Effect of the Definition of Non-Exposed Population in Risk Pattern Mining

Giulia Toti[1], Ricardo Vilalta[1], Peggy Lindner[2], and Daniel Price[2]

[1]Department of Computer Science, University of Houston
[2]Honors College, University of Houston

Abstract

Odds ratios, and other traditional metrics used to quantify the correlation between illness and risk factors, rely on the correct definition of exposed and non-exposed populations. This definition has always been straightforward in conventional epidemiological studies that focus on the effect of a single risk factor. Current data mining techniques, like association rule mining, allow the evaluation of the effect of combinations of multiple risk factors. In this new scenario, what would be, the optimal definition of non-exposed population?

So far in the literature, the non-exposed group included every subject who was not exposed to *all* of the risk factors under analysis. Alternatively, we may decide to include in the non-exposed group subjects who were not exposed to *any* of them. A study to determine which definition should be favored in differing circumstances is currently missing. In this paper, we discuss possible advantages and disadvantages in using one definition or the other. We also show the differences in results obtained when the two definitions are implemented in an association rule mining algorithm and used to extract rules from a group of datasets. We conclude that these differences should not be ignored and proper attention should be given to finding the correct definition of non-exposed population in risk assessment studies.

1 Introduction.

In recent years we have observed an exponential growth in the amount of data available in the medical field. This trend creates an opportunity for modern data mining techniques, which can be employed to extract meaningful and useful information from massive repositories [11]. Techniques that can extract and report the information in the form of rules are particularly favored, because they are readily interpretable for all health practitioners, even those who do not have a background in data analytics. The most popular algorithms for extraction of rules from data can be grouped into two families: Decision Trees and Association Rule Mining (ARM). In this paper, we will focus on ARM.

Association rule mining was originally designed to find frequent associations between items in large databases [1]. Since then, different formulation of ARM have been studied, and many have also been applied in clinical environments. One of the first applications was presented by Brossette *et al.* in 1998, to study the association between hospital infections and public health surveillance [4]. Other publications include studies on chronic hepatitis, septic shock, heart disease, association deficit disorder, cancer prevention, response to drugs and general lifestyle risk behaviors [6, 12, 17, 18, 19, 20, 21, 22]. In its early formulation, ARM was designed to find rules with high support and confidence, that is, groups of elements that appear frequently in the dataset and that are highly correlated. However, sometimes associations of interest in the medical domain can be infrequent and not particularly highly correlated. Therefore, it was necessary to introduce new metrics to evaluate the interestingness of a rule. A list of these metrics and an evaluation of their effectiveness is presented in [18].

The list of objective measures available to evaluate rules includes Risk Ratio (RR) and Odds Ratio (OR). These two measures are largely used in the field of medicine and public health to establish a correlation between one or more factors and the health outcome under study. The factors implicated in the health outcome may differ widely (from genetic, to demographic, to environmental) and are generally called *exposures*. By computing risk ratio and odds ratio, it is possible to compare the exposed and non exposed populations to determine if one of them has higher chances to develop the outcome under study.

In 2009, Li *et al.* [13] presented a variant of association rule mining that abandoned the traditional support-confidence framework in favor of a pattern search guided by risk ratio. The proposed method was more efficient in covering the search space, and produced a smaller number of rules. But the number of rules in the output could still be too large for easy interpretation. Later, another paper by Li *et al.* [14] presented a method to prune redundant rules based on overlapping of the confidence interval of the odds ratio. The odds ratio is usually reported with its confidence interval to show the accuracy of the estimate. Li *et al.* used confidence intervals to determine if a rule and its parent are statistically different. If the confidence intervals do not overlap, the rules must carry different

