Data 630 9040

Machine Learning 2215

Professor Bati Firdu

Melissa Hunfalvay

Date: 6-8-2021

Assignment 1

Introduction

Objective

The dataset used for this project was the thoracic surgery data set from Wroclaw Thoracic Surgery Centre. Patients in this data set underwent major lung resections due to cancer in the lungs.

The objective of the analysis was to determine co-morbid symptoms for patients who had lung cancer and were required to undergo a lung resection. Understanding which symptoms to be co-morbid can help to improve patient care and medicine prescription (https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html).

The type of analysis is association rule learning, whereby associations, or sets of frequent items, between variables are explored to try and find insights and hidden relationships in large datasets (Han, Kamber, Pei, 2011). One example may be, that a person who smokes (lhs, antecedent), is frequently coupled with a symptom of cough and weakness (rhs, consequent).

Problem Domain

Lung cancer is a highly prevalent disease and is increasing. Lung cancer is the most common cancer in men worldwide with an age-standardized rate (ASR) of 33.8 per 100,000, and it is the fourth most frequent cancer in women (13.5 per 100,000; Ridge, McErlean, Ginsberg, 2013). The incidence and mortality attributed to lung cancer has risen steadily since the 1930s (Ridge, McErlean, Ginsberg, 2013).

There has been greater research and understanding on the causes of lung cancer (Ridge, McErlean, Ginsberg, 2013) which is important in educating the public to prevent the disease and to be diligent in recognizing early signs and symptoms for diagnosis.

When a person is diagnosed with lung cancer there are several options for treatment. However, one of the problems in deciding if surgery is the right option is understanding which patients are most likely to have success (ZiÄ, Tomczak, Lubicz, et al., 2013). When determining if surgery is an option for a patient, short- and long-term risks and benefits to the patient's mortality and quality of post-operative life are less well understood (Shapiro, et al., 2010; Aydogmus, et al., 2010; Icard, et al., 2013; Shahian & Edwards, 2008).

Therefore, the purpose of this analysis is to examine pre-operative symptoms that may be co-morbid for people who have undergone lung surgery due to lung cancer to determine if further insights or patterns can be understood to assist doctors, researchers, and medical educators.

Method Rationale

The methodology chosen is association mining via the Apriori algorithm. The rationale for this methodology is as a first step in understanding patterns in the dataset which may then lead to more targeted analysis and further questioning of the dataset. Association learning helps understand how the many variables, occur together which is difficult to do by simply looking at the dataset.

This methodology would be a beginning step in the analytic process. As this dataset is medical the most useful information involves causal and predictive analysis. Nevertheless, understanding relationships between symptoms may inform medical professionals and medical educators, in the following ways:

1. Determination of medications
2. Educating patients of common risk associated with behaviors such as smoking

3. When assessing patients, relationships between symptoms may assist in narrowing a diagnosis.
4. Medical researchers can use associations of symptoms for pre-screening inclusion/exclusion criteria

Analysis

Data

This data set was collected retrospectively between 2007-2011 at Wroclaw Thoracic Surgery Centre in association with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland (ZiA, Tomczak, Lubicz, et al., 2013). This the research database is a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

Patients in the dataset had primary lung cancer and a major lung resection. Cancer that begins in the lungs is called primary lung cancer (National Health System, 2019). A lung resection is a surgical procedure where all, or part, of the lung is removed (https://intermountainhealthcare.org/services/respiratory-care/treatment-and-detection-methods/lung-resection/).

Variables in the data set include some pre-operative symptomology such as pain levels (PRE7), haemoptysis (coughing up blood, PRE8), dyspnoea (difficulty breathing, PRE9), cough (PRE10) and weakness (PRE11). Overall performance stat is measured via the Zubrod scale. Zubrod or ECOG (Eastern Cooperative Oncology Group) scale. This scale ranges from 0 to 4, with 0 being fully functional and asymptomatic, and 4 being bedridden (West & Jin, 2015).

