 doi:10.20944/preprints201911.0117.v1

*Article*

# Identification of Malignant Mesothelioma Risk Factors through Association Rule Mining

**Talha Mahboob Alam**
Computer Science and Engineering Department
University of Engineering and Technology
Lahore, Pakistan
Talhamahboob95@gmail.com

**Abstract:** Malignant mesothelioma is a rare proliferative cancer that develops in the thin layer of tissues surrounding the lungs. Malignant mesothelioma is associated with an extremely poor prognosis and the majority of patients do not show symptoms. The epidemiology of mesothelioma is important for the identification of disease. The primary aim of this study is to explore the risk factors associated with mesothelioma. The dataset consists of healthy and mesothelioma patients but only mesothelioma patients were selected for the identification of symptoms. The raw data set has been pre-processed and then the Apriori method was utilized for association rules with various configurations. The pre-processing task involved the removal of duplicated and irrelevant attributes, balanced the dataset, numerical to the nominal conversion of attributes in the dataset and creating the association rules in the dataset. Strong associations of disease's factors; asbestos exposure, duration of asbestos exposure, duration of symptoms, erythrocyte sedimentation rate and Pleural to serum LDH ratio determined via Apriori algorithm. The identification of risk factors associated with mesothelioma may prevent patients from going into the high danger of the disease. This will also help to control the comorbidities associated with mesothelioma which are cardiovascular diseases, cancer-related emotional distress, diabetes, anemia, and hypothyroidism.

***Keywords:*** Malignant mesothelioma, Epidemiology, Association rule mining, Apriori method, Imbalanced dataset

## I. Introduction

The second half of the twentieth century and the start of the new era is a period in which life expectancy in the world has increased. Modem medicine and improved sanitation have eradicated many contagious diseases like cholera [1]despite the massive research work and prompt development, cancer remains a deadly disease in the world. In 2018 there were 20.7 million new cancer patients diagnosed globally and 11.6 million people would die due to cancer in the same year [2]. Malignant mesothelioma is a rare but highly aggressive tumor which arising from mesothelial cells that cover the lungs, pericardium, and peritoneum [3]. Infections and inflammation are common symptoms for diagnosis of malignant mesothelioma [4]. The incidence of malignant mesothelioma is still increasing due to the long latency time [5]. Australia has the highest reported incidence of malignant mesothelioma with annual rates of 40 cases per million [6]. In the United States, the incidence is 15 cases per million with statistically significant elevation of incidence in shipbuilding region[7]. In Europe, the average incidence is 18 cases per million, with the highest incidence in Great Britain and the Netherlands with 33 and 30 cases per million respectively. The incidence is especially high in areas with large shipyards [8]. The worldwide mortality of malignant mesothelioma from 1980-2020 shows the elevation in a number of deaths each year[2]. Epidemiology is the founding science of public health. Epidemiology of mesothelioma is important for the identification of disease. 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 [9].

With this large data resource in medical, the manual analysis had become very arduous [10]. The potential for extracting useful information from the biological data sets is immense. A large amount of data in bioinformatics and medicine has changed the field into an information science which makes processing and interpretation of such large data sets a challenging area for computer scientists,

 doi:10.20944/preprints201911.0117.v1



biologists and medical practitioners [11]. Advances in DNA sequencing and disease diagnosis techniques have created many opportunities in bioinformatics [12]. Data mining is the process of discovering useful knowledge from data. It includes the application of various methodologies and algorithmic approach to pre-process, cluster, classify and associate the information for useful knowledge retrieval [13].World health organization, World Lung Foundation, and other international health organizations are working together for early detection of malignant mesothelioma. The detection of mesothelioma was done with the help of nomograms [14],CT images [15-17] and blood samples [18]. With the rise of computing methods for disease prediction, data mining also played a vital rule to predict malignant mesothelioma. Several data mining models were utilized for early diagnosis of malignant mesothelioma [19, 20]. Geographical, Socioeconomic and clinical factors were used for early prediction of mesothelioma with the help of data mining techniques [21, 22]. Epidemiology of malignant mesothelioma is a very important aspect of the disease. Without knowing the epidemiology disease, prevention and proper medication is a difficult task for epidemiologist. Limited work has done on the epidemiology of malignant mesothelioma because previous researches focused on detection of disease rather than associated risk factors. Furth more, a huge work was done based on CT images for detection but it is not possible to extract the associated factors with malignant mesothelioma by using images because images contained less information. Some previous researches showed that asbestos exposure is a major cause of malignant mesothelioma [23, 24] but various studies negated this claim that there is no relation between malignant mesothelioma and asbestos exposure [25, 26]. The levels of neutrophil/lymphocyte ratio (NLR) and platelet/lymphocyte ratio (PLR) in the blood are also associated with malignant mesothelioma [27] But association other than complete blood count parameters levels like hemoglobin, blood lactic dehydrogenase (LDH) and glucose is still unknown. Other factors related to pleura like pleural LDH, pleural glucose, pleural effusion, and pleural thickness also need to be investigated. The impact of socioeconomic and geographic factors upon survival in patients of malignant mesothelioma is not clear yet [28]. In this study, we identified the risk factors that cause the disease of malignant mesothelioma.

