bring the threshold toward the lower bound (support) or the upper bound (OR, in some cases). It is computed as

$$Extremity = \frac{\sum_{k=1}^{N} |T_k - T_{1/2}|}{max(T_{1/2} - T_{LOW}, T_{HIGH} - T_{1/2})} \Big/ N \tag{3.9}$$

We have also considered the necessity to avoid repeated and redundant rules in the population. The following two metrics have been proposed to penalize repetition and redundancy. They have not been included in the list of fitness function candidates because, unlike the others, these metrics can not be computed on the basis of the rule alone, but they depend on a comparison between the rule under evaluation and the other rules in the population. For this reason, they are proposed as fitness *adjustments* and are computed after all the rules have been scored and sorted based on the other metrics.

- **Repetition:** rules that are identical to other rules in the population receive a penalty based on the ranking of the rule they are replicating. Being identical

Rank	E_1	E_2	E_3	E_4	E_5
1	1	0	0	1	0
2	0	0	0	1	1
3	1	0	0	1	0
4	0	1	1	0	0
5	1	0	0	1	0

Penalty score	Received
1.00	0.00
0.75	0.00
0.50	-1.00
0.25	0.00
0.00	-1.50

Figure 3.5: Rules are penalized if they replicate rules with a higher rank. The penalty scores decrease linearly from 1 (highest ranked rule) to 0 (lowest ranked rule). In this example, Rule #3 receives a penalty of -1 for replicating the highest ranked rule. Rule #5 receives a penalty of $-1 + (-0.5) = -1.5$, because it is identical to Rule #1 and #3.

to the highest ranked rule brings a penalty of -1 to the fitness. The penalty scores reduce linearly with the rank until the last ranked rule, which carries no penalty. Two rules are identical if they include the same exposures ($X_1 = X_2$), independently of the thresholds used. If a rule is identical to two higher ranked rules, it receives both penalties, therefore the total penalty score can be higher than 1. This penalty enforces variety in the gene pool and prevents the whole population from converging toward one solution. For an example of how penalties are calculated and assigned, see Figure 3.5.

- **Redundancy:** rules also receive a penalty if their 95% CI overlaps with that of an existing parent in the population. Given a rule $X \rightarrow case$, a rule $Y \rightarrow case$ is said to be a parent if $Y \subset X$. If $Y = X$ the two rules are actually identical and are not penalized at this stage, since they will be receiving a penalty for repetition (previous point). The penalty received depends on if and how the confidence interval of the child rule (CI_c) overlaps with that of the parent rule

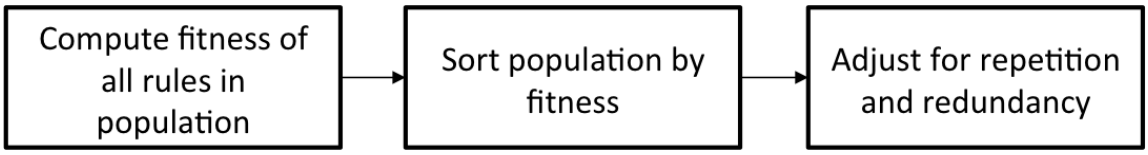

Figure 3.6: This flowcharts explains the steps to follow to compute and adjust the fitness scores of a population of rules.

($\mathrm{CI_p}$).

$$Redundancy = \begin{cases} 0, & \text{if } \mathrm{CI_c} \not\subset \mathrm{CI_p} \text{(no overlap)} \\ 1, & \text{if } \mathrm{CI_c} \subseteq \mathrm{CI_p} \text{(complete overlap)} \\ \sum_{k=1}^{P} \mathrm{CI_c} \cap \mathrm{CI_p}/P, & \text{otherwise (partial overlap)} \end{cases}$$

Where P is the number of parents of the rule that is being scored.

All the proposed metrics can potentially contribute to find in a large database the rules associated with the most relevant changes in odds ratio. However, it is possible that not all of the proposed metrics are necessary for the operation to be successful. In the following chapters, we will discuss how we tested combinations of the proposed objective functions on a series of datasets in order to find the most effective multi-objective optimization. The final fitness function is actually:

$$fitness = OR_{fitness} - length_{penalty} \tag{3.10}$$

Followed by the two adjustments, for repetition and for redundancy (Figure 3.6).

Chapter 4

Methods

4.1 Data

4.1.1 Synthetic datasets

Before testing the algorithms described in Chapter 3 on real data, we decided to test them on a group of synthetic databases with controlled effect of the exposures on the outcome. The databases were designed for two purposes: (i) to provide a bench mark for the fine tuning of the genetic algorithm fitness function, and (ii) to verify that the ARM algorithms for risk assessment are able to capture the real impact of each exposure on the health outcome.

The guideline to create a dataset with one or more embedded rules are the following:

1. First, we create the features, organized in columns. Each feature can be numerical, integer or logical, and with a different distribution (i.e. uniform, Bernoulli, normal).

2. Features of choice are selected to form the group of exposures. Exposures cause an increase in the chances of presenting the health outcome. The impact of the exposures varies (i.e. linear or step function). Furthermore, they can act alone or have synergic effects with others.

3. For each row, we determine the probability of the subject to manifest the simulated health outcome, depending on its history of exposure. Multiple exposures add up: for example, if E_1 causes a 20% risk and E_2 a 30% risk, the chances of having the health outcome will be 50%. If a subject is not exposed, it can still manifest the health outcome with a certain baseline probability.

The first 4 out of the 5 datasets have been implemented using RStudio (R version 3.1.2 "Pumpkin Helmet"), while the last one has been designed in Matlab R2014a.

- **Dataset 1:** This dataset includes 10 continuous variables of uniform distribution between 0 and 10. The only active exposure is E_1, which causes a chance of health outcome of 50% if above the threshold of 6.0. The baseline chance is 10%. 259 are cases, the remaining 741 subjects are controls.

- **Dataset 2:** This dataset includes 5 continuous variables ($E_1 - E_5$), 2 integer ($E_6 - E_7$) and 3 logical ($E_8 - E_10$). The continuous variables have uniform distribution between 0 and 10. The integer variables are uniform between 0

and 20. The logical variables have a Bernoulli distribution with p = 0.4. The baseline chance is 10%. It increases to 50% if a subject is exposed to $E_1 > 6.0$ and $E_10 = true$. 154 out of 1000 subjects are cases.

- **Dataset 3:** The feature columns of this dataset are identical to Dataset 2, with the exception of the logical variables, which have a Bernoulli distribution with a higher p (0.7). The baseline chance is 10%. It increases to 60% if a subject is exposed to $E_3 > 4.0$, $E_6 > 10$ and $E_9 = true$. 222 out of 1000 subjects are cases.

- **Dataset 4:** The feature columns of this dataset are identical to GA dataset 3. Two rules are embedded in this dataset: a subject has 40% chances of presenting the health outcome if exposed to E_8 and E_10 (versus a 10% baseline). The chances are also worsen by E_4, which has a linear impact, from 0 to +20%. The resulting dataset includes 333 cases out of 1000.

- **Dataset 5:** the most complex of the 5 datasets, it embeds 5 rules of interest. The features have been given names for ease of understanding and memorization, however they are completely artificial and are not to be interpreted as representative of a real clinical study:

  - **Age**: continuous, uniform distribution from 20 to 80 years.
  - **Gender**: binary (male = *true*), p(male)=0.5.
  - **Smoker**: continuous, from 0 to 30 cigarettes per day; p(0 = non smoker) = 0.6; remaining 40% is uniformly distributed.

- **Systolic blood pressure (SBP)**: continuous, normal (mu = 130, sigma = 25).
- **Diabetes**: binary (diabetes = *true*), p(diabetes) = 0.2.
- **Daily exercise**: categorical (none = 0, light = 1, intense = 2), uniformly distributed.

The database has a total of 7 columns (cases/controls + 6 exposures). The different features have different effects on the simulated health outcome:

- Baseline probability = 0.05.
- Age: the probability increases by 0.0025 by year of age, starting at 0 for age = 20 and ending at +0.15 for age = 80.
- Gender: no effect.
- Smoker: the impact of cigarettes has been designed as a step function. No impact up to 20 cigarettes per day, then the probability of having the health outcome goes up by 0.4 (+40%).
- Systolic blood pressure *and* diabetes: these two features have no impact unless they happen together (diabetes = *true* and pressure $\geq$ 150). If this condition is verified, the probability of having the health outcome goes up by 0.2 (+20%).
- Exercise reduces the risk of cases by 0.2 if light and 0.4 if intense. However, exercise has no effect in case of high blood pressure.

This database comprises 20,000 subjects. 3220 have been included in the group of cases, and the remaining 16780 in the controls.

Table 4.1: Summary of every rule embedded in each different dataset. The tuples (E_k, T_k) indicate the exposures and the thresholds necessary to cause an impact on the odds of experiencing the health outcome. Exposures with linear impact have no definite thresholds and are marked as $(E_k, -)$.