Information regarding the diagnosis (DGN) is also available including the International Classification of Diseases (ICD) – 10 (tenth revision) codes for primary (1), secondary (2) and multiple (3-8) tumors. Size of the tumor (PRE14) is included via a TMN code where T refers to the size and extent of the main tumor. Four levels are categorized whereby OC11 is the smallest tumors and OC14 is the largest.

Various medical conditions are also present in the dataset. These included diabetes (Type 2, PRE17), heart attack (Myocardial Infarction, MI, PRE19), Peripheral Arterial Disease (PAD, PRE25) and asthma (PRE32).

Medical state of the individual was measured prior to surgery to include Forced Vital Capacity (FVC, PRE4) which is the amount of air that can be forcibly exhaled from the lungs (https://www.verywellhealth.com/forced-expiratory-capacity) and is measured by a spirometer. Forced Expiration Volume (FEV) at the end of the first second of forced expiration, that is FEV1, (PRE5) is included in the data set. Medical state data is numeric and includes two decimal places.

Finally, age of the patient at the time of surgery (AGE) and survival 1 year post surgery (Risk1Yr) is also part of the data set. Original variable names, transformed variable names, definitions, types, and further details can be found in Appendix A. As the variables in the original dataset were numbered, for example PRE4, PRE5 an so on, they were changed (transformed) in the code to be more descriptive when trying to explain the results of the analysis.

Exploratory Analysis

There are 17 variables in the data set, as shown by the str function (Figure 1). The attributes in this dataset include both continuous data (such as age) and classification data (such as true and false for 1 year risk). Variable types include factors, which are discrete (as opposed to continuous) with pre-determined labels, such as PRZ which are three levels within the Zubrod scale.

```
> # Perform exploratory analysis
>
> # a) Provide basic description of the data
> str(ThoraricSurgerery)
'data.frame':   470 obs. of  17 variables:
 $ Diagnosis  : Factor w/ 7 levels "DGN1","DGN2",..: 2 3 3 3 3 3 3 2 3 3 ..
 $ FVC        : num  2.88 3.4 2.76 3.68 2.44 2.48 4.36 3.19 3.16 2.32 ...
 $ FEV1       : num  2.16 1.88 2.08 3.04 0.96 1.88 3.28 2.5 2.64 2.16 ...
 $ Zubrod     : Factor w/ 3 levels "PRZ0","PRZ1",..: 2 1 2 1 3 2 2 2 3 2 ..
 $ Pain       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Haemoptysis: logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
 $ Dyspnoea   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Cough      : logi  TRUE FALSE TRUE FALSE TRUE TRUE ...
 $ Weakness   : logi  TRUE FALSE FALSE FALSE TRUE FALSE ...
 $ Tumor_Size : Factor w/ 4 levels "OC11","OC12",..: 4 2 1 1 1 1 2 1 1 1 ..
 $ Diabetes   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Heart_A    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ PAD        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Smoking    : logi  TRUE TRUE TRUE FALSE TRUE FALSE ...
 $ Asthma     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ AGE        : int  60 51 59 54 73 51 59 66 68 54 ...
 $ Risk1Yr    : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
```

Figure 1: str function output

The data set also has “num” variables, which are real numbers, that have a value of a continuous quantity, that can represent a distance along a line (or alternatively, a quantity that can be represented as an infinite decimal expansion; Feferman, 1989). An example of a num variable in the data set is Forced Vital Capacity.

The final type of variable is a logi variable, which has only two logical outcomes. In this data set these are true/False and are found in multiple variables, such as pain, or cough or weakness before surgery.

There are 470 rows (or observations) of data. Each row represents one unique patient.