## II. Related Work

Shuai Wang et al. [14]established a nomogram based method to forecast the survival of mesothelioma patients. The data set included surveillance, epidemiology, and results of patient disease. Only 7.1% (1092) patients were selected with mesothelioma from whole SEER database to analyzed the clinical and pathological features. The survival rate of each patient was calculated with the time duration of 1, 2 and 3 year. Their developed nomogram model was accurately predicted the survival of a patient the help of the clinical features of the patient. Maria Bonomi et al. [15] classified mesothelioma patients into different stages according to their disease severity. The obtained dataset also contained patient history, treatment details, demographics and laboratory results along with CT images. The obtained images were analyzed based on tumor volume and fissural thickness of the tumor. Mitchell Chen et al. [16] proposed a computer-based semi-automated method to classify mesothelioma CT images of patients into different groups. A walk based method was applied to segment the images according to their tumor. The proposed method handled the boundaries of weak tissues and abnormal shapes of the tumor. Their proposed method outperformed radiologist in the sense of time. Xue Hu and Zebo Yu [19] automatically diagnosed mesothelioma patients by using deep learning. Significant attributes were selected from the dataset with the help of a genetic algorithm and relieff method. Several deep learning models were applied on the dataset which is the back propagation algorithm, extreme learning algorithm, and stacked sparse autoencoder (SSAE) algorithm. Genetic algorithm+SSAE efficiently classified the dataset in terms of ROC curve, F-measure, accuracy, specificity, and sensitivity and given the best results. Kemal Tutuncu and Ozcan Cataltas[20] diagnosed mesothelioma disease by the utilization of different classification methods. 9 different types of classification methods were applied on the dataset which obtained from Dicle University, faculty of medicine. These applied methods were an artificial neural network, J48, Multiclass classifier, SMO, PART, Bayes net, random committee, logistic and LMT. An artificial neural network was given the highest accuracy among other methods. Sabyasachi Mukherjee [21] predicted mesothelioma disease by using two different data mining

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0117.v1



methods. Support vector machine and MLPE neural network methods were applied on dataset. To validate the results Cross-validation method used because the data spilt method only used some portion of data. By comparing the results of both models, ANN was given better results. The results of both models were computed on the bases of F1-measure, AUC curve, accuracy, and ROC curve. Mehrbakhsh Nilashi et al. [22] developed an intelligent system by using data mining techniques to enhance the classification accuracy of mesothelioma disease. They used clustering and classification methods to develop their system. A hybrid model was developed for clustering and classification tasks by using expectation maximization and naïve Bayes. 10- Fold cross-validation was used to obtain the unbiased results. Their hybrid method outperformed other classifiers in terms of accuracy. Luigi vimercati et al. [23] studied the association between mesothelioma and asbestos exposure. They included three mesothelioma patients from a regional operational unit of the Apulia region. All the information related to patient disease was collected through face-to-face interviews. Their results stated that asbestos is associated with mesothelioma. Tommaso A.Dragani et al. [24]established that the diagnosis of mesothelioma at the younger age is associated with the heavier exposure of asbestos. The patient's data included age, sex, age when diagnosis, location of tumor and whether the patient has exposure of asbestos in his childhood. Results showed that asbestos in lung tissues at younger age associated with quantity of asbestos as well as mesothelioma. Seda Tural Onur et al. [27] investigated the viability of NLR and PLR as prognostic indicators in Mesothelioma. The data of patients included complete blood count parameters, NLR, MPV, platelet count, and red blood cell distribution width. PLR was a significant prognostic factor for mesothelioma patients but not NLR. Furthermore, the reliability of these parameters is not still proved in a large number of studies. Venkites waran Muralidhar et al. [25]claimed that there is no relation in mesothelioma and asbestos exposure on the basis of certain evidence. Asbestos has consumed in India more than any other country in the world. Since, India consumed asbestos more than any country in the world but there was only one reported case of mesothelioma. Bharat Jasani et al. [26] established a relation that Mesothelioma is not associated with asbestos exposure by using recent researches. Only 10 to 20 percent people exposed by asbestos leads to mesothelioma which is not large ratio as well as after detailed assessment, 20% people have diagnosed mesothelioma who have never exposed by asbestos in their life. It was concluded that asbestos exposure was not the only parameter which causes mesothelioma, some other factors are also contributed to causing malignant mesothelioma.

## III. Methodology

In this study, a data mining framework for identification of risk factors which contains five steps as shown in figure 1. These steps consist of a selection of dataset, data pre-processing and modeling.

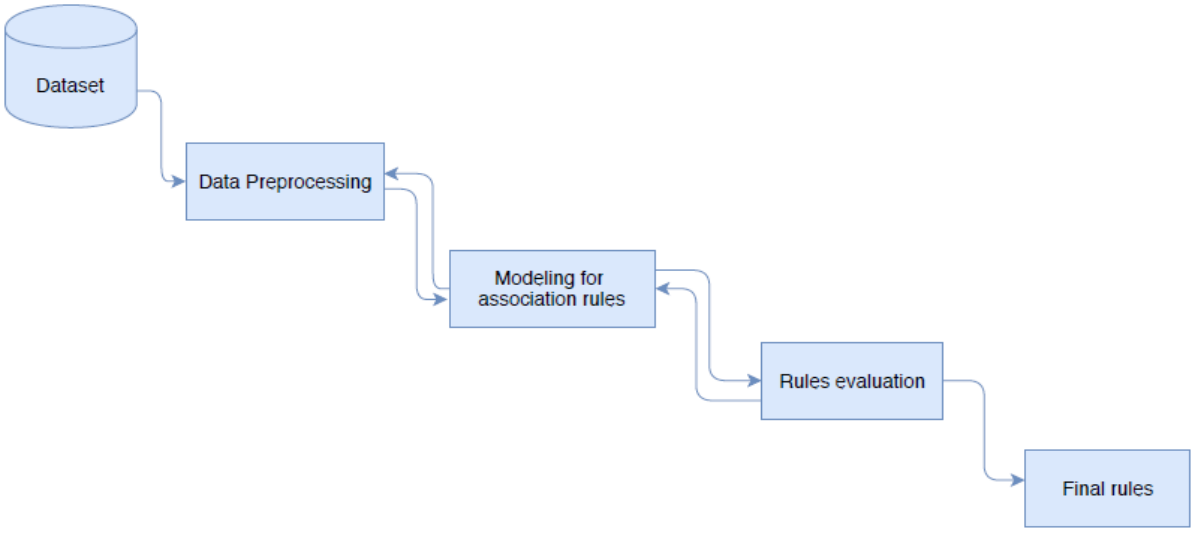


Figure 1: Proposed methodology of our work

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0117.v1



### 3.1 Dataset

The dataset consists of medicinal histories of 324 patients gathered from the University of California (Irvine, CA, USA) machine learning database [29]. There were 96 mesothelioma patients as well as 228 healthy individuals which indicated imbalanced dataset [30]. Regarding with imbalanced dataset, it includes 29.63% patients of mesothelioma and 70.37% healthy individuals. The names and complete detail of features have presented in Table 1. The dataset included 11 Boolean, 14 continuous, 6 categorical, 3-time values features and class label.