Datasets and embedded rules	
Dataset 1	$\{(E_1, 6.0)\} \rightarrow case$
Dataset 2	$\{(E_1, 6.0), (E_10, true)\} \rightarrow case$
Dataset 3	$\{(E_3, 4.0), (E_6, 10), (E_9, true)\} \rightarrow case$
Dataset 4	$\{(E_8, true), (E_10, true)\} \rightarrow case$ $\{(E_4, -)\} \rightarrow case$
Dataset 5	$\{(Smoker, 20)\} \rightarrow case$ $\{(SBP, 150), (Diabetes, true)\} \rightarrow case$ $\{(Age, -)\} \rightarrow case$ $\{(Exercise, 1)\} \rightarrow case$ $\{(Exercise, 1), (SBP, 150)\} \rightarrow case$

Table 4.1 summarizes the rules embedded in each dataset. In a later section of this chapter (4.2), we will illustrate how these data have been used to fine tune the fitness function of the genetic algorithm and to test the two proposed methods.

4.1.2 TEDAS-TCEQ data

The clinical data used in this study come from the Texas Emergency Department Asthma Surveillance (TEDAS), which we already mention in Section 2.2.2. The network shared the data related to 20,959 pediatric ED visits from 01/01/02 to 31/12/12

Figure 4.1: Representation of dataset expansion following case-crossover study design. The assumption made is that a subject who visits the emergency department on a given day did not visit again 1 or 2 weeks before and after the event. A similar approach has been used by Raun *et al.* [66]

(an estimate of 11,000 patients under the age of 18 in the Houston-Galveston area). Only excluded patients were those diagnosed with cardiovascular or pulmonary disease. The features of the database include demographics, insurance status, primary care provider, diagnosis and severity assessment performed by the physician, and other information.

The database has been expanded following the design of a case-crossover study to include controls (days when patients did not experience an asthma attack and/or did not visit the ED). We made the assumption that every subject did not experienced severe asthma 1 and 2 weeks before, and 1 and 2 weeks after the date of the ED visit (Figure 4.1). Therefore for every subject we have four control days. The database size after expansions includes 104,795 events.

The environmental data used in this study have been collected and shared by the Texas Commission on Environmental Quality (TCEQ). Over the Houston area (TCEQ Region 12, Figure 4.3), a total of 103 monitors are available, including 6 for CO, 7 for SO_2, 21 for NO and NO_2, 42 for O_3, and 6 for $PM_{2.5}$ (particulate matter of diameter of 2.5 μm or smaller). Each sensor records the level of pollutant(s) at 5 minute intervals. These raw data have been grouped according to date and

hour and averaged into a single hourly value, in order to reduce noise. In the next step, for every sensor and every pollutant, the maximum value recorded for each day has been annotated. The result of this operation was a complete database tracing the maximum level reached by every pollutant every day and at any available location during the desired period of time (01/01/02 to 31/12/12). A summary of the distribution of the six pollutants of interest over the Houston area from January 1^{st} 2002 to December 31^{st} 2012 is presented in Table 4.2. Figure 4.2 shows instead the correlation between the pollutant distributions.

Table 4.2: Summary of the distribution of the six pollutants under analysis over the Houston area from January 1^{st} 2002 to December 31^{st} 2012. All measures are in ppb (parts per billion), with the exception of $PM_{2.5}$.

	CO	SO_2	NO	NO_2	O_3	$PM_{2.5}$ ($\mu g/m^3$)
1^{st} quart.	297.61	1.36	2.11	9.21	33.70	14.07
Median	473.17	3.45	6.57	16.84	43.98	19.04
Mean	612.15	6.67	21.59	19.69	47.78	21.20
3^{rd} quart.	760.12	7.95	21.33	27.70	58.53	25.79
St. Dev.	522.23	9.60	42.11	13.33	20.12	12.09

In order to understand the relationship between ED visits due to asthma attacks and air quality, each subject of the expanded patients dataset has been associated to pollutant levels recorded on the date of the visit (or control dates, for non-asthma events). The TEDAS database reports the location of the patient domicile (as zip code), so it is possible to associate the patient to the closest sensors. We do not know where the patient was at the moment of the event, but since the database is related to young subjects, we assumed a close distance to their domicile. We imposed a limit of 20 km between sensor and subject zip code centroid, and picked

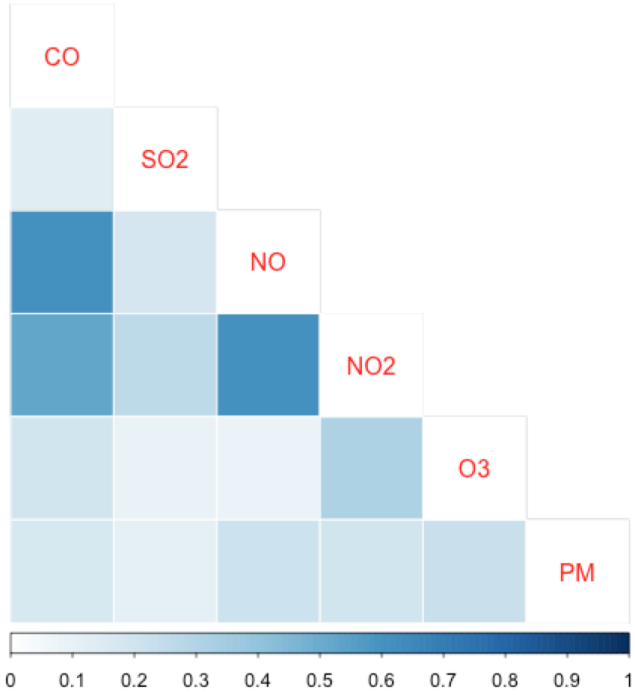

Figure 4.2: Colormap of correlation between daily pollutant distributions. Darker values indicate high correlation, while light values indicate no correlation (independence).

the closest sensor when more than one was available within that radius. This is a reasonable approximation and it is more reliable than trying to interpolate between sensors locations, because movements of chemicals in the atmosphere are influenced by a very high number of factors, such as wind and precipitation. This way we have an approximate knowledge of the air quality surrounding the patient location at the moment of the event. If a sensor is not available within the limit radius, the field for that pollutant is left blank. The total number of subjects whose zip code is within the Houston city limits and with no missing pollutant data is 14704, including 2973 cases and 11731 controls. Because some sensors were activated later in time, no event

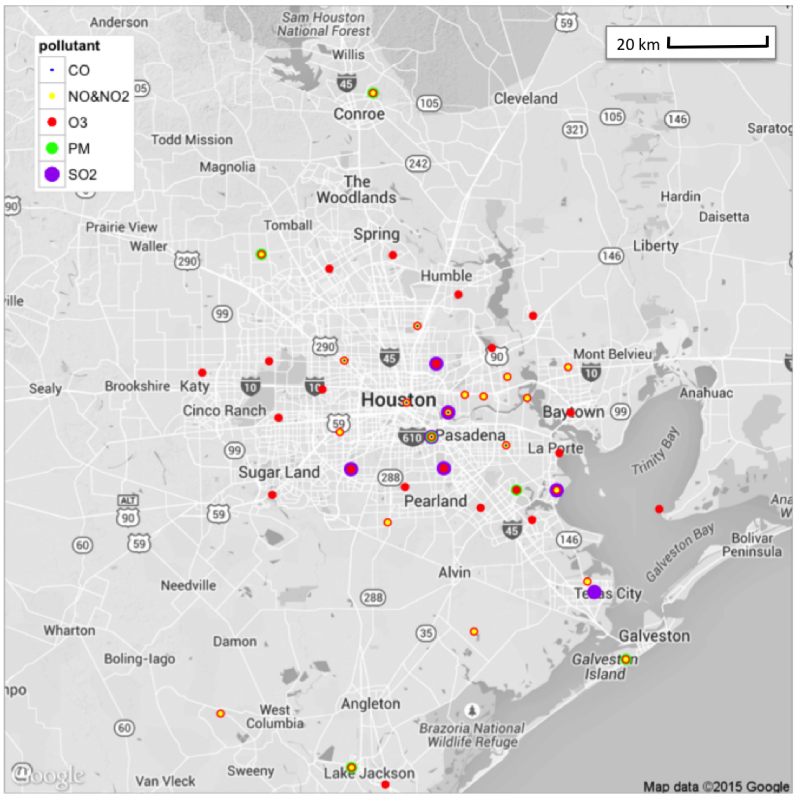

Figure 4.3: Representation of distribution of sensors of the TCEQ network over the Houston area. Some sensors are located beyond the map boundaries, and they are too far from any registered patient to be of interest, therefore were not included in this map.

antecedent to 27/02/06 had complete pollution records.

In the literature delayed actions of chemicals on the human body have been observed. For this reason not only the pollutant levels recorded on the same day of the event have been included in the subject data, but also the levels recorded from 1 to 4 days before, for a total of 30 possible exposures per subject.

The database described in this section has been used to test the proposed method for OR assessment and to gain a better understanding of interaction between asthma

exacerbation and exposure to air pollution.

4.2 Experiments

4.2.1 Apriori-OR assessment of Dataset 5

Initially, we tested our Apriori-based method for risk assessment (Method I) on the synthetic Dataset 5 described in Section 4.1.1. The goal of this experiment was two-fold: first, we wanted to verify that the method was able to capture the effect of the different exposures on the health outcome. Second, we wanted to show that the characterization provided by the ARM algorithm was more effective than traditional logistic regression.

The ARM algorithm was compared with logistic regression as implemented in the R environment (version 3.1.2). The dataset was loaded into the R console and analyzed using the `glm` command (generalized linear model). The coefficient obtained through logistic regression are interpreted and compared with the rules obtained by the ARM algorithm in Section 5.1. Categorical and numerical features had to be converted to binary form to have a valid input for the item set mining method. We chose to consider values in the highest quartile as *exposed*, while the first 3 quartiles represent a non-exposed subject. The fourth quartile thresholds for the numerical features are: cigarettes > 18; SBP > 147; age > 65. For the categorical feature *exercise*, only intense exercise is considered as exposure. The binarized dataset was inputed into the ARM algorithm.