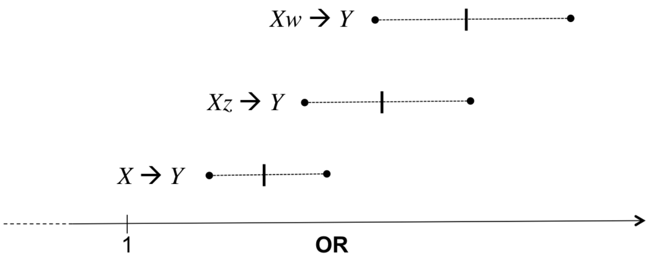

Figure 1: Schematic representation of OR confidence interval of different rules. Rule $X \rightarrow Y$ is the parent. By adding other exposures to the parent rule, we obtaine the new rules $Xz \rightarrow Y$ and $Xw \rightarrow Y$. Because only the confidence interval of $Xw \rightarrow Y$ does not 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.

information; otherwise, they are considered equivalent and the subrule is pruned. A schematic representation of this concept is visible in Figure 1.

The odds ratio measures the correlation between the exposure under study and a particular health outcome by comparing two groups of subjects, exposed and non-exposed (Eq. 2.2). Traditional epidemiological studies analyze one risk factor at a time. While other factors can be included in the data collection to control for confounders and interactions; the interest is usually limited to one new exposure that has not been studied before. In this way, defining exposed and non-exposed populations becomes straightforward. But in ARM, multiple risk factors are often combined to form a rule, therefore a new question arises: given a set of exposures, which subjects should be included in the non-exposed group?

So far in the literature [14, 18], researchers have chosen to define non-exposed subjects as all those subjects other than those exposed to *all* the factors included in the rule. This definition is also implemented in popular software for ARM such as the *arules* package in R [8, 9]. An alternative definition, which to our knowledge has not been discussed before, includes in the non-exposed population only those subjects who have not been exposed to *any* of the factors included in the rule. In this paper, we will discuss possible advantages and disadvantages of the two definitions and we will show their impact when the pruning criteria described in [14] is used.

2 Problem definition.

We will start this section with some of the foundations of rules and rule mining. A rule represents an association between two sets of items, i.e. X and Y. The notation $X \rightarrow Y$ indicates that when X occurs Y also occurs with a probability $P(Y|X)$. In association rule mining, rules are extracted from large binary databases. The columns of the database represent the set of possible items $I = \{i_1, i_2, ..., i_m\}$. A subset of items $X \subseteq I$ is called an *itemset.* The rows of the database represent all instances, or transactions, that occurred in the dataset. In a medical study, each row represents a different subject of study, columns are used to represent characteristics or conditions of the corresponding subject. Possible items could be $\{Male\}$ or $\{Age : 30 \div 40\}$ or $\{Smoker\}$. Medical data are mined to find significant rules such as $\{Smoker, Age : 30 \div 40\} \rightarrow \{Lung\ cancer\}$. Similar rules may have low confidence (not all smokers in their 30s have lung cancer!), but help finding a significant change in risk for exposed and non-exposed population.

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 its first ARM formulation [1], forms the support-confidence framework. The support of a rule represents how often the items of the rule occur together in the dataset ($supp(XY) = P(Y \wedge X)$). The confidence of a rule measures the chance of finding the itemset Y (also called *consequent* or *RHS, right hand side*) in a transaction, given the presence of the itemset X (also called *antecedent* or *LHS, left hand side*). Therefore, the confidence is simply a conditional probability:

$$conf(X \rightarrow Y) = P(Y|X) = \frac{supp(XY)}{supp(X)} \tag{2.1}$$

The support-confidence framework requires selected rules to have support and confidence larger than some minimum thresholds, normally imposed by the user according to the situation. But in a public health study this framework represents a limitation. Some of the interactions between exposures and health outcome can be infrequent or not very strong, but still significant, especially if they capture an important difference between who is exposed and who is not. Fortunately other metrics can be used to determine if a rule is significant for medical purposes. One of them is the odds ratio:

(2.2)

$$\begin{aligned} \mathrm{OR}(X \rightarrow Y) &= \frac{P(X \wedge Y)/(1 - P(X \wedge Y))}{P(\neg X \wedge Y)/(1 - P(\neg X \wedge Y))} = \\ &= \frac{supp(XY)/supp(X\neg Y)}{supp(\neg XY)/supp(\neg X \neg Y)} \\ &= \frac{supp(XY)supp(\neg X \neg Y)}{supp(\neg XY)supp(X \neg Y)} \end{aligned}$$