TABLE 1: SOME UPDATED FEATURE NAMES TO ADD CLARITY: KEEP SIDE INTO the LUNG SIDE, CYTOLOGY INTO PLEURAL FLUID CYTOLOGY TEST, CELL COUNT INTO WHITE BLOOD CELL (WBC) COUNT, WHITE BLOOD INTO PLEURAL FLUID WBC COUNT, HEMOGLOBIN INTO HEMOGLOBIN NORMALITY TEST, SEDIMENTATION INTO ERYTHROCYTE SEDIMENTATION RATE, ALBUMIN INTO SERUM ALBUMIN AND GLUCOSE INTO BLOOD GLUCOSE

| **Feature name with unit** | **Meaning** |
|---|---|
| Age (years) | Age took in numerical form |
| Gender (category) | Male=1,Female=0 |
| City (category) | Diyarbakır=0, Elazıg=1, Binggl=2, Şanliurfa=3, Mardin=4, Bitlis=5, Siirt=6, Batman=7 |
| Asbestos exposure (Bool) | Yes=1,No=0 |
| Type of MM(category) | Pleural mesothelioma=0,Peritoneal mesothelioma=1, testicular mesothelioma=2 |
| Duration of asbestos exposure (years) | The duration of time in which asbestos exposed in patient |
| Diagnosis method(Bool) | Mesothelioma patient=1,non-mesothelioma patient=0 |
| Lung side(category) | Left side =0, Right side =1, both sides =2 |
| Pleural fluid cytology test (Bool) | Affected=1,clear =0 |
| Duration of symptoms (years) | The time duration of symptoms of individual patient |
| Dyspnoea (Bool) | Present=1, Not-Present =0 |
| Ache on chest (Bool) | Present=1, Not-Present =0 |
| Weakness (Bool) | Present=1, Not-Present =0 |
| Habit of cigarette(category) | Non-smoker=0, Rare smoker=1, Regular smoker=2, Frequent smoker=3 |
| Performance status (Bool) | Not able to perform everyday task=0, Able to perform everyday task=1 |
| White blood cell (WBC) count | The test measures the number of white blood cells and its quality |
| | |
| Pleural fluid WBC count (K/mcL) | This test measures leukocytosis count in a pleural fluid to detect infections and undiagnosed medical conditions |
| Hemoglobin normality test(Bool) | A test to check hemoglobin level is lower or normal |
| Platelet count (PLT) kilo platelets per mcL | A lab test to check quantity of platelets have in a blood |
| Erythrocyte sedimentation rate (mm/hr) | Blood test measures how quickly red blood cells settle in one hour in a test tube |
| Blood lactic dehydrogenase (LDH) (IU/L) | A test to check protein value that helps to produce energy in body |
| Alkaline phosphatase (ALP) (IU/L) | ALP test has used for detection of cancers that have spread to the bones |
| Total serum protein (g/dL) | A biochemical test to measure the total protein amount in serum |
| Serum albumin (g/dL) | This test measures the amount of protein albumin in the blood |
| Blood glucose (mg/dL) | A blood glucose test measures the glucose level in the blood |
| Pleural lactic dehydrogenase (IU/L) | Its levels indicate if the fluid is transudate or exudates |
| Pleural protein (g/L) | Classification of pleural effusions are based on the fluid protein level |
| Pleural albumin (g/dL) | The Albumin Level in the pleural fluid |
| Pleural glucose (mg/dL) | The value of pleural glucose linked with infection |
| Dead or not(Bool) | Dead=1,Alive=0 |
| Pleural effusion(Bool) | Present=1, Not-Present =0 |
| Pleural thickness on tomography(Bool) | Involved=1, Not-Involved=0 |
| Pleural level of acidity (pH) (Bool) | Yes=1, No=0 |
| C-reactive protein (CRP) (mg/L) | A blood test to detect inflammation in the body |
| Class of diagnosis (Bool) | Normal=1,Patients=2 |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0117.v1



3.2 Data Pre-processing

Data pre-processing is a crucial task in any data mining process as it has a direct impact on the success rate of results. This reduces the complexity of the data under analysis as data in real-world is unclean. Data pre-processing includes data cleaning, dimensionality reduction and transformation of data [31]. Step I: Data cleaning removes noise and inconsistent data by filtering, aggregating, and filling in missing values. The completeness of the dataset is exceptional quality in electronic medical records which enables to make a more exact and precise analysis than different conditions where a few values are missing [32]. The "diagnosis method" attribute has duplicate values to "class of diagnosis" [30] then attribute was removed from the data set.

Step II: Biomedical data is often high-dimensional, and not all variables are useful for a particular task such as association rule mining [33]. It is also necessary to select potential relevant risk factors and patient cases associated with a particular disease. This also helps in reducing the dimensionality of the dataset which makes it easier to discard the attributes which are not of much relevance to the disease. Some attributes were deleted from the data set due to irrelevancy with mesothelioma. The attributes "Type of MM", "Dead or not" and "Cytology" were also deleted from the dataset because these attributes were not associated with etiology of mesothelioma.

Step III: The other issue with the dataset was imbalance because patients were 29.63% and healthy was 70.37%. In a classification task, models see more healthy instances during training, they are better at predicting healthy instances during testing. Imbalanced data occurs when one (or more) of classes that are trying to label has many more instances than the other. Synthetic minority oversampling technique (SMOTE) was utilized to overcome the difficulty of imbalanced data. After SMOTE, the total numbers of instances are 456 where 228 are healthy and 228 are mesothelioma patients. For classification of the risk factors associated with the disease of mesothelioma, the 228 instances affected with mesothelioma are separated and healthy individuals are deleted.