4.2.2 Apriori-OR assessment of TEDAS-TCEQ dataset

Once we were convinced that Method I was able to capture the effects of exposures on the subjects, thanks to the experiment on the synthetic dataset, we moved to mining the real data collected for the study of asthma and pollution. Because the pollutant levels are continuous numerical features, they had to be converted to binary before being used as input for our method. We used the same quartile techniques used for the artificial dataset and defined an exposure in relation to monitor readings in the top quartile of the distribution of each pollutant. The threshold values used to bin each pollutant are reported in Table 4.3. This decision mirrors the approach currently used by the EPA, which uses single pollutant levels to issues air quality warnings. However, we decided not to use the threshold values indicated by the EPA because they were established to account for a range of health effects (particularly mortality rates), and less influenced by the more recent literature on asthma attacks. Note that, because not all pollutant data were included in the rule mining database, the top quartile is different from the one listed in Table 4.2.

With the sole exception of CO, all the top quartile thresholds are well within the range of suspected health effects on the human body around which the monitoring regime was designed. CO pollution has decreased dramatically because of the catalytic converter and the reported threshold of 696 ppb is significantly lower than EPA air quality standards. We are not confident in the precision of CO measurements at this level, given that the monitors were designed to give meaningful results around much higher concentrations [42]. Because we could not guarantee the necessary precision of CO, or its plausible pathway to health effects, those measurements have not

Table 4.3: Thresholds above which a subject is considered to have been exposed to a particular pollutant, compared with most recent EPA standards for 1-hour average value regulation (with the exception of $PM_{2.5}$, for which only a 24-hour average limit has been established). NO is not currently regulated [1].

Pollutant	4^{th} Quartile threshold	EPA standard
CO	696 ppb	35,000 ppb
SO_2	7 ppb	75 ppb
NO	37 ppb	-
NO_2	34 ppb	100 ppb
O_3	55 ppb	120 ppb
$PM_{2.5}$	24 $\mu g/m^3$	35 $\mu g/m^3$ (24-hour avg)

been included in further analysis.

We evaluated the chance of overfitting the information carried by the entire database of 14704 entries. We wanted to avoid extracting from this database erroneous information produced by factors such as noise and statistical variance. To protect us from this possibility, we relied on a training/testing strategy. The database was randomly split into training and testing sets (5000 and 9704 entries, respectively). The training set was used to compute the binning thresholds and to produce subtraining sets of different sizes. Particularly, we randomly sampled 10 sets for each size considered: 10, 20, 50, 100, 200, 500, 1000, 2000 and 5000 entries. In total, 90 different sets were generated from the original training set of 5000 entries, using a sampling with replacement strategy. Rules were mined in each of the training sets. The output was then validated on the 9704 entries forming the testing set. If all the criteria listed above were respected also in the testing phase, the rule was approved and included in the final results, which will be discussed in the Section 5.2.

4.2.3 Fine tuning of GA fitness function

In Section 3.2 we described the details of the genetic algorithm design. As we saw, different metrics were implemented and were possible candidates to include in the final fitness function. As part of a multi-objective optimization problem, all of these metrics have the potential of carrying important bits of information. However, it is not sure that all of these metrics are useful or necessary. We conducted a series of tests to determine which one of the proposed objective functions and adjustments should be included in the final GA implementation.

The fitness function was designed using an iterative process. During the first iteration, the fitness function includes only one of the five objective functions (support, confidence, OR fitness, length or extremity). This minimal fitness function is included in the GA, which is then used to mine the five synthetic datasets. By observing the rules mined using the different fitness functions, we can assign to each metric a score that reflects how effectively the algorithm was able to find the target rules. The scores, together with other qualitative observations on the results, are used to select the winner metric, which is definitively included in the fitness function. During the next iteration, the remaining metrics are added, in turns, to the winner of the previous round to create a new set of fitness functions, which are then tested on the five datasets. The winner of this round is added to the final formulation of the fitness function. The process is repeated until adding new metrics to the fitness function does not produce a visible improvement in the results. Note that the two fitness adjustments (redundancy and repetition) can only be tested from the second iteration.

To score the results of each fitness function, we start by selecting the best candidates from the final population, that is, those more similar to the target rules listed in Table 4.1. We can then assign the following penalties:

- +10 for a missed rule, in case no one of the final presented rules resembles the target rule.
- +1 for added/missed feature, in case a resembling candidate exists but it includes a non-necessary feature or omits a required one.
- $+(t - t_0)$/range for threshold values, where t is the threshold proposed by the GA, while t_0 is the target threshold. When a threshold is not specified (exposures with linear impact) this penalty is not assigned.

A candidate can be matched with one rule and one rule only. If the GA is tested over a dataset including two embedded and the final population is composed by 30 identical rules, it is clearly a miss, and it should receive a 10 points penalty. The winning fitness function is the one with the lowest score, unless other qualitative considerations on the results suggest that a different winner should be selected.

Other parameters of the GA were kept constant during this testing phase. In particular, we imposed 200 generations with a population size of 30. The initial population was created using a P_{in} of 0.25. The probability of crossover was set to 0.6, with a P_{cu} of 0.5, implicating that the two children should receive about 50% of their genes from each parent. The probability of mutation (P_m) ranges between 0.001 and 0.01, while the P_t (probability of tournament selection) ranged between

0.6 and 0.9. The size of the tournament was 4. No weights were applied to the objective metrics of the fitness function or the adjustments. For more information about the parameters of the GA, refer to Section 3.2.

4.2.4 GA-OR assessment of TEDAS-TCEQ dataset

Once the fitness function to use in the genetic algorithm was finalized, the algorithm was used to mine quantitative association rules from the TEDAS-TCEQ database, in order to find evidence of correlation between exposure to air pollutants and asthma exacerbation.

We set the GA using the same parameters specified during the tuning of the fitness function (200 generations, $P_{in} = 0.25$, $P_c = 0.6$, $P_{cu} = 0.5$, $P_m \in [0.001, 0.01]$, $P_t \in [0.6, 0.9]$. The only difference was the size of the population, which was set to 50. We chose this value because it is large enough to start seeing repeated rules in the final population, which suggests that most of the significant rules have already been found, and further increasing the population size would not help retrieving additional information.

Because this is a stochastic method, it was executed 5 times, each time changing the seed for the initial population generation. Not all of the 50 rules in the final population are necessarily suitable candidates. Some times the same candidate is repeated more than once. Other times the confidence interval of reported rules includes the value 1. This is still a possibility even though the $OR_{fitness}$ of these rules is 0. Because these associations are not significant in terms of change of risk,

they should be discharged. After the 5 algorithm executions, we preserve as valid rules those that had a large enough support (> 0.001, or at least 15 exposed cases) and that have been reported in the final population by at least two executions. The results of this experiment are described in Section 5.4.

Chapter 5

Results

5.1 Apriori-OR: results on Dataset 5

In this section, we will discuss the risk assessment obtained by the proposed ARM method in comparison with traditional logistic regression as implemented in the R environment.

After importing Dataset 5 in the R console, we used the command `summary` to visualize important descriptive of the dataset and verify that they were in line with the design we wished to implement. The results are visible in Figure 5.1. They are in line with the synthetic database design. This summary also shows that some values are not realistic (i.e. systolic blood pressure values below 40 mmHg), but they are acceptable for this test goals.

Performing basic logistic regression in R is quite simple. We used the function

```
      Case             Age            Gender         N_cigarettes
 Min.   :0.000   Min.   :20.00   Min.   :0.0000   Min.   : 0.000
 1st Qu.:0.000   1st Qu.:34.00   1st Qu.:0.0000   1st Qu.: 0.000
 Median :0.000   Median :50.00   Median :0.0000   Median : 6.000
 Mean   :0.161   Mean   :49.77   Mean   :0.4975   Mean   : 9.338
 3rd Qu.:0.000   3rd Qu.:65.00   3rd Qu.:1.0000   3rd Qu.:18.000
 Max.   :1.000   Max.   :80.00   Max.   :1.0000   Max.   :30.000
      SBP            Diabetes        Exercise
 Min.   : 33.0   Min.   :0.0000   Min.   :0.000
 1st Qu.:113.0   1st Qu.:0.0000   1st Qu.:0.000
 Median :130.0   Median :0.0000   Median :1.000
 Mean   :129.9   Mean   :0.1999   Mean   :1.003
 3rd Qu.:147.0   3rd Qu.:0.0000   3rd Qu.:2.000
 Max.   :225.0   Max.   :1.0000   Max.   :2.000
```

Figure 5.1: Descriptive statistics of the synthetic dataset, calculated by R using the command `summary`. This brief reports includes minimum and maximum value, median and mean, and first and third quartile for each column of the dataset.

`glm` for generalized linear model with the following parameters:

- `formula = Case ~ Age + Gender + N_cigarettes + SBP + Diabetes + Exercise`
- `family = ‘‘binomial’’`
- `data = Dataset 5`

The parameter `formula` indicates what function should be fitted. In this case, the log odds of the health outcome given all 6 available exposures. The second parameter specifies that the model produced is a logistic regression. And `data` simply indicates the table to use for the regression. Figure 5.2 reports the regression coefficients resulting from executing the command.