The odds ratio allows us to compare health outcomes in two populations differentiated by some exposure(s). A positive (negative) correlation between outcome and exposure exists if the OR is greater (less) than 1. OR = 1 indicates no correlation. Epidemiological studies normally report odds ratios with their confidence interval (CI) $[exp(log(\text{OR}) - \omega), exp(log(\text{OR}) + \omega)]$, where

$$\begin{aligned}\omega = & z_{\alpha/2}\sqrt{\frac{1}{supp(XY)}+\frac{1}{supp(\neg XY)}+} \\ & \overline{+\frac{1}{supp(X\neg Y)}+\frac{1}{supp(\neg X\neg Y)}}\end{aligned} \tag{2.3}$$

z is the critical value of the confidence interval, and it is typically equal to 1.96, for a 95% level of significance.

ARM can be used to mine rules of the form $X \rightarrow Y$ with a significant odds ratio (that is, an odds ratio which confidence interval does not cross 1). X represents a set including one or more exposures and Y is the health outcome under study. The great value of ARM is that it allows us to explore the impact of all possible combinations of exposures and report only those that produce an interesting OR.

Unfortunately, this application, like many other forms of ARM, is affected by the problem of redundant rules. A redundant rule is a rule whose LHS could be simplified by reducing the number of items without any loss of information. For example, the rule $\{Pregnant\} \rightarrow \{Age : 20 \div 40\}$ is just as informative as the rule $\{Pregnant, Female\} \rightarrow \{Age : 20 \div 40\}$. Or, in the case of a medical study, adding an exposure may not change the odds of having the health outcome. Consider for example the rules $\{Smoker, Female\} \rightarrow \{Lung\ cancer\}$ and $\{Smoker\} \rightarrow \{Lung\ cancer\}$, resulting in the same odds ratio. Clearly, smoking is responsible for the health outcome. The fact that some subjects were females and smokers did not worsen their odds, even if women and men have different levels of risk. Therefore, the simpler rule should be preferred. Not controlling for redundant rules can cause the number of output rules to grow exponentially and make the results impossible to understand. Kotsiantis and Kanellopoulos [10] offer a good overview of association rule mining techniques and open questions, including a paragraph on redundant association rules. The most popular methods include selection of k best rules [2, 5], mining only maximal itemsets [3, 7], and integration of external knowledge to facilitate the search. This latter became particularly popular in mining relationships in gene expression data [16, 15].

Li *et al.* built an algorithm [14] based on the following assumption: if adding an exposure to a rule does not produce a significant change in OR, the rule should not be reported. The odds ratio between two rules is significantly different if their 95% confidence intervals do not overlap. The assumption seems reasonable, but it is affected by how the non-exposed population is defined in the presence of multiple risk factors. Changing the non-exposed population results in different odds ratio, and as a consequence must be designed carefully. The definition of the exposed population when the antecedent X includes more than one exposure is straightforward: all risk factors must be present at the same time. However, in the current literature, $\neg X$ includes every other possible scenario. This may create some confusion in the interpretation of the rule, because the comparison group is non-homogenous (it includes partially exposed and completely non-exposed subjects). We also argue that it results in a wrong comparison between rules. Consider the rule $\{X\} \rightarrow \{Y\}$ and its child $\{Xw\} \rightarrow \{Y\}$, where w is a single new exposure ($w \notin X$). We observe the confidence intervals of the rule to determine if they are statistically different. The odds ratio of the parent rule is computed against a non-exposed population composed by the union of $(\neg Xw)$ and $(\neg X\neg w)$. But the odds ratio of the child rule includes in the non-exposed population $(\neg Xw)$, $(\neg X\neg w)$, and also $(X\neg w)$. The non-exposed group of the child rule is not included in the non-exposed group of its parents. This makes the rule intrinsically different.