Using the summary command, and various visualizations, including histograms, pie charts, box plots, bar charts, various observations of the data were found (Figure 2).

```
> # b) Descriptive Statistics
> # Run the summary command to display the descriptive statistics for all variables
> summary (ThoraricSurgerery)
Diagnosis       FVC              FEV1           Zubrod       Pain            Haemoptysis     Dyspnoea        Cough           Weakness        Tumor_Size  Diabetes
DGN1:  1   Min.   :1.440   Min.    : 0.960   PRZ0:130   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   OC11:177   Mode :logica
DGN2: 52   1st Qu.:2.600   1st Qu.: 1.960   PRZ1:313   FALSE:439       FALSE:402       FALSE:439       FALSE:147       FALSE:392       OC12:257   FALSE:435
DGN3:349   Median :3.160   Median : 2.400   PRZ2: 27   TRUE :31        TRUE :68        TRUE :31        TRUE :323       TRUE :78        OC13: 19   TRUE :35
DGN4: 47   Mean   :3.282   Mean   : 4.569                                                                                              OC14: 17
DGN5: 15   3rd Qu.:3.808   3rd Qu.: 3.080
DGN6:  4   Max.   :6.300   Max.    :86.300
DGN8:  2
 Heart_A          PAD             Smoking          Asthma              AGE           Risk1Yr
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Min.   :21.00   Mode :logical
FALSE:468       FALSE:462       FALSE:84        FALSE:468       1st Qu.:57.00   FALSE:400
TRUE :2         TRUE :8         TRUE :386       TRUE :2         Median :62.00   TRUE :70
                                                                Mean   :62.53
                                                                3rd Qu.:69.00
                                                                Max.   :87.00
```

Figure 2: Summary Command showing descriptive statistics for all variables in the data set.

For number and integer variables (AGE, FEV1 and FVC) standard deviation was used to help further explore both distribution of the data, skewness, and possible later decisions for groupings during discretization. To visualize the numeric data histograms and boxplots were used.

Missing data values were checked. None were found in this data set.

From these summary and standard deviations, we can suspect that FVC is normally distributed as the mean and median are close together, 3.28 and 3.16 respectively, and the standard deviation is low (*SD* = 0.87). This is confirmed when viewing the histogram (Figure 3). Results for the age variable were like those of the FVC. Further exploration of the FVC variable using the boxplot visualization shows there are also some outliers in the data (Figure 4).

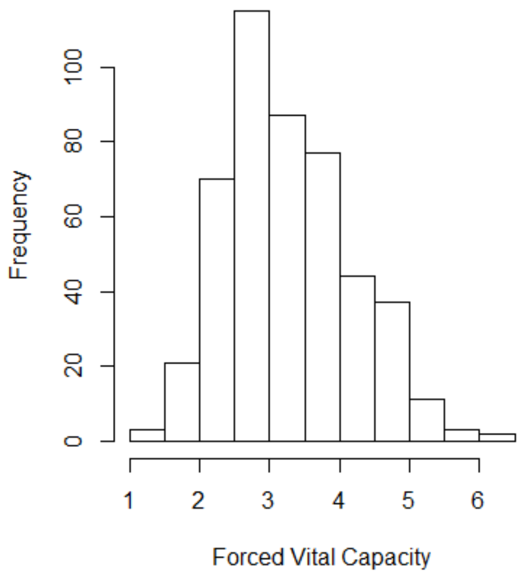


Figure 3: Distribution of the Forced Vital Capacity

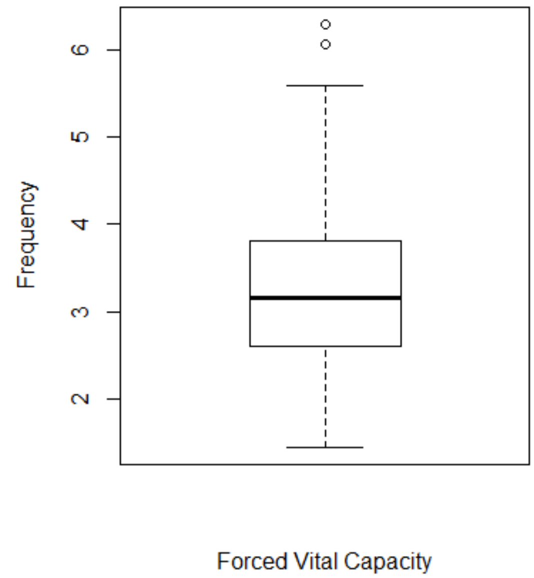


Figure 4: Box plot of Forced Vital Capacity

Similar analysis for the FEV1 show very different outcomes to Age and FVC. The range in the FEV1 variable is very large (minimum = 0.96, maximum = 86.30, range = 85.34). The standard deviation is also very large (*SD* = 11.77). When visualizing the data via a histogram we can see most of the data falls within 0-10 (Figure 5).