Significant association is found between the health outcome and every feature available, except gender, which was designed to have no impact. Overall, the signs

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.2171422  0.1456008 -35.832   <2e-16 ***
Age           0.0165619  0.0012121  13.664   <2e-16 ***
Gender       -0.0306010  0.0421942  -0.725    0.468
N_cigarettes  0.0928910  0.0020849  44.554   <2e-16 ***
SBP           0.0144544  0.0008593  16.821   <2e-16 ***
Diabetes      0.4600584  0.0497345   9.250   <2e-16 ***
Exercise1    -0.4925011  0.0493411  -9.982   <2e-16 ***
Exercise2    -0.9323239  0.0534097 -17.456   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.2: Estimates of the coefficients obtained using the `glm` function on the synthetic case-control study, together with their respective standard error, z value and associated p value. Every variable except *gender* is marked with the symbol "***", indicating strong correlation between the variable and the outcome. Notice that the feature *exercise*, being categorical, had to be handled as two separate variables.

of the coefficients are correct, which means that they capture correctly the quality of the interaction. R is also able to estimate the Odds Ratio associated with each variable, as reported in Figure 5.3. Note that the OR for the intercept is computed but not usually interpreted.

Despite the regression coefficients and the OR values indicating that the nature of the interaction between exposures and outcome has been captured correctly, this model can not be trusted for quantitative estimate of risk. This becomes very clear through an example, in which the coefficients are used to calculate the odds of having the health outcome in a particular group of subjects. The subjects in this scenario are 20 year-old female patients, with median blood pressure, no diabetes and not exercising. The following equation can be used to estimate the log odds for this

	OR	2.5 %	97.5 %
(Intercept)	0.005422804	0.004076508	0.007213725
Age	1.016699838	1.014287337	1.019118077
Gender	0.969862431	0.892882604	1.053479069
N_cigarettes	1.097342087	1.092867157	1.101835340
SBP	1.014559366	1.012852038	1.016269571
Diabetes	1.584166465	1.437033020	1.746364457
Exercise1	0.611096094	0.554766539	0.673145205
Exercise2	0.393637876	0.354514901	0.437078319

Figure 5.3: OR of the different exposures as calculated using logistic regression, with associated 95% confidence interval.

group of people:

$$\ln\left(\frac{p}{1-p}\right) = -5.217 + (0.017 \times 20) + 0.093x_{cig} + (0.014 \times 130) = -3.057 + 0.093x_{cig} \tag{5.1}$$

where x_{cig} is the number of cigarettes smoked by the different subjects in this group. The equation can be used to estimate the OR distribution in this group by inputing different values for x_{cig}. The results of this estimate are reported in Table 5.1

Cigarettes per day	Odds	OR (to 0 cig.)
0	0.047	-
5	0.075	1.596
10	0.119	2.532
15	0.190	4.043
20	0.302	6.426
25	0.481	10.234

Table 5.1: odds and OR for 20 year-old female patients, with median blood pressure, no diabetes and not exercising, smoking a different number of cigarettes per day, computed using equation 5.1. The OR is calculated using 0 cigarettes as reference.

According to these values, a person smoking 10 cigarettes has 2.532 times the odds of presenting the health outcome than a person who does not smoke, which

contradicts the design of the synthetic dataset (they should have the same odds). A simple logistic regression such as the one used to fit this model is not capable of capturing the step function representing the effect of cigarettes in this fictional scenario. Additionally, the synergic action of high blood pressure and diabetes, and of diabetes and exercise, are not visible at all.

This simple example shows that basic logistic regression can not be assumed to represent the real impact of each exposure on the health outcome. Clinical studies through traditional regression models are usually long trial-and-error processes. If the model does not fit the data in a satisfactory manner, the researcher can try changing some of the parameters, or increasing the complexity of the model, for example by adding splines or interaction terms. This operation is not only time consuming, but it can result in a model too complicated to be interpreted and therefore useless.

The ARM method described in Section 3.1 was used to perform risk assessment on the synthetic dataset. It reported a total of 16 rules, listed in Table 5.2, including the 5 target rules listed in Table 4.1. We can see how the method identifies the smokers in the higher quartiles as the most at risk (Rule 4). Also, the joint effect of SBP and Diabetes is correctly identified, with an odds increase of 5.87 (Rule 7). Exercise is proven to have a protective effect (Rule 5), but not when the subject also has diabetes (Rule 9): the confidence interval of this rule includes 1, and the odds of people who have diabetes and exercise are comparable with those of people who do not. Besides the 5 embedded rules, we can observe other combinations of exposures that result in different odds of experiencing the health outcome. These rules do not represent real interactions, but rather combinatory effects of features with different

effects on the odds. For example, age and smoking do not interact, but being old and smoking cause a worsening of the odds significant enough to be reported as a rule. These combinatory effects can be of interest, but in the future it would be ideal to differentiate them from actual interactions. The combination at highest risk is brought by smoking, high blood pressure and diabetes: this group has more than 40 times the odds of presenting the health outcome.

#	Exp. 1	Exp. 2	Exp. 3	Supp.	Conf.	OR	low CI	high CI
1	Diabetes			0.04	0.20	1.45	1.33	1.58
2	Age			0.05	0.20	1.46	1.35	1.59
3	SBP			0.06	0.26	2.27	2.098	2.46
4	Smoker			0.10	0.41	7.92	7.30	8.60
5	Exercise			0.04	0.11	0.55	0.50	0.60
6	Age	Diabetes		0.01	0.23	1.88	1.60	2.20
7	SBP	Diabetes		0.02	0.47	5.87	5.10	6.74
8	Smoker	Diabetes		0.02	0.44	10.46	9.04	12.10
9	Diabetes	Exercise		0.01	0.16	0.87	0.75	1.02
10	Age	SBP		0.02	0.31	3.30	2.88	3.78
11	Age	Smoker		0.03	0.46	11.66	10.18	13.36
12	Age	Exercise		0.01	0.13	0.76	0.65	0.88
13	Smoker	SBP		0.03	0.53	20.14	17.47	23.21
14	SBP	Exercise		0.02	0.25	1.68	1.49	1.90
15	Smoker	Exercise		0.03	0.34	4.50	3.98	5.08
16	Smoker	SBP	Diabetes	0.01	0.72	45.41	33.55	61.45

Table 5.2: 16 rules generated by the ARM algorithm for risk assessment when used to mine Dataset 5.

It is worth mentioning that the algorithm identified rules of interest of up to 3 exposures combined, despite being set to look for up to 6 interactions. This proves that the method was able to classify rules having four exposures combined as not significantly different from their three-exposures parents. Although division into quartiles means that we cannot identify the exact exposure thresholds, the algorithm clearly identifies the non-linear and synergic effects, which went unnoticed in the

logistic regression.

5.2 Apriori-OR: results on TEDAS-TCEQ dataset

After validation, the algorithm reported 27 rules that fit the criteria of minimum support, statistical significance, significant OR interval, non-redundancy and valid lift. Using the False Discovery Rate (FDR) controlling procedure proposed by Benjamini and Hochberg [10] we verified that the total FDR was less than 13%. Table 5.3 reports the 10 rules found more often across different training sets, while Table 5.4 reports those with the highest support. The tag "day_" before a pollutant indicates how many days before the ED visit the value was recorded.

Rule	Exposures	OR	Frequency
1	day1_O_3	1.14 (1.02 - 1.27)	8
2	day0_O_3, day0_PM	1.20 (1.02 - 1.41)	5
3	day3_NO, day0_NO_2, day2_NO_2	1.34 (1.05 - 1.70)	3
4	day0_O_3, day4_O_3	1.21 (1.03 - 1.73)	3
5	day0_NO_2, day2_O_3, day0_PM	1.33 (1.00 - 1.65)	3
6	day1_NO_2, day2_O_3, day0_PM	1.29 (1.03 - 1.61)	3
7	day0_SO_2, day0_O_3	1.23 (1.03 - 1.46)	2
8	day0_O_3, day1_PM	1.22 (1.21 - 2.18)	2
9	day3_NO, day4_NO, day1_NO_2	1.34 (1.02 - 1.75)	2
10	day1_SO_2, day3_NO_2, day2_O_3	1.36 (1.01 - 1.81)	2

Table 5.3: Set of 10 rules with highest frequency across training sets.

The support of the 27 rules varies from 0.54% to 5.82%, which means that every rule was validated on a 2x2 table including at least 52 and at most 564 exposed cases. The rule with the highest support is $\{\text{day1\_O}_3\} \rightarrow \{case\}$, which is also the

Rule	Exposures	OR	Supp
1	day1_O_3	1.14 (1.02 - 1.27)	0.06
2	day0_O_3, day4_O_3	1.21 (1.03 - 1.42)	0.02
3	day0_O_3, day0_PM	1.20 (1.02 - 1.41)	0.02
4	day0_O_3, day1_PM	1.22 (1.03 - 1.44)	0.02
5	day0_SO_2, day0_O_3	1.23 (1.03 - 1.46)	0.02
6	day3_NO_2, day1_PM	1.30 (1.08 - 1.57)	0.02
7	day1_NO, day4_O_3	1.25 (1.02 - 1.54)	0.01
8	day4_SO_2, day0_PM	1.26 (1.02 - 1.55)	0.01
9	day3_NO, day4_NO, day2_NO_2	1.28 (1.03 - 1.59)	0.01
10	day2_O_3, day0_PM, day2_PM	1.27 (1.02 - 1.58)	0.01

Table 5.4: Set of 10 rules with highest support across training sets.

only rule including only one pollutant exposure. The one with the lowest support is {day0_NO, day1_NO, day1_NO_2, day0_PM, day1_PM} $\rightarrow$ {*case*}. Naturally, rules including more exposures tend to have smaller support. The rule length varies from 1 to 5 risk factors in combination, with an average length of 2.81. The highest OR (1.54, 95% CI 1.14 - 2.08) is associated with the rule {day0_SO_2, day0_NO, day0_NO_2, day1_PM} $\rightarrow$ {*case*}. Some exposures tend to appear across rules more often than others, as visible in Figure 5.4.