Step IV: Data transformation is the process of converting data from one format to another for modeling. Association rule mining is only applicable when data attributes are categorical not continuous[34]. The attribute values in the dataset were continuous so these numerical values changed into nominal or categorical values. All Boolean and categorical values have also converted into nominal values. The attribute "Age" has converted into three categories: less than forty years, between forty to sixty-five years and greater than sixty years. The risk for malignant mesothelioma after first exposure is more than 40 years. "Duration of asbestos exposure" has also categorized into two classes: less than twenty years and more than twenty years. The other numerical attribute "duration of symptoms" has also converted into two classes, Less than ten years and greater than ten years which has also concluded in [35]. Normally the total WBC count for an adult ranges from 5,000 to 10,000 [36, 37]. White blood cell (WBC) count and pleural fluid WBC count were also divided into three categories: Less than 5000, 5000 to 10000 and more than 10000. The normal platelet count is widely quoted as 150–400×$10^9$/L in blood [38]. The attribute "platelet count (PLT)" has categorized into three classes: less than 150, between 150to 400 and above 400. The normal erythrocyte sedimentation rate is less than 42 mm/hr[39]. The attribute of "Erythrocyte sedimentation rate" has divided into two categories: less than 42 and above 42. Alkaline phosphatase (ALP) normal range is 40–160 [40]. ALP attribute contains only one category which has normal because all values in this attribute lied in the normal range. Alkaline phosphatase (ALP) has constant value so it made so the effect on the results so this attribute has also deleted. Glucose level between 70 and 100 is considered normal and the higher value is linked to a non-healthy condition of the patient [41]. The attribute "blood glucose" has divided into two classes: less than 100 and above 100. Particularly a normal pleural fluid glucose value above 60 mg/dl is not helpful; however, a low pleural fluid glucose level below 60 mg /dl will be helpful to narrow the exudative pleural effusion differential diagnosis [42, 43]. The attribute "pleural glucose" has divided into two categories: less than 60 and above 60. C - reactive protein (CRP) is an acute-phase protein, synthesized in response to various stimuli by the liver. Several cytokines which released in the inflammatory region triggered the induction of CRP synthesis. A neutrophilic exudate with pleural fluid C-reactive protein (CRP)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0117.v1



levels >45 mg/L will be most likely be parapneumonic and if pleural fluid CRP >100 mg/dL, its complicated parapneumonic effusion[42, 43]. The attribute "C-reactive protein" has categorized into three classes: less than 45, between 45 to 100 and greater than 45. Light's criteria identify exudates and transudates pleural effusion has many ways. The ratio of pleural fluid lactate dehydrogenase (LDH) to serum LDH is exudative effusion if its value is greater than 0.6. Pleural fluid protein divided by serum protein greater than 0.5 also indicates exudative effusion [44]. Serum albumin minus pleural fluid albumin is known as Albumin gradient. Using 1.2 g/dl or less Albumin gradient to indicate exudates and greater than 1.2 g/dl to indicate transudates [45, 46]. Some continuous attributes like Blood lactic dehydrogenise, Pleural lactic dehydrogenise, Total serum protein, Pleural protein, Serum albumin, and Pleural albumin has also transformed with the help of previous literature [44-46]. The visualization of pre-processed data has shown in figure 2.