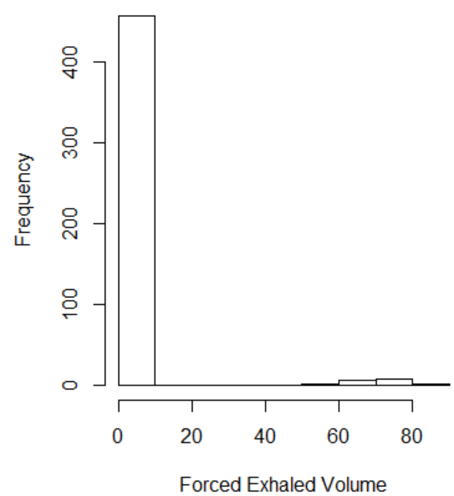
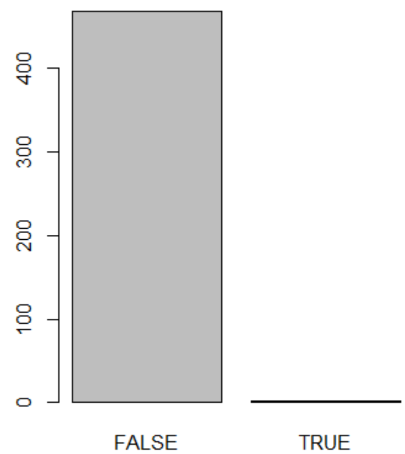


Figure 5: Distribution of the Volume Exhaled (FEV1)

Logi variables were explored via the summary command, and percentages. Visualizations included bar plots and pie charts (see Figure 6 & 7). It can be observed that PAD, Heart_A and

Asthma all have highly skewed data showing mostly false values. Percentages of false values in all three categories were greater than 98%.



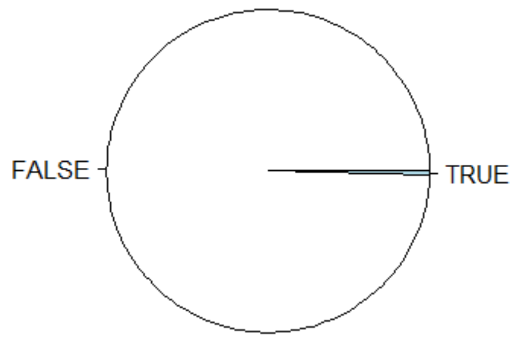


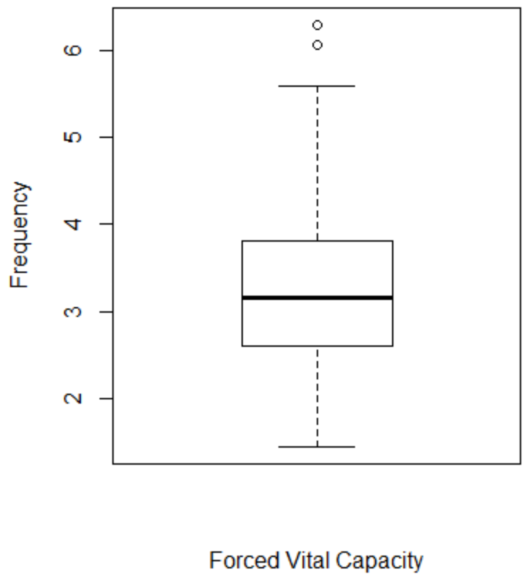
Figure 6: Bar chart; Figure 7: Pie Chart of Asthma variable

Factor variables were explored via the summary command and visualized via pie charts and bar plots. Observations showed that most patients had a diagnostic code of 3 (DGN3 = 349) with smaller sized tumors (OC11 = 177, OC12 = 257).