Finally, we checked for correlation of pollutants within the 27 rules. Some results were expected, because of the correlation between pollutants previously shown in Figure 4.2. For example, we found that NO and NO_2 appear together in many rules. Other occurrences are more surprising. For example, we found that every rule including day0_NO always includes also day1_PM (100%), and many rules including day0_NO_2 also include day0_PM (75%).

The rules produced by this study indicate that significant correlation between

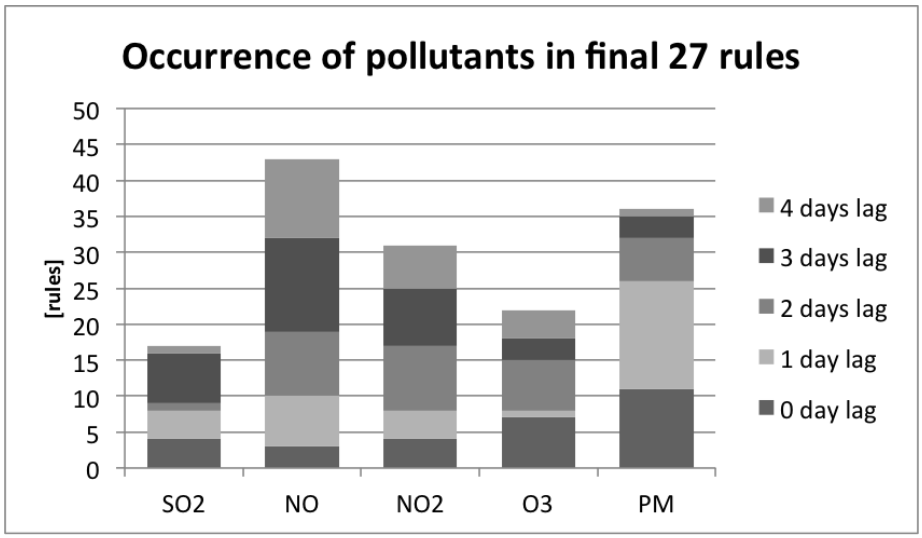

Figure 5.4: Frequency with which different pollutants at different day lags appear in the final group of 27 rules.

pediatric asthma and outdoor air pollution exists, when exposure is defined as proximity to chemical levels in the top quartile of their distribution. The rule {day1_O3} $\rightarrow$ {*case*} supports previous results in the literature [22, 53, 66], thus reinforcing the hypothesis of dangerousness associated with exposure to high levels of ozone. Ozone levels higher than 54.72 ppb have been associated with increased risk of asthma exacerbation in children. The threshold used to identify high ozone levels is below the threshold currently employed by the EPA (120 ppb for the 1-hour average exposure, although their primary metric for compliance is 70 ppb for an 8-our average [2]).

The algorithm did not find a significant correlation between exposure to NO_2 alone and odds of asthma exacerbation, which differs from some previous findings [22, 53, 66, 71]. However, NO_2 appears to cause an increase in risk when associated with other pollutants. Analogous behavior was found for NO and PM.

The rules point to hidden correlations in the data and can be ground for further analysis. The method proposed requires minimal data preprocessing and no human intervention in selecting the combination of exposures to be tested. It is particularly suitable when multiple interactions between risk factors are suspected and need to be investigated. The combinatory rules produced by the ARM method represent possible chemical interactions and it would have been challenging to identify them using interaction terms in a logistic model, particularly when more than 2 exposures are involved. The interactions found using the modified ARM can be further investigated using other methods, but initial identification is much simpler.

The constraints added to the basic Apriori rule search are effective in limiting the number of associations outputted by the algorithm. In Figure 5.5 it is possible to see how the number of rules found in the training sets is reduced at each step of the algorithm, thanks to the different filters implemented.

On the methodological side, we should also report that the training/testing strategy involving training sets of multiple sizes was necessary to find all the final 27 rules. By using the training set at the fixed size of 5000 entries, only a subset of the final 27 rules would be found. Significant rules have been found using sets of different sizes, from 5000 to 20 entries. The only sets that did not produce significant results were the smallest ones, containing only 10 subjects. We believe this strategy to be an acceptable compromise between an inclusive analysis of the possible rules included in the training set and the risk of overfitting.

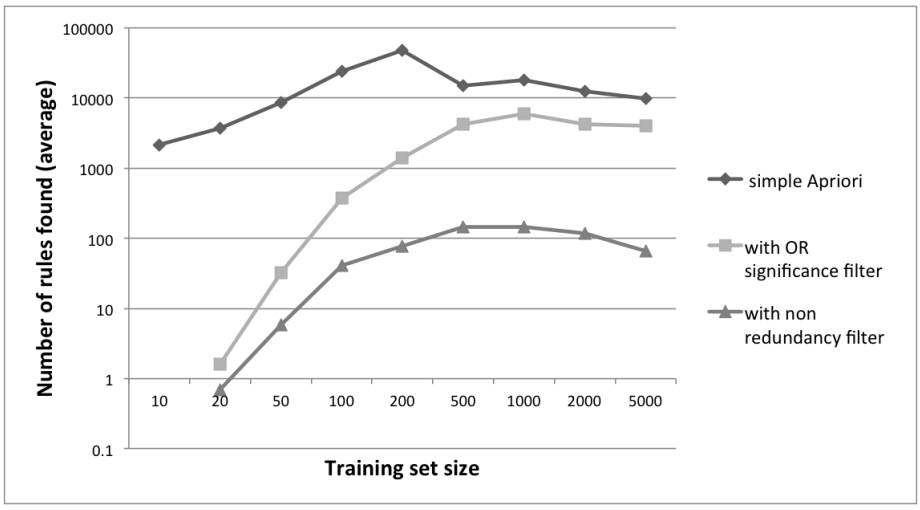

Figure 5.5: Average number of rules found at each iteration using training sets of different size. When basic Apriori search is used, thousands of associations are reported. The other lines of the chart represent the effect of adding additional filters (in sequence). When all filters are used, less than 100 rules need to be validated.

5.3 Fine tuning of GA fitness function: results

In this section, we present the results of the iterative process described in Section 4.2.3, which allowed us to select the necessary and most effective metrics among the seven originally designed and described in Section 3.2.1.

Table 5.5 summarizes the scores obtained by the different metrics at each iteration. The winner of the iteration is marked in bold characters. The best objective metric of each iteration is maintained in the fitness function. *Repetition* and *redundancy* have not been tested in the first iteration, because they are fitness adjustments and they should be applied to an existing fitness score.

The metric that received less penalties was not always selected as winner. In the

first iteration, *length* received the lowest score, but was not selected because, observing the resulting rules, it was clear that they were the result of a random selection based solely on obtaining rules of $length = 1$, in order to minimize the penalty. This random selection produced better results than other metrics in this early stage of testing. However, this is an effect of chance and increased variability in the final population. Because of lack of other criteria to measure fitness, significant rules are also hard to distinguish from the other randomly selected chromosomes in the population. For these reasons, we decided to opt for a different winner ($OR_{fitness}$), whose resulting rules were meaningful and not outputted as result of an entirely random process. A similar situation occurred during Iteration 2 with *repetition*. During the first iteration, we also decided to favor $OR_{fitness}$ over *confidence*, because the resulting scores are not significantly different, and rules associated with changes in odds ratio are the real target of this genetic algorithm.

As visible in Table 5.5, the selected metrics were $OR_{fitness}$, then *redundancy*, then *repetition*, and finally *length*. After the fourth iteration, the addition of other metrics to the fitness function actually resulted in a worsening of the performance over the 5 synthetic datasets, therefore we decided to limit the fitness function and the following adjustments to the aforementioned four.