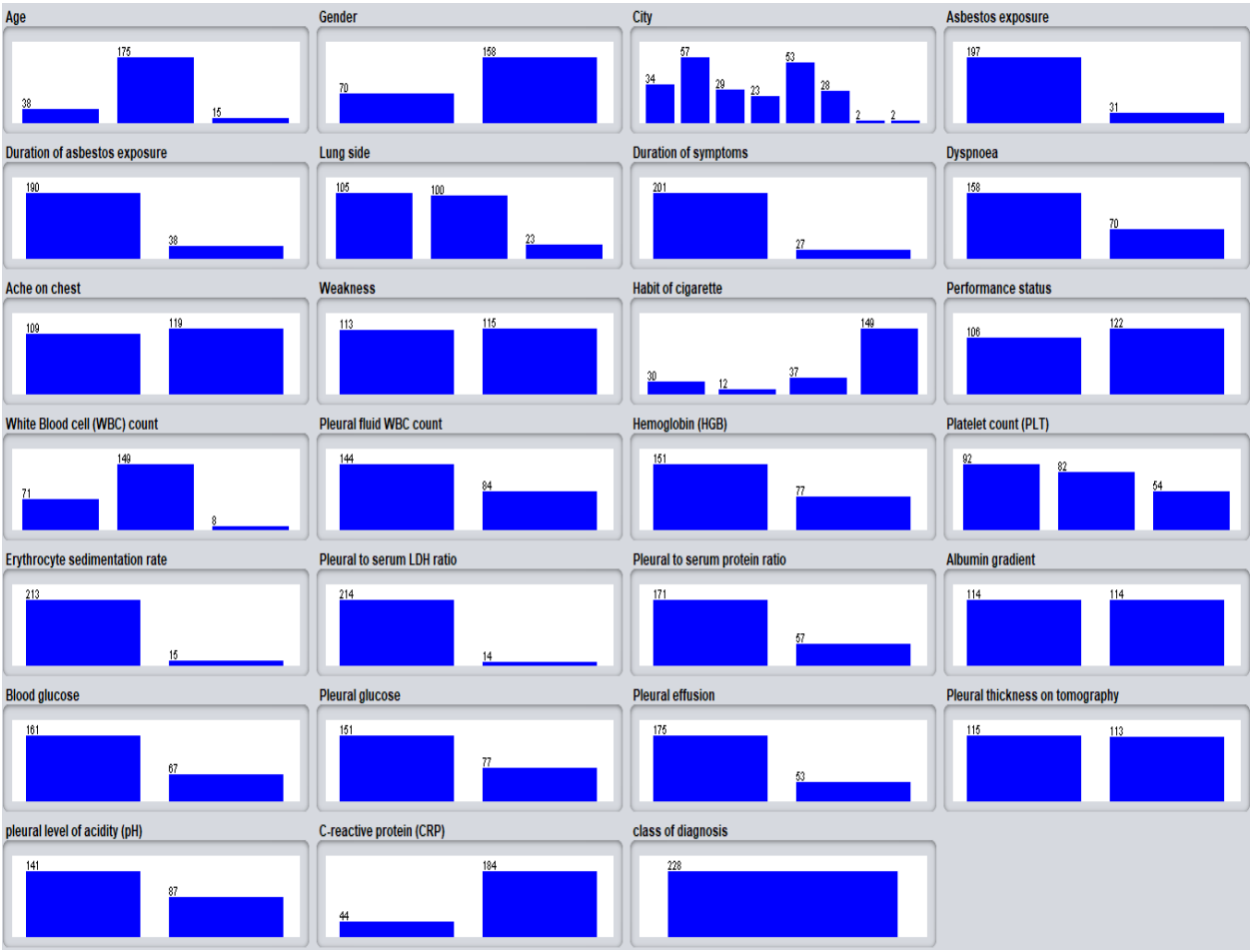


Figure 2: Visualization of pre-processed dataset

## 3.3 Modeling

### 3.3.1 Association Rule Mining

Association Rule Mining (ARM) is important and commonly used data mining technique. It was firstly introduced by Agrawal et al. [47] and useful for discovering hidden relationships in large data sets. It is popular for its use in recommender systems, promotional bundling, cross-selling and customer relationship management [48]. However, it has been utilized, extracting useful patterns from protein-protein interaction data set and protein function prediction[49]. Association rules help to uncover hidden relationships from large data sets. The goal of finding association rules is to extract co-occurrence patterns between the items of a data set. Association rules present statements in the form of "if…else" statements, where a set of items co-occurs with a separate set of items. First, the frequent itemsets are generated from candidate itemsets, and then rules are generated from these frequent

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0117.v1



itemsets[50]. Apriori algorithm has a widely used technique to find associations in biomedical data [34, 51, 52].

3.3.1.1 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 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. If $X, Y \subseteq I$ and $X \subseteq Y$ then the support of $Y$ can not be greater than the support of($supp(X) \geq supp(Y)$). This means that if an itemset $X$ is not frequent, than every superset of $X$ cannot 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 [50]. The Apriori algorithm was proposed in 1993 by Agrawal et al. [47]. 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.

o extract meaningful itemsets and rules, quality measures are used. The common quality measures are support and confidence. Fixed support (minsup) and confidence as threshold (minconf) are used to remove the uninteresting rules. Only the itemsets that have support greater or equal to minimum support are used as frequent item sets. Then, minimum confidence is used to filter the important rules, where rules having confidence greater than or equal to minimum confidence are considered [50].

Support: Support of an item set X is the probability that a randomly chosen transaction will contain X. For instance, an item-set X has the support a support of 0.1 means 10% of the transaction has all the items in X.

$$support = \frac{(X \cup Y).count}{n}$$

Confidence: Confidence of a rule X → Y is the conditional probability that a randomly chosen transaction will contain all the items in Y if the transaction contains all the items in X[34]. For instance, a rule X→ Y has the confidence of 0.9 means 90% of transactions that feature X also features Y.

$$confidence = \frac{(X \cup Y).count}{X.count}$$

IV. Results and Discussion

In this work, multiple analyses have performed on the dataset to discover the association rules. The dataset consists of healthy and mesothelioma patients but only mesothelioma patients were selected for the identification of symptoms. Most popular association rule mining algorithm known as Apriori has been used in this study. Results of the experiment have shown in table 2. The minimum threshold values of confidence and minsupport were 90% and 75%, were established. Many rules were obtained after experiment but only those rules having mesothelioma patient's class on the right-hand side (RHS) were considered. After going through the dataset, several rules were obtained based on minimum support and confidence level. Different researches utilized the estimations of support and confidence to find strong associations like in [53, 54], we change similar measures and show the estimations of support and confidence for each rule. As the features in the dataset were numerous, a limited number of these features have become beneficial in this study. Hence, new results have not been extracted, confirming the previous information. We saw that some rules were generated considered a common

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0117.v1



knowledge amongst the members of the medical fraternity, for example, the relation between mesothelioma and asbestos exposure was established [23] but still not accepted around the globe [25].

TABLE 2: RULES EXTRACTION BY USING APRIORI ALGORITHM

| Rule# | Association Rules | Support (%) | Confidence (%) |
|---|---|---|---|
| Rule#1 | Erythrocyte sedimentation rate=Above 42 AND Pleural to serum LDH ratio=Exudative ==> class of diagnosis=Patient | 87 | 100 |
| Rule#2 | Duration of symptoms=Less than ten years AND Erythrocyte sedimentation rate=Above 42 ==> class of diagnosis=Patient | 82.8 | 100 |
| Rule#3 | Asbestos exposure=Yes AND Duration of asbestos exposure=Above twenty years ==> class of diagnosis=Patient | 79.3 | 100 |
| Rule#4 | Asbestos exposure=Yes AND Duration of symptoms=Less than ten years ==> class of diagnosis=Patient | 77.1 | 100 |
| Rule#5 | Duration of asbestos exposure=Above twenty years AND Erythrocyte sedimentation rate=Above 42 ==> class of diagnosis=Patient | 77.1 | 100 |
| Rule#6 | Duration of symptoms=Less than ten years AND Erythrocyte sedimentation rate=Above 42 AND Pleural to serum LDH ratio=Exudative ==> class of diagnosis=Patient | 77.1 | 100 |
| Rule#7 | Duration of symptoms=Less than ten years AND Erythrocyte sedimentation rate=Above 42 AND Pleural to serum LDH ratio=Exudative ==> class of diagnosis=Patient | 77.1 | 100 |
| Rule#8 | Asbestos exposure=Yes AND Duration of asbestos exposure=Above twenty years AND Pleural to serum LDH ratio=Exudative ==> class of diagnosis=Patient | 76.3 | 100 |
| Rule#9 | Asbestos exposure=Yes AND Erythrocyte sedimentation rate=Above 42 AND Pleural to serum LDH ratio=Exudative ==> class of diagnosis=Patient | 76.3 | 100 |

Above association rules shows that asbestos exposure, duration of asbestos exposure, duration of symptoms, erythrocyte sedimentation rate and pleural to serum LDH ratio indicates the presence of mesothelioma. In the modern world, the huge growth of data in several fields has produced. Medical practitioners expect new computing methods to invent the knowledge contained in the database so that they can take a good decision and provide better services to the patients. With the increasing growth of bid data in medical, the manual analysis had become very arduous and led to the development of several automated data mining methods because useful information for medical practitioners is important. In cancer treatment, association rule mining through theApriori algorithm was utilized to explore the relationship between survival of the cancer patient and treatment preferences [34]. Other than cancer, Apriori algorithm was also utilized to find risk factors associated with diabetes [51] and heart disease [52]. Any rule with confidence greater or equal than 90%, can be beneficial in medical knowledge [55].