The alternative definition of non-exposed population, which includes only subjects that have not been exposed to any of the risk factors under examination $((\neg X\neg w))$, results in easily interpretable rules, because the non-exposed group is homogenous. It also offers a more consistent comparison between child and parent rule, because now the non-exposed group of the child rule is a subset of the parent non-exposed group. However, this definition may be problematic as it reduces significantly the size of the non-exposed group, thus reducing the power of the analysis. Furthermore, the comparison with a completely non-exposed population results in higher odds ratios that must be interpreted carefully.

We performed a set of tests to observe the differences in performance obtained by the two definitions of non-exposed group. The following section describes the data used for the experiments and the different tests conducted, followed by a summary of results and discussion.

3 Experimental setting.

3.1 Method. We used a basic A Priori association rule mining algorithm to extract rules from the data (described in details in the following section). We conducted our experiment using the Rstudio environment and the *arules* package [8, 9]. We used three different variations of the basic algorithm to evaluate the effect of pruning and of using different population definitions:

- **Traditional (Trad):** the first method uses the traditional definition of non-exposed population, that is, any subject who has not been exposed to all the risk factors included in the LHS. No pruning criteria is used to filter redundant rules.
- **Traditional + Pruning (TradP):** this method adds to the traditional definition of non-exposed population a pruning criteria of redundant rules based on overlapping of 95% CI.
- **Alternative + Pruning (AltP):** the last method uses the same CI based pruning criteria, but the non-exposed population used to compute the OR is limited to subjects who have not been exposed to any of the risk factors included in the LHS.

All methods have low thresholds for minimum support (1%) and confidence (0.0%) to preserve a large number of rules and observe the differences in the results. The rules that satisfy the requirements of minimum support and confidence were checked for statistically significant OR confidence interval. Only rules with an interval that does not cross 1 were included in the output (for all three methods).

3.2 Data. First, we tested the three methods described in the previous section on six synthetic datasets including 20,000 subjects and 51 features (one indicating whether the subject is a case or a control, the other 50 describing the exposure history). A single rule was embedded in each of the six datasets. By knowing in advance what rule should be found in each dataset, it was possible to evaluate the performance of each algorithm. Embedded rules have different lengths (from 1 to 3 risk factors in combination) and different strengths (weak, $P(Y|X) = 0.4$ or strong, $P(Y|X) = 0.8$). All the features not included in the embedded rule have no impact on the outcome and are potential sources of noise. A baseline probability $P(Y|\neg X) = 0.1$ was introduced to create a population of exposed controls, the absence of which would result in infinite OR. Table 1 offers a summary of the six datasets.

ID	Rule length	Strength	Cases
1	1	weak	3795
2	1	strong	6197
3	2	weak	2486
4	2	strong	3184
5	3	weak	2124
6	3	strong	2342

Table 1: :
Description of the six single-rule synthetic datasets used in the experiment. Each dataset was embedded with a rule of different length and strength. The last column reports how many of the 20,000 subjects included in each datasets are cases.

We also tested the three methods on a more complex synthetic dataset designed to have a controlled interaction between exposure and health outcome. The data represent a case-control study including 3220 cases and 16780 controls. The database includes six exposures designed to have a different impact on the chances of developing the health outcome. Features are named for ease of understanding. However, the data is not representative of a real clinical study. We gave subjects the following features across the database:

- Age; continuous, uniform distribution from 20 to 80 years.
- Gender; binary (male = 1), 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 = 1), p(diabetes) = 0.2.
- Daily exercise; categorical (none = 0, light = 1, intense = 2), uniformly distributed.

The features have been designed to have different impact on the chances of contracting the disease. Every subject starts from a baseline probability of 5%. Exposures can have a gradual impact or only act after a certain threshold. They can also be affected by other exposures. Here is the complete list:

- 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 developing the health outcome rises by 0.4 (+40%).

- High SBP and diabetes: these two features have no impact unless they happen together (diabetes = true and pressure ≥ 150). If this condition is verified, the probability of the event 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.