The exclusion of the metrics *confidence* and *extremity* from the final fitness function is not entirely surprising. The strength of an association is reflected well enough in $OR_{fitness}$, which is also of higher interest given the goal of our genetic algorithm. *Extremity* is a newly designed metric that was meant to prevent distortion of selected thresholds toward the range boundaries, but it was proven unnecessary, and

	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5
OR fitness	**61.22**	-	-	-	-
Support	81.95	51.34	53.17	1.52	2.39
Confidence	59.66†	68.76	52.43	2.43	1.46
Length	26.40†	53.39	53.09	**1.12**	-
Extremity	66.00	66.14	53.89	1.71	1.81
Repetition	NA	20.33†	**2.10**	-	-
Redundancy	NA	**51.17**	-	-	-

Table 5.5: Scores obtained by the different objective metrics during the iterative process used to determine which of them should be included in the fitness function. The scores represent penalties assigned when the algorithm was not able to find the rules embedded in the dataset, therefore lowest scores indicate better performance. Occasionally, the lowest score in an iteration was not selected as winner, and they are marked by †. The reasons of these choices are explained in the paragraph. Once a metric is selected, it is finalized as part of the fitness function and it does not need to receive further scores. Because the metrics *repetition* and *redundancy* are adjustment to the fitness score, they have not been tested singularly in the first iteration.

occasionally harmful, especially when dealing with exposures with heavily skewed distributions. The fact that *support* was not selected was more unexpected. A measure of support is indirectly reflected in *length*, because shorter rules tend to have higher support, and in $OR_{fitness}$, because higher support helps narrowing the associated confidence interval. However, how we will see in the next section, it is still possible to find in the final population rules with very low support. In this case, the user may decide to discard those rules as not interesting. In the future, we may look into better solutions to incorporate the support of a rule in the implementation of the GA, but we are confident that this metric should not be included in the proposed fitness function. The presence of *support* in the fitness function resulted in negative effects such as lower performance and difficulty in determining the correct threshold,

since this metric heavily favors thresholds of lower values.

5.4 GA-OR: results on TEDAS-TCEQ dataset

In Section 4.2.4, we explained how we planned to use the proposed implementation of genetic algorithm to mine significant rules from our data on asthma exacerbation and outdoor air pollution. We run the algorithm five times on the entire dataset of 14704 subjects, changing the initial randomly generated population. Thanks to the novel algorithm, it was no longer necessary to pre-bin the pollutant values included in the dataset. Each of the five executions produced between 17 and 24 valid candidates (non-repeated rules with a OR 95% confidence interval above or below 1). We then compared the resulting list of candidates and decided to preserve rules that appeared in at least two final population, to minimize the risk of capturing associations due to noise. We also decided to impose a minimum support of 0.001% (15 or more exposed cases). Eleven rules satisfied all the required conditions, and are listed in Table 5.6.

Even if a rule is reported by different execution of the GA, it is unlikely that the exposure threshold will converge toward the exact same number. This is why in Table 5.6 thresholds are reported in terms of mean and standard deviations. A smaller standard deviation is preferable, as it indicates high agreement between different GA executions on that particular value. Large standard deviations suggest that it may not be possible to identify a more dangerous exposure threshold, and the chemical is likely to have a linear impact on increasing the odds of an asthma attack.

Summarizing the results of different GA executions using mean and standard

Rule	Exposure	Threshold	Support	OR	Frequency
1	day0_NO_2	37.61 ± 12.97	0.054	1.16 (1.03 - 1.30)	5
2	day0_O_3	62.95 ± 32.54	0.062	1.30 (1.04 - 1.67)	4
3	day1_O_3	55.38 ± 0.23	0.050	1.13 (1.03 - 1.25)	2
4	day1_PM	28.25 ± 14.47	0.060	1.20 (1.03 - 1.39)	2
5	day2_NO	235.96 ± 98.94	0.006	1.57 (1.11 - 2.28)	5
6	day2_NO_2	45.59 ± 4.80	0.020	1.19 (1.04 - 1.38)	5
7	day3_NO_2	52.00 ± 0.41	0.010	1.24 (1.03 - 1.50)	3
8	day3_O_3	79.21 ± 0.81	0.01	1.21 (1.03 - 1.43)	4
9	day4_NO	268.04 ± 37.35	0.002	1.77 (1.23 - 2.55)	5
10	day4_NO_2	26.86 ± 9.82	0.088	1.12 (1.02 - 1.22)	5
11	day4_O_3	39.12 ± 6.92	0.116	1.13 (1.04 - 1.23)	5

Table 5.6: Set of 11 rules selected by the proposed genetic algorithm and satisfying the conditions of significant OR, $supp \geq 0.001$ and occurring in at least two final populations.

deviation may not be the optimal way to preserve all the information. Consider for example the rule {day1_PM} $\rightarrow$ *case*. This rule appeared in the final population of two different runs, once with the threshold of 18.02 and once of 38.48. Observe Table 5.7 to see the values of OR and 95% CI associated with the two different rules and the statistics we would actually obtain if we were to compute OR and 95% CI after binning day1_PM with the average threshold of 28.25. The rule produced using the average threshold no longer shows an increased risk of adverse health effect. The method appears to be sensitive to small variation. In the future, this issue will need to be addressed, to guarantee both robustness and interpretability of the results.

If we compare the results obtained by Apriori-OR (5.2) with those of the genetic algorithm, we notice that the only common rule is {day1_O_3} $\rightarrow$ *case*. Curiously,

day1 PM threshold	OR	p-value
38.48 ppb	1.26 (1.02 - 1.55)	0.029
18.02 ppb	1.14 (1.05 - 1.23)	0.002
28.25 ppb	1.01 (0.90 - 1.13)	0.897

Table 5.7: OR, 95% CI and p-value associated with the rule {day1_PM} $\rightarrow$ *case* when different binning thresholds are used. The first two rows appeared in the final population produced by the genetic algorithm and show a significant risk in the odds of having an asthma attack. The last rule was obtained by binning the dataset using the resulting average threshold, and then computing the required statistics. The rule with the average threshold is no longer significant.

the GA also agrees on the binning threshold and places it at 55 ppb. This is entirely coincidental, as the GA was designed to look for the threshold associated with the highest risk in OR, while the binning threshold used in Apriori-OR was chosen without knowledge of its relationship with the outcome. None of the 11 final rules found using the GA includes more than one exposures. Rules including multiple exposures appeared in the final population of different runs, but their support was always very low (< 0.001), so they have not been deemed valid and included in the final report. Their support is lower than those of rules presented by Apriori-OR because they were normally associated with higher thresholds. Again an indication that support should be somehow incorporated in the genetic search in order to increase the chance to find this kind of information. We also also think it is harder to find significant rules including multiple exposures when the traditional definition of non-exposed population is used (exposed to *not all* the listed chemicals), as it is in the implementation of the GA. A rule with multiple exposures would need to prove to be very different from their single parents to avoid being marked as redundant.

The GA generates many more rules including one exposures only, a fact that is justified by three main differences:

- Most of the rules proposed by the GA have been assigned different thresholds than those used to mine with Apriori-OR, and we have already mentioned how small changes in the binning thresholds can have large impact on the resulting statistics.

- Validating the rules on a larger dataset (14704 versus 9704 subjects) can help producing more narrow confidence intervals for the odds ratio.

- Apriori-OR uses the alternative definition of non-exposed population when multiple exposures are involved (not exposed to *any* of the listed chemicals), so differences between the two algorithms are to be expected. We can assume the GA to be, in this sense, more strict against multi-exposure rules.

From an air chemistry point of view, the proposed rules confirm the suspicions against the dangers of exposure to ozone. O_3 is associated with higher risk of asthma exacerbation, even in presence of a large lag (4 days). NO_2 seems to be similarly dangerous. In both cases, the thresholds proposed by the GA are almost always higher than those used to bin the dataset before mining with Apriori-OR. Exposure to fine particulate matter appears to bring a short term risk increase (1 day lag). NO is the only other chemical to appear in the final list. The thresholds associated with this exposures are very high, much higher than the 75^{th} percentile. It is possible that NO has negative effects on risk of asthma only at very high concentration levels. SO_2

and CO do not appear in the final list and should probably be considered unrelated to the risk of asthma exacerbation.

Chapter 6

Conclusion

6.1 Summary of Contributions

In this dissertation, we addressed some of the difficulties associated with risk assessment in epidemiological studies. In particular, we illustrated the limitations of traditional statical tools, such as logistic regression analysis, when the data to model presents high correlation and interaction between the different exposures under consideration. These issues became evident in the recent studies on correlation between asthma exacerbation and outdoor air pollution. Interdependency and interaction between chemicals in the atmosphere, together with the impossibility of controlling subjects' exposure, make the identification of the pollutants responsible for higher risk of asthma (if they exist) very challenging. A variety of studies on the subject have been published (2.3), but the results are occasionally contradictory and no common agreement has been reached. Members of the epidemiology community

invoke the need for new multi-pollutant methods [11, 32, 75], able to overcome the limitations of traditional statistical analysis.

In response to this open problem, we developed two novel methods for risk assessment in epidemiology based on association rule mining. Both methods have been designed to find combinations of exposures associated with higher or lower odds of presenting the health outcome of interest from data collected in case-control studies. The first proposed method is based on a combination of the Apriori algorithm for frequent rule mining and a set of post-processing criteria that help preserving and highlighting the information of interest. The Apriori algorithm helps identifying associations of the form $\{exp_1, ..., exp_n\} \rightarrow \{case\}$, where $\{exp_1, ..., exp_n\} \in E$, the set of all exposures under analysis. Because some associations of interest can be unfrequent, the minimum support required for Apriori to function should be set to low values, resulting in a very high number of reported rules. Post-processing criteria are then used to find, in this long list, rules associated with a significant change in odds ratio, non-redundant and statistically significant.

This first method (Apriori-OR) has been validated on a synthetic dataset and then used to mine a collection of real data related to asthma attack events in pediatric patients and outdoor air pollution in the Houston area. The method reported 27 rules associated with significant changes in the odds of experiencing an asthma attack. In particular, it confirmed existing suspects of the danger of exposure to ozone. Rules including combinations of pollutants have also been reported and should be further investigated to understand if they represent interactions between pollutants or additive effects caused by the simultaneous presence of multiple chemicals.