This study aim is to explore the risk factors of mesothelioma. All of the association rules extracted in this study revolve around the asbestos exposure, one of the major global risk factor of mesothelioma disease which has extracted from data. Support of most of the association rules is above 75% which inferring the combination of these conditions prevails amongst more than half of the patients present in our dataset. In our study, most of the extracted association rules depict the important underlying association with mesothelioma as consequent and risk factors as antecedent. The second most contributing factor was the duration of asbestos exposure which has also associated with asbestos exposure as antecedent. The latency period for mesothelioma is very long and it is also highly diverse in nature so it can range from 13 to 70 years [56]. The greater understanding of mesothelioma factors that determine the duration between exposure to asbestos and development of mesothelioma would help to predict the number of cases in future; it would also help to attribute new cases in past exposures and it would help in understanding the process of disease [35]. The long latency period of malignant mesothelioma means the disease usually develop after many years the asbestos exposure that cause it.

In most of the cases, the symptoms of mesothelioma were 20 to 30 years. The latency period depends on the duration of exposure and its intensity[57]. The people who suffered from heavier exposure like asbestos miners, the latency period lies between 12 and 20 years and even shorter latency periods are also possible [58]. There are some other diseases whose major cause is also asbestos exposure. Asbestos-related diseases other than mesothelioma are lung, larynx, and ovary cancer [59]. In clinics, ESR is the most extensively used biomarker for determining the inflammation in many diseases like autoimmune and malignant diseases. Cancer patients have elevated level of ESR. ESR is an evident prognostic factor badly affected the survival of cancer patients[60]. ESR is also a strong prognostic factor in cancers[61] like malignant mesothelioma.

The traditional criteria of classifying exudate were developed by Light et al 1972. Pleural fluid is categorized into exudates and transudates. When the balance of hydrostatic forces influences the formation and desorption of the pleural fluid then a transudative effusion occurs. However, an exudative pleural effusion develops when the pleural surface or the local capillary permeability is altered [44]. In addition to the quantitative changes in pleural fluid analyses in diseases causing an exudate formation, inflammatory changes of the pleural capillaries may result in a qualitative alteration in pleural fluid constituents [62]. The enzyme LDH is found in almost all the cells of the body. When cells are damaged or destroyed, LDH is released into the blood. A higher than normal level of LDH means that there is tissue or cell damage somewhere in the body. People with mesothelioma who have a higher than normal LDH level have a less favorable prognosis than people who have normal or low LDH levels [46]. Pleural to serum LDH ratio greater 0.6 indicates exudative effusion [62] which is also an important factor for the prognosis of malignant mesothelioma and its epidemiology. The identification of risk factors associated with mesothelioma may prevent patients from gone into the high danger of disease. The comorbidities associated with malignant mesothelioma are cardiovascular diseases [63],cancer-related emotional distress [64], pulmonary, neurologic, diabetes, anemia, psychiatric disorders, hypothyroidism, metastatic cancer, weight loss, and electrolyte disorders[65].

## V. Conclusion

Epidemiology of diseases at early stages assumes as a substantial measure for patient's appropriate treatment. While it is true that there is no sure way to prevent the majority of cancers; the best strategy is, based on the knowledge of preventive factors, to avoid the risk factors that can be managed and through following a healthy lifestyle as much as possible. In this study, a dataset of mesothelioma has been obtained from UCI repository and Apriori based association rule mining method has utilized. Computational intelligence-based technique has been used to find the significant risk factors associated with mesothelioma. It is observed that not a single factor is the cause of malignant mesothelioma, but it is the result of generally several factors. The results showed that asbestos exposure, duration of asbestos exposure and duration of symptoms having a considerable impact on the incidence of malignant mesothelioma. Further, erythrocyte sedimentation rate and Pleural to serum LDH ratio indicates also noted as important factors. These outcomes are expected to assist, through a systematic approach, researchers, medical practitioners and the general public in the diagnosis and assessment of the disease.

References

[1] U. Schlipköter and A. Flahault, "Communicable diseases: achievements and challenges for public health," *Public Health Reviews,* vol. 32, p. 90, 2010.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 68, pp. 394-424, 2018.

[3] N. Inagaki, K. Kibata, T. Tamaki, T. Shimizu, and S. Nomura, "Primary intrahepatic malignant mesothelioma with multiple lymphadenopathies due to non-tuberculous mycobacteria: A case report and review of the literature," *Oncology letters,* vol. 6, pp. 676-680, 2013.

[4] C. Couture, "Embryology, anatomy, and histology of the lung," in *Applied Respiratory Pathophysiology*, ed: CRC Press, 2017, pp. 1-14.

[5] B. W. Robinson, A. W. Musk, and R. A. Lake, "Malignant mesothelioma," *The Lancet*, vol. 366, pp. 397-408, 2005.

[6] V. Delgermaa, K. Takahashi, E.-K. Park, G. V. Le, T. Hara, and T. Sorahan, "Global mesothelioma deaths reported to the World Health Organization between 1994 and 2008," *Bulletin of the World Health Organization,* vol. 89, pp. 716-724, 2011.