The database described above includes 5 embedded meaningful rules: 3 caused by single exposures (Age, Smoker and Exercise), and 2 caused by interaction between exposures (high SBP with Diabetes, and high SBP with Exercise). A good rule miner should capture all these rules and avoid other less meaningful rules. Less meaningful rules can be divided into two categories: rules caused by simultaneous presence of two or more risk factors, and truly redundant rules. The first category includes those rules that do not represent true interaction between risk factors, but produce a different odds ratio because of their simultaneous presence. For example, we expect the rule $\{Age, Smoker\} \rightarrow \{Event\}$ to result in a higher odds ratio than its single parent rules. Although the rule is not representative of a real interaction between risk factors, it can still be of interest for the study; therefore, we do not penalize methods that output these associations. Truly redundant rules include risk factors whose removal would result in no changes in odds ratio. For example, we expect the rule $\{Male, Smoker\} \rightarrow \{Event\}$ to have approximatively the same OR of $\{Smoker\} \rightarrow \{Event\}$, because gender has no impact. In this case, the longer rule has no added utility and should be avoided.

4 Results.

We recorded the number of rules reported by the different methods when they were used to mine the six one-rule datasets. Ideally, the output should be limited to the one embedded rule. However, this is highly unlikely because of the noise in the data and correlations introduced when embedding the rule. A good output should include the embedded rule and limit the number of other associations.

Every method was able to find the embedded rule in all of the six datasets. The total number of rules found was variable, as visible in Figure 2. Because of the low support and confidence thresholds used and the absence of a pruning criterion for redundant rules, the Traditional method reports a very high number of rules, sometimes over a thousand. This proves that pruning for redundancy can be very useful in lowering the number of output rules, when other selection criteria are missing or less strict.

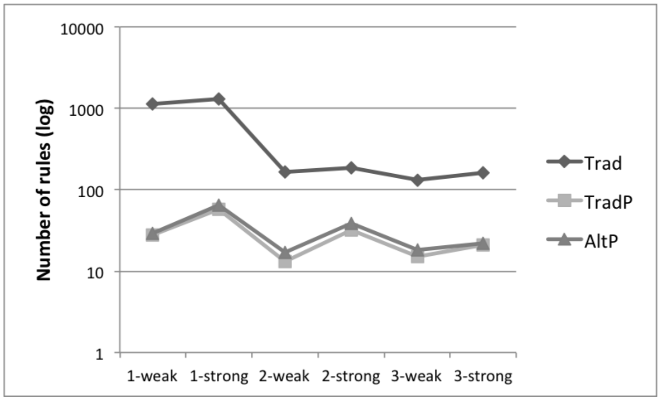

Figure 2: Number of rules found by the different methods in the six one-rule datasets. Datasets are labeled on the x-axis using length of embedded rule and strength of association.

TradP and AltP report a very similar number of rules, although the first method appears slightly more effective at filtering rules and returns in average 3.83 rules less than AltP, per trial. No remarkable difference was found in the overall value of the reported odds ratios and p-values.

When tested on the more complex synthetic dataset, Trad output 23 associations, including the embedded five. 9 rules represented additive effects between risk factors. And 9 of the 23 reported rules were redundant, as they were composed by simpler rules and the risk factor $\{Male\}$, which we know by design has no effect. Again, the high number of redundant rules output shows that this method alone is not effective for the task.

TradP output 14 associations, including the 5 embedded in the synthetic set. 7 rules represented additive effects between risk factors. Two redundant rule were also included: $\{Male, Smoker\} \rightarrow \{Event\}$, and $\{Male, Diabetes\} \rightarrow \{Event\}$.

Alt3 reports a total of 14 rules: the 5 most significant plus 9 additive effects. No redundant rules are reported. A summary of the rules found by each algorithm is visible in Figure 3.