The merits of Apriori-OR lay in the limited data preprocessing necessary, and in the fact that all possible combinations of exposures are tested automatically, with no need of human intervention. To find the same combinations using logistic regression, the user would have to create and test *ad-hoc* interaction terms, a challenging and time-consuming operation, especially when 3 or more exposures need to be tested. The inclusion of the post-processing criteria is also very effective in reducing the initial number of rules reported by Apriori alone, limiting the output to a manageable number of associations of interest.

The second method (GA-OR) was designed to accomplish the same task, but with a major improvement: avoiding the binning necessary to transform continuous variables into binary. We achieved this goal by implementing a novel genetic algorithm for quantitative association rule mining, specialized on finding rules associated with significant changes in the odds of presenting a given health outcome. GA-OR can handle continuous variables and assess automatically the most critical threshold of a given exposure.

The core of the method is the fitness function. A correctly designed fitness function is key to find appropriate solutions to the problem at hand. In order to produce the function that would more effectively output rules of interest, we started by considering desirable qualities in the final rules. The list included traditional metrics such as support, confidence and, of course, odds ratio. We also designed penalties for long, redundant, and repeated rules. Finally, we considered adding a penalty for rules with exposure threshold too far from the median value of the distribution, with the goal of preventing other metrics (OR and support) from skewing the threshold

toward extreme values. All the objective metrics have been designed to have values limited in the range [0,1]. After tuning the GA on five synthetic datasets, we determined that the fitness function more effective at reporting all of the embedded rules was

$$fitness = OR_{fitness} - length_{penalty} \tag{6.1}$$

Followed by the two adjustments for repetition and for redundancy (Figure 3.6). Repetition and redundancy penalties differ from the other metrics in that they need to be evaluated comparing the rule to the rest of the population. For this reason, they have not been included in the fitness function, but they are considered as a posteriori adjustments.

The finalized algorithm has then be used to mine the same data on asthma and pollution previously analyzed with Apriori-OR. The results were rather different: all of the final 11 rules produced by GA-OR are limited to a single exposure. We believe this to be a direct consequence of the ability of GA-OR to set the exposure threshold independently, which brings the algorithm to favor shorter rules and adjust the exposure threshold to the most critical value. Longer rules are rejected because they are redundant or have very low support. The only rule that appears identical in the results of the two methods is $\{day1\_O_3\} \rightarrow \{case\}$, with a threshold of 55 ppb. Other differences between the results of the two methods are attributable to the different size of the testing set and to the different definition of non-exposed population used to compute the odds ratio.

6.2 Future Work

Future research should focus on further development of the GA-OR method. In particular, it is evident that rule support should be incorporated in the genetic search to avoid the generation of very unfrequent rules. Adding rule support directly to the fitness function produced negative effects, such as a tendency to set the exposure threshold to values lower than the embedded ones. Therefore, an alternative approach will be required.

The method also appears to be sensitive to small variations in the threshold value. As illustrated in Section 5.4, we found cases of rules reported as interesting associated with different threshold values. However, when the resulting average value was used to bin the dataset, the rule no longer produced significant changes in OR. To address this issue, we will investigate a new strategy where a window around the proposed threshold values is also tested for significance, or otherwise accounted for when determining the most significant exposure levels. This should also improve the interpretability of the reported rules, as we have shown that reporting threshold values in terms of mean and standard deviation is an imperfect solution.

Regarding the particular study of association between asthma and outdoor air pollution, further steps should be taken to gain better information about the actual exposure of each subject, and about their health status at different times. At the moment, we have very approximate knowledge of the air quality around the alleged subject location, because the readings come from sensors placed above ground level and at a maximum distance of 20 km. We also had to estimate days during which

a subject was probably not experiencing an asthma attack. Although these are reasonable assumptions, a much larger amount of noise could be eliminated from the data if subjects were equipped with portable chemical sensors and followed closely to record their health conditions, possibly including other informations regarding their demographic and lifestyle.

Bibliography

[1] http://www3.epa.gov/airquality/. [Last accessed on March 10th 2016].

[2] https://www.epa.gov/criteria-air-pollutants/naaqs-table. [Last accessed on April 1st 2016].

[3] Nata - national air toxics assessments. http://www.epa.gov/ttn/atw/natamain/index.html. [Last accessed on April 8th 2015].

[4] Tedas - texas emergency department asthma surveillance. http://www.pediatricasthma.org/emergency_departments/houston. [Last accessed on April 14th 2015].

[5] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[6] B. Alataş and E. Akin. An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules. *Soft Computing*, 10(3):230–237, 2006.

[7] J. Alcala-Fdez, N. Flugy-Pape, A. Bonarini, and F. Herrera. Analysis of the effectiveness of the genetic algorithms based on extraction of association rules. *Fundamenta Informaticae*, 98(1):1–14, January 2010.

[8] Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 261–270. ACM, 1999.

[9] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, February 2008.

[10] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[11] C. Billionnet, D. Sherrill, and I. Annesi-Maesano. Estimating the health effects of exposure to multi-pollutant mixture. *Annals of epidemiology*, 22(2):126–41, February 2012.

[12] R. Brook, S. Rajagopalan, C. r. Pope, J. Brook, A. Bhatnagar, A. Diez-Roux, F. Holguin, Y. Hong, R. Luepker, M. Mittleman, A. Peters, D. Siscovick, S. J. Smith, L. Whitsel, and J. Kaufman. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the american heart association. *Circulation*, 121:2331–2378, 2012.

[13] S. Brossette, A. Sprague, J. Hardin, K. Waites, W. Jones, and S. Moser. Association rules and data mining in hospital infection control and public health surveillance. *Journal of Americal Medical Informatics Association*, 5(4):373–81, 1998.

[14] CDC. Asthma in the US, May 2011.

[15] CDC. Monitoring selected national hiv prevention and care objectives by using hiv surveillance data - united states and 6 dependent areas - 2011. *HIV Surveillance Supplemental Report*, 18(5), 2013.

[16] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang. Data mining curriculum: A proposal. In *Intensive Working Group of ACM SIGKDD Curriculum Committee*, 2006.

[17] J. Chen, H. He, G. Williams, and H. Jin. Temporal sequence associations for rare events. In H. Dai, R. Srikant, and C. Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 235–239. Springer Berlin Heidelberg, 2004.

[18] F. Chew, D. Goh, B. Ooi, and B. Lee. Time trends and seasonal variation in acute childhood asthma in tropical singapore. *Respiratory Medicine*, 92(2):345–350, February 1998.

[19] B.-C. Chien, Z.-L. Lin, and T.-P. Hong. An efficient clustering algorithm for mining fuzzy quantitative association rules. In *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, volume 3, pages 1306–1311. IEEE, 2001.

[20] P. Clark and T. Niblett. The cn2 induction algorithm. *Machine Learning*, 3(4):261–283, March 1989.

[21] S. R. Dhrubajit Adhikary. Trends in quantitative association rule mining techniques. In *IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, 2015.

[22] J. Gasana, D. Dillikar, A. Mendy, E. Forno, and E. R. Vieira. Motor vehicle air pollution and asthma in children: A meta-analysis. *Environmental Research*, 117:36–45, 2012.

[23] K. Gass, M. Klein, H. H. Chang, W. D. Flanders, and M. J. Strickland. Classification and regression trees for epidemiologic research: an air pollution example. *Environmental Health*, 13(17), March 2014.

[24] GINA. Global strategy for asthma management and prevention, 2014.

[25] M. Guarnieri and J. Balmes. Outdoor air pollution and asthma. *The Lancet*, 383, May 2014.

[26] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik. *arules: Mining Association Rules and Frequent Itemsets*, 2015.

[27] M. Hahsler, B. Gruen, and K. Hornik. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25, October 2005.

[28] Y. Han, Y. Lee, and Y. Guo. Indoor environmental risk factors and seasonal variation of childhood asthma. *Pediatric Allergy and Immunology*, 20(9):748–56, December 2009.

[29] T. Harju, T. Keistinen, T. Tuuponen, and S.-L. Kivela. Seasonal variation in childhood asthma hospitalizations in finland, 1972-1992. *European Journal of Pediatrics*, 156:436–439, 1997.

[30] W. L. J. Holcomb, T. Chaiworapongsa, D. A. Luke, and K. D. Burgdorf. An odd measure of risk: Use and misuse of the odds ratio. *Obstetrics and Gynecology*, 98(4):685–688, October 2001.

[31] J. H. Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[32] D. O. Johns, L. W. Stanek, K. Walker, S. Benromdhane, B. Hubbell, M. Ross, R. B. Devlin, D. L. Costa, , and D. S. Greenbaum. Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution. *Environmental Health Perspectives*, 120(9):1238–42, September 2012.

[33] S. B. Johnson. *The Ghost Map: The Story of London's Most Terrifying Epidemic – and How it Changed Science, Cities and the Modern World*. Riverhead, 2006.

[34] G. M. Kang, Y. S. Moon, H. Y. Choi, and J. Kim. Bipartition techniques for quantitative attributes in association rule mining. In *TENCON 2009 - 2009 IEEE Region 10 Conference*, pages 1–6, 2009.

[35] F. J. Kelly and J. C. Fussell. Air pollution and airway disease. *Clinical and Experimental Allergy*, 41:1059–1071, 2011.