[7] B. Price and A. Ware, "Time trend of mesothelioma incidence in the United States and projection of future cases: an update based on SEER data for 1973 through 2005," *Critical reviews in toxicology,* vol. 39, pp. 576-588, 2009.

[8] D. Marsili, B. Terracini, V. Santana, J. Ramos-Bonilla, R. Pasetto, A. Mazzeo, *et al.*, "Prevention of asbestos-related disease in countries currently using asbestos," *International journal of environmental research and public health,* vol. 13, p. 494, 2016.

[9] N. Halfon and M. Hochstein, "Life course health development: an integrated framework for developing health, policy, and research," *The Milbank Quarterly,* vol. 80, pp. 433-479, 2002.

[10] N. Jothi and W. Husain, "Data mining in healthcare–a review," *Procedia Computer Science,* vol. 72, pp. 306-313, 2015.

[11] J. Y. Chen and S. Lonardi, *Biological data mining*: CRC Press, 2009.

[12] M. J. Zaki, J. T.-L. Wang, and H. Toivonen, "BIOKDD01: workshop on Data Mining in Bioinformatics," *SIGKDD Explorations,* vol. 3, pp. 71-73, 2002.

[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 17, pp. 37-37, 1996.

[14] S. Wang, K. Ma, Z. Chen, X. Yang, F. Sun, Y. Jin, *et al.*, "A nomogram to predict prognosis in malignant pleural mesothelioma," *World journal of surgery,* vol. 42, pp. 2134-2142, 2018.

[15] R. R. Gill, B. Y. Yeap, R. Bueno, and W. G. Richards, "Quantitative clinical staging for patients with malignant pleural mesothelioma," *JNCI: Journal of the National Cancer Institute,* vol. 110, pp. 258-264, 2017.

[16] M. Chen, E. Helm, N. Joshi, F. Gleeson, and M. Brady, "Computer-aided volumetric assessment of malignant pleural mesothelioma on CT using a random walk-based method," *International journal of computer assisted radiology and surgery,* vol. 12, pp. 529-538, 2017.

[17] W. Brahim, M. Mestiri, N. Betrouni, and K. Hamrouni, "Semi-automated rib cage segmentation in CT images for mesothelioma detection," in *2016 International Image Processing, Applications and Systems (IPAS)*, 2016, pp. 1-6.

[18] M. Demir, H. Kaya, M. Taylan, A. Ekinci, S. Yılmaz, F. Teke, *et al.*, "Evaluation of new biomarkers in the prediction of malignant mesothelioma in subjects with environmental asbestos exposure," *Lung,* vol. 194, pp. 409-417, 2016.

[19] X. Hu and Z. Yu, "Diagnosis of mesothelioma with deep learning," *Oncology letters,* vol. 17, pp. 1483-1490, 2019.

[20] K. Tutuncu and O. Cataltas, "Diagnosis of Mesothelioma Disease Using Different Classification Techniques," *International Journal of Intelligent Systems and Applications in Engineering,* pp. 7-11, 2017.

[21] S. Mukherjee, "Malignant Mesothelioma Disease Diagnosis using Data Mining Techniques," *Applied Artificial Intelligence,* vol. 32, pp. 293-308, 2018.

[22] M. Nilashi, M. Z. Roudbaraki, and M. Farahmand, "A Predictive Method for Mesothelioma Disease Classification Using Naïve Bayes Classifier," *Journal of Soft Computing and Decision Support Systems,* vol. 4, pp. 7-14, 2017.

[23] L. Vimercati, D. Cavone, P. Lovreglio, L. De Maria, A. Caputi, G. M. Ferri, *et al.*, "Environmental asbestos exposure and mesothelioma cases in Bari, Apulia region, southern Italy: a national interest site for land reclamation," *Environmental Science and Pollution Research,* pp. 1-10, 2018.

[24] T. A. Dragani, F. Colombo, E. N. Pavlisko, and V. L. Roggli, "Malignant mesothelioma diagnosed at a younger age is associated with heavier asbestos exposure," *Carcinogenesis,* vol. 39, pp. 1151-1156, 2018.

[25] V. Muralidhar, P. Raghav, P. Das, and A. Goel, "A case from India of pleural malignant mesothelioma probably due to domestic and environmental asbestos exposure: a posthumous report," *BMJ Case Reports CP,* vol. 12, p. e227882, 2019.

[26] B. Jasani and A. Gibbs, "Mesothelioma not associated with asbestos exposure," *Archives of pathology & laboratory medicine,* vol. 136, pp. 262-267, 2012.

[27] S. T. Onur, S. N. Sokucu, L. Dalar, S. İliaz, K. Kara, S. Buyukkale, *et al.*, "Are neutrophil/lymphocyte ratio and platelet/lymphocyte ratio reliable parameters as prognostic indicators in malignant mesothelioma?," *Therapeutics and clinical risk management,* vol. 12, p. 651, 2016.

[28] A. Linton, M. Soeberg, R. Broome, S. Kao, and N. van Zandwijk, "Geographic and socioeconomic factors in patients with malignant pleural mesothelioma in N ew S outh W ales and their impact upon clinical outcomes," *Respirology,* vol. 22, pp. 978-985, 2017.

[29] O. Er, A. C. Tanrikulu, A. Abakay, and F. Temurtas, "An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease," *Computers & Electrical Engineering,* vol. 38, pp. 75-81, 2012.

[30] D. Chicco and C. Rovelli, "Computational prediction of diagnosis and feature selection on mesothelioma patient health records," *PloS one,* vol. 14, p. e0208737, 2019.

[31] D. Pyle, *Data preparation for data mining*: morgan kaufmann, 1999.

[32] B. K. Beaulieu-Jones, D. R. Lavage, J. W. Snyder, J. H. Moore, S. A. Pendergrass, and C. R. Bauer, "Characterizing and managing missing structured data in electronic health records: data analysis," *JMIR medical informatics,* vol. 6, p. e11, 2018.

[33] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics,* vol. 3, pp. 119-131, 2016.

[34] Q. Fan, C.-J. Zhu, J.-Y. Xiao, B.-H. Wang, L. Yin, X.-L. Xu, *et al.*, "An application of apriori algorithm in SEER breast cancer data," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, 2010, pp. 114-116.