5 Conclusions.

We confirmed that mining with no pruning criteria produces a high number of redundant rules, thus proving the necessity of a process for their elimination. TradP and AltP were both effective in reducing the number of rules and the size of their output is almost identical. However, AltP appeared to be slightly more effective at eliminating redundant rules in a more complex scenario. TradP produced some undesired results, in

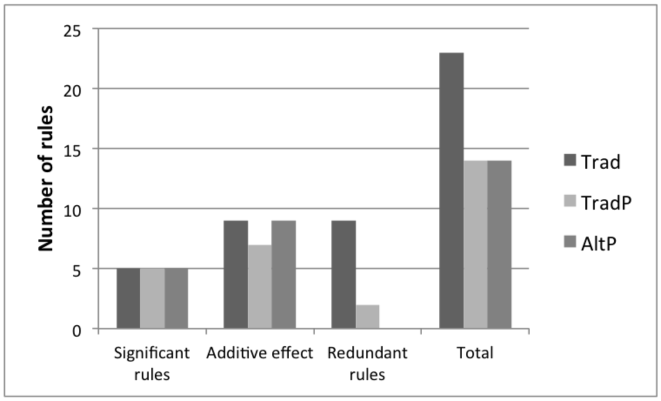

Figure 3: Number of rules found by the different methods. The first three groups of columns represent rules of different quality. The significant rules are important and should be preserved. Additive effects are tolerable. Redundant rules should be avoided.

the form of the rules $\{Male, Smoker\} \rightarrow \{Event\}$, and $\{Male, Diabetes\} \rightarrow \{Event\}$. As suspected, comparing the combination $\{Male, Smoker\}$ against a mixed non-exposed population resulted in an OR different from the parent rules (it is more than $\{Male\}$, but less than $\{Smoker\}$), tricking the algorithm into thinking they were significantly different. If the comparison were made against the uniform non-exposed population, the OR would be similar to the rule with the single $\{Smoker\}$ risk factor and would be pruned. Similar circumstances occurred for $\{Male, Diabetes\} \rightarrow \{Event\}$.

AltP was the only one capable of avoiding all truly redundant rules when mining the more complex database, thanks to a more consistent comparison between populations of child rules with their parents. However, it reported a slightly higher number of rules in the tests done using the one-rule datasets.

TradP appeared to be more resistant against interaction and produced fewer rules caused by additive effects between exposures than AltP, possibly because comparing non-homogenous populations may require more significant differences to be present to produce the necessary change in the odds ratio.

This experiment shows that different definitions of non-exposed groups can be used when using ARM for risk estimate. The differences in using one or the other definition may seem unimportant in these simple mining scenarios, however they represents on a small scale the risk of using the wrong method when mining association rule in large medical databases. In the future, the three methods should be tested on real datasets to better understand their performance when mining perturbed data. We currently do not know what is causing the differences in performance over the proposed datasets. We believe that exploring this question would be beneficial for the development of medical data mining for risk evaluation and of interest for the participants of the workshop.

References

[1] 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.

[2] M. Atzmueller and F. Puppe. Sd-map – a fast algorithm for exhaustive subgroup discovery. In *Knowledge Discovery in Databases: PKDD 2006*, pages 6–17. Springer, 2006.

[3] J. Bayardo and R. J. Efficiently mining long patterns from databases. *SIGMOD Rec.*, 27(2):85–93, jun 1998.

[4] 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.

[5] R. Cai, Z. Hao, W. Wen, and H. Huang. Kernel based gene expression pattern discovery and its application on cancer classification. *Neurocomputing*, 73(13-15):2562–2570, aug 2010.

[6] 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.

[7] K. Gouda and M. J. Zaki. Genmax: An efficient algorithm for mining maximal frequent itemsets. *Data Min. Knowl. Discov.*, 11(2):223–242, nov 2005.

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

[9] 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.

[10] S. Kotsiantis and D. Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006.

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

[12] 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.

[13] 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.

[14] 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.

[15] Y.-C. Liu, C.-P. Cheng, and V. S. Tseng. Discovering relational-based association rules with multiple minimum supports on microarray datasets. *Bioinformatics*, 27(22):3142–3148, 2011.

[16] R. Martinez, N. Pasquier, and C. Pasquier. Mining association rule bases from integrated genomic data and annotations. In *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 78–90. Springer, 2008.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.