[36] M. Kinell-Dunn, N. Pearce, and R. Beasley. Seasonal variation in asthma hospitalizations and death rates in new zealand. *Respirology*, 5(3):241–6, September 2000.

[37] W. LaMorte. Prospective and retrospective cohort studies. `http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_AnalyticOverview/EP713_AnalyticOverview3.html`. [Last accessed on April 10th 2015].

[38] J. M. Last. *A Dictionary of Public Health*. Oxford University Press, 2007.

[39] N. Lavrac. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, 16(1):3–23, May 1999.

[40] C. Lee, J. Chen, and V. Tseng. A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. *Computer Methods and Programs in Biomedicine*, 101:44–61, January 2011.

[41] D. Lee, K. Ryu, M. Bashir, J.-W. Bae, and K. Ryu. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. *Journal of Medical Systems*, 37(2), 2013.

[42] B. Lefer, B. Rappenglck, J. Flynn, and C. Haman. Photochemical and meteorological relationships during the texas-ii radical and aerosol measurement project (tramp). *Atmospheric Environment*, 44(33):4005–4013, 2010.

[43] J. Li, A. W. chee Fu, and P. Fahey. Efficient discovery of risk patterns in medical data. *Artificial Intelligence in Medicine*, 45:77–89, 2009.

[44] J. Li, J. Liu, H. Toiovonen, K. Satou, Y. Sun, and B. Sun. Discovering statistically non-redundant subgroups. *Knowledge-Based Discovery*, 67:315–327, 2014.

[45] W. Lian, D. W. Cheung, and S. Yiu. An efficient algorithm for finding dense regions for mining quantitative association rules. *Computers & Mathematics with Applications*, 50(3):471–490, 2005.

[46] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 125–134, New York, NY, USA, 1999. ACM.

[47] H. Liu, H. Lu, L. Feng, and F. Hussain. Efficient search of reliable exceptions. In *Methodologies for Knowledge Discovery and Data Mining*, volume 1574 of *Lecture Notes in Computer Science*, pages 194–204, 1999.

[48] M. Maclure. The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133(2):144–153, 1991.

[49] D. Martín, A. Rosete, J. Alcalá-Fdez, and F. Herrera. A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. *IEEE Transactions on Evolutionary Computation*, 18(1):54–69, 2014.

[50] F. D. Martinez. Genes, environments, development and asthma: a reappraisal. *European Respiratory Journal*, 29(1):179–184, January 2007.

[51] J. Mata, J. L. Alvarez, and J. C. Riquelme. *Artificial Neural Nets and Genetic Algorithms: Proceedings of the International Conference in Prague, Czech Republic, 2001*, chapter Mining Numeric Association Rules with Genetic Algorithms, pages 264–267. Springer Vienna, 2001.

[52] J. Mata, J.-L. Alvarez, and J.-C. Riquelme. Discovering numeric association rules via evolutionary algorithm. In *Advances in knowledge discovery and data mining*, pages 40–51. Springer, 2002.

[53] R. McConnell, K. Berhane, F. Gilliland, S. J. London, T. Islam, W. J. Gauderman, E. Avol, H. G. Margolis, and J. M. Peters. Asthma in exercising children exposed to ozone: a cohort study. *The Lancet*, 359(9304):386–391, 2002.

[54] R. J. Miller and Y. Yang. Association rules over interval data. *SIGMOD Rec.*, 26(2):452–461, June 1997.

[55] R. L. Miller and S.-m. Ho. Environmental epigenetics and asthma. *American Journal of Respiratory and Critical Care Medicine*, 177(6):567–573, 2015/04/07 2008.

[56] J. Nahar, K. Tickle, A. Ali, and Y.-P. Chen. Significant cancer prevention factor extraction: An association rule discovery approach. *Journal of Medical Systems*, 35(3):353–367, 2011.

[57] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi. A rule discovery support system for sequential medical data in the case study of a chronic hepatitis dataset. In *Proceedings of the ECML/PKDD-2003 discovery challenge workshop*, pages 154–165, 2002.

[58] C. Ordonez, N. Ezquerra, and C. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3):1–2, 2006.

[59] J. Paetz and R. Brause. A frequent patterns tree approach for rule generation with categorical septic shock patient data. In J. Crespo, V. Maojo, and F. Martin, editors, *Medical Data Analysis*, volume 2199 of *Lecture Notes in Computer Science*, pages 207–213. Springer Berlin Heidelberg, 2001.

[60] E. Park, P. Hopke, M. Oh, E. Symanski, D. Han, and C. Spiegelman. Assessment of source-specific health effects associated with an unknown number of major sources of multiple air pollutants: a unified bayesian approach. *Biostatistics*, 15(3):484–97, July 2014.

[61] S. Park, S. Jang, H. Kim, and S. Lee. An association rule mining-based framework for understanding lifestyle risk behaviors. *PLoS One*, 9(2), February 2014.

[62] M. Patel, J. Quinn, K. Jung, L. Hoepner, D. Diaz, M. Perzanowski, A. Rundle, P. Kinney, F. Perera, and R. Miller. Traffic density and stationary sources of air pollution associated with wheeze, asthma, and immunoglobulin e from birth to age 5 years among new york city children. *Environmental Research*, 111(8):1222–1229, November 2011.

[63] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, volume 229-238, 1991.

[64] H. R. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Systems with applications*, 38(1):288–298, 2011.

[65] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[66] L. H. Raun, K. B. Ensor, and D. Persse. Using community level strategies to reduce asthma attacks triggered by outdoor air pollution: a case crossover analysis. *Environmental Health*, 13(58), July 2014.

[67] L. Rodrigues and B. Kirkwood. Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. *International Journal of Epidemiology*, 19(1):205–13, March 1990.

[68] A. Sá-Sousa, T. Jacinto, L. F. Azevedo, M. Morais-Almeida, C. Robalo-Cordeiro, A. Bugalho-Almeida, J. Bousquet, and J. A. Fonseca. Operational definitions of asthma in recent epidemiological studies are inconsistent. *Cliinical and Translational Allergy*, 4(24), 2014.

[69] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transaction. In *14th IEEE International Conference on Data Engineering*, 1998.

[70] SCHER. Opinion on risk assessment on indoor air quality, May 2007.

[71] J. S. Schildcrout, L. Sheppard, T. Lumley, J. C. Slaughter, J. Q. Koenig, and G. G. Shapiro. Ambient air pollution and asthma exacerbations in children: an eight-city analysis. *American Journal of Epidemiology*, 164(6):505–517, 2006.

[72] R. Sram, B. Binkova, M. Dostal, M. Merkerova-Dostalova, H. Libalova, A. Milcova, P. J. Rossner, A. Rossnerova, J. Schmuczerova, V. Svecova, J. Topinka, and H. Votavova. Health impact of air pollution to children. *International Journal of Hygiene and Environmental Health*, 216(5):533–40, August 2013.

[73] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. *SIGMOD Rec.*, 25(2):1–12, June 1996.

[74] M. Srinivas and L. M. Patnaik. Genetic algorithms: A survey. *Computer*, 27(6):17–26, 1994.

[75] Z. Sun, Y. Tao, S. Li, K. Ferguson, J. Meeker, S. Park, S. Batterman, and B. Mukherjee. Statistical strategies for constructing health risk models with

multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health*, 12(85), October 2013.

[76] Y. Tai and H. Chiu. Comorbidity study of adhd: applying association rule mining (arm) to national health insurance database of taiwan. *International Journal of Medical Informatics*, 78(12):75–83, December 2009.

[77] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, June 2004.

[78] Q. Tong, B. Yan, and Y. Zhou. Mining quantitative association rules on overlapped intervals. In *Advanced Data Mining and Applications*, pages 43–50. Springer, 2005.

[79] G. Toti, R. Vilalta, P. Lindner, and D. Price. Effect of the definition of non-exposed population in risk pattern mining. In *5th Workshop on Data Mining for Medicine and Healthcare*. SIAM International Conference on Data Mining, May 2016.

[80] D. Vachon. Doctor john snow blames water pollution for cholera epidemic. *Old News*, 16(8):8–10, 2005.

[81] J. Wendt, E. Symanski, T. Stock, W. Chan, and X. Du. Association of short-term increases in ambient air pollution and timing of initial asthma diagnosis among medicaid-enrolled children in a metropolitan area. *Environmental Research*, 131:50–58, May 2014.

[82] WHO. Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach, 2007.

[83] X. Yan, C. Zhang, and S. Zhang. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Systems with Applications*, 36(2):3066–3076, 2009.

[84] J. Yang and F. Zhang. An effective algorithm for mining quantitative associations based on subspace clustering. In *Networking and Digital Society (ICNDS), 2010 2nd International Conference on*, volume 1, pages 175–178, 2010.

[85] C. Zhang and S. Zhang. *Association Rule Mining*. Springer, 2002.

[86] W. Zhang. Mining fuzzy quantitative association rules. In *ictai*, page 99. IEEE, 1999.

[87] H. Zheng, J. He, G. Huang, and Y. Zhang. Optimized fuzzy association rule mining for quantitative data. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 396–403, 2014.

[88] J. E. Zora, S. E. Sarnat, A. U. Raysoni, B. A. Johnson, W.-W. Li, R. Greenwald, F. Holguin, T. H. Stock, and J. A. Sarnat. Associations between urban air pollution and pediatric asthma control in el paso, texas. *Science of the Total Environment*, 448:56–65, 2013.