[35] G. Frost, "The latency period of mesothelioma among a cohort of British asbestos workers (1978–2005)," *British journal of cancer,* vol. 109, p. 1965, 2013.

[36] E. L. GEORGE and A. Panos, "Does a high WBC count signal infection?," *Nursing2018,* vol. 35, pp. 20-21, 2005.

[37] V. C. Broaddus and R. W. Light, "Pleural effusion," in *Murray and Nadel's textbook of respiratory medicine*, ed: Elsevier, 2016, pp. 1396-1424. e10.

[38] M. F. Buckley, J. W. James, D. E. Brown, G. S. Whyte, M. G. Dean, C. N. Chesterman, *et al.*, "A novel approach to the assessment of variations in the human platelet count," *Thrombosis and haemostasis,* vol. 83, pp. 480-484, 2000.

[39] P. Elmes and M. J. Simpson, "The clinical aspects of mesothelioma," *QJM: An International Journal of Medicine,* vol. 45, pp. 427-449, 1976.

[40] M. W. Saif, D. Alexander, and C. M. Wicox, "Serum alkaline phosphatase level as a prognostic tool in colorectal cancer: a study of 105 patients," *The journal of applied research,* vol. 5, p. 88, 2005.

[41] E. F. Goljan, *Rapid Review Pathology E-Book*: Elsevier Health Sciences, 2018.

[42] M. J. Na, "Diagnostic tools of pleural effusion," *Tuberculosis and respiratory diseases,* vol. 76, pp. 199-210, 2014.

[43] J. M. Porcel and R. W. Light, "Diagnostic approach to pleural effusion in adults."

[44] R. W. Light, M. I. Macgregor, P. C. Luchsinger, and W. C. Ball, "Pleural effusions: the diagnostic separation of transudates and exudates," *Annals of internal medicine,* vol. 77, pp. 507-513, 1972.

[45] B. J. Roth, T. F. O'Meara, and W. H. Cragun, "The serum-effusion albumin gradient in the evaluation of pleural effusions," *Chest,* vol. 98, pp. 546-549, 1990.

[46] J. Joseph, P. Badrinath, G. S. Basran, and S. A. Sahn, "Is albumin gradient or fluid to serum albumin ratio better than the pleural fluid lactate dehydroginase in the diagnostic of separation of pleural effusion?," *BMC pulmonary medicine,* vol. 2, p. 1, 2002.

[47] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, 1993, pp. 207-216.

[48] S. Venkateswari and R. Suresh, "Association Rule Mining in E-commerce: A Survey."

[49] J. Wang, *Encyclopedia of data warehousing and mining*: iGi Global, 2005.

[50] C. Zhang and S. Zhang, *Association rule mining: models and algorithms*: Springer-Verlag, 2002.

[51] S. M. Kang and P. W. Wagacha, "Extracting diagnosis patterns in electronic medical records using association rule mining," *International Journal of Computer Applications,* vol. 108, 2014.

[52] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," *Expert Systems with Applications,* vol. 40, pp. 1086-1093, 2013.

[53] Y.-M. Tai and H.-W. Chiu, "Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan," *International journal of medical informatics,* vol. 78, pp. e75-e83, 2009.

[54] H. S. Kim, A. M. Shin, M. K. Kim, and Y. N. Kim, "Comorbidity study on type 2 diabetes mellitus using data mining," *The Korean journal of internal medicine,* vol. 27, p. 197, 2012.

[55] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," *IEEE Transactions on Information Technology in Biomedicine,* vol. 10, pp. 334-343, 2006.

[56] B. P. Lanphear and C. R. Buncher, "Latent period for malignant mesothelioma of occupational origin," *JOM,* vol. 34, pp. 718-721, 1992.

[57] I. Ahmed, S. A. Tipu, and S. Ishtiaq, "Malignant mesothelioma," *Pakistan journal of medical sciences,* vol. 29, p. 1433, 2013.

[58] S. B. Markowitz, "Occupational disease surveillance and reporting systems."

[59] K. Takahashi, P. J. Landrigan, and C. Ramazzini, "The global health dimensions of asbestos and asbestos-related diseases," *Annals of global health,* vol. 82, pp. 209-213, 2016.

[60] K. Bochen, A. Krasowska, S. Milaniuk, M. Kulczynska, A. Prystupa, and G. Dzida, "Erythrocyte sedimentation rate–an old marker with new applications," *Journal of Pre-clinical and Clinical Research,* vol. 5, 2011.

[61] F. Tas and K. Erturk, "Elevated erythrocyte sedimentation rate is associated with metastatic disease and worse survival in patients with cutaneous malignant melanoma," *Molecular and clinical oncology,* vol. 7, pp. 1142-1146, 2017.

[62] S. P. Chubb and R. A. Williams, "Biochemical Analysis of Pleural Fluid and Ascites," *The Clinical Biochemist Reviews,* vol. 39, p. 39, 2018.

[63] G. Zalcman, J. Mazieres, J. Margery, L. Greillier, C. Audigier-Valette, D. Moro-Sibilot, *et al.*, "Bevacizumab for newly diagnosed pleural mesothelioma in the Mesothelioma Avastin Cisplatin Pemetrexed Study (MAPS): a randomised, controlled, open-label, phase 3 trial," *The Lancet,* vol. 387, pp. 1405-1414, 2016.

[64] S. Lelorain, A. Cortot, V. Christophe, C. Pinçon, and Y. Gidron, "Physician Empathy Interacts with Breaking Bad News in Predicting Lung Cancer and Pleural Mesothelioma Patient Survival: Timing May Be Crucial," *Journal of clinical medicine,* vol. 7, p. 364, 2018.

[65] M. van Gerwen, A. Wolf, B. Liu, R. Flores, and E. Taioli, "Short-term outcomes of pleurectomy decortication and extrapleural pneumonectomy in mesothelioma," *Journal of surgical oncology,* vol. 118, pp. 1178-1187, 2018.