Preprocessing

Armed with the objective and exploratory analysis, as well as the model type preprocessing was conducted. This involved removal of outliers, removal of some variables and discretization of numeric variables.

Outlier Removal. Outliers were removed from all numeric variables. Outliers were defined as those falling outside the 25th and 75th percentiles (Figure 8 & 9).



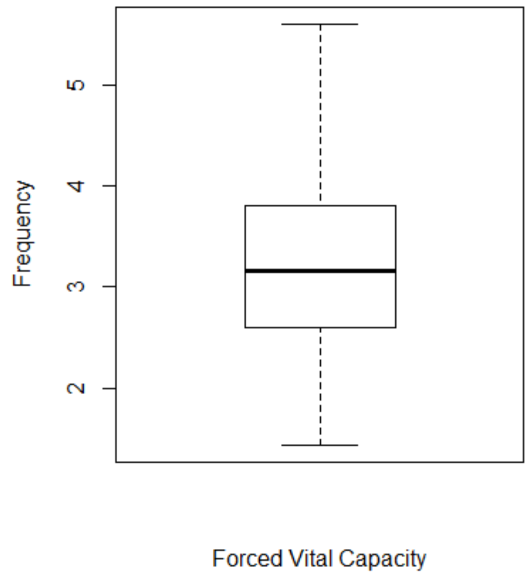


Figure 8: Box plot for FVC with outliers. Figure 9: Box plot for FVC after outlier removal

Removal of variables. Heart_A, PAD, and Asthma variables were removed because the data was not diverse enough to provide additional insights into the results.

Discretizing Variables. All numeric variables were discretized and made into factors to be used in the model analysis. Groupings were determined for age based on the distribution across decades. Groupings for FEV1 and FVC were used via standard deviations on the cleaned data (that is after outliers were removed the standard deviation was rerun and the new standard deviation was used for the grouping determination).

Algorithm Intuition

An Apriori algorithm was used to generate the model. The Apriori principle is based on how item sets are generated. Apriori algorithm is based on the frequency (or count) of the item set and the principle is “If an itemset is frequent, then all of its subsets are frequent (Han, Kamber, Pei, 2011).” This is important because it improves the performance and efficiency of the model by reducing the number of transactions needed to form item sets. The algorithm intuits that if the itemset is frequent then it does not need to go and determine if all similar (subset) items are also frequent.

The Apriori algorithm is based on an association rule, that is, an implication of an expression (Han, Kamber, Pei, 2011). Together, if the frequency of variables occur often, then they form a itemset, for example, smoking, cough and pain *may*, logically form an item set.

The logic behind the formula is in the form of an antecedent (rhs) and consequence (lhs). The antecedent is the "before" part of the equation, essentially stating the "if" this. The consequence is the "after," or other half of the statement and says, "then the probability of that." For example, *if* a large tumor (e.g. OC14) *then* there is some probability that you were also smoking and coughing. We could assume the size of a tumor the antecedent, could adversely affect symptoms such as pain (the consequent). However, we cannot infer any causality or prediction from a Apriori model. We can only say that a group of variables frequently occur together.

When put together, the antecedent and consequence form the Apriori algorithm and an association rule. The association uncovers relationships, represented by the frequency of item sets occurring together. The relationships can be expressed as such:

Antecedent (lhs)   Consequent (rhs)

X -> *Y*; where X and Y are item sets.

Associations are then generated as rules. The strength of the association is measured in terms of support and confidence and form part of the key input parameters for the Apriori model.

Support is how often the rule is applied to a specific data set. Whereas confidence determines how frequently items in Y appear in transactions with X (Han, Kamber, Pei, 2011).

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N};$$
$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

U = union, σ = count of the number of transactions where this union exists.

Length is also a key input parameter. Length of the rule is the number of items in the left-hand side (lhs) plus those on the right-hand side (rhs).

Rule length = lhs + rhs

Lift is another important metric is deciphering the relationship rule. Lift is the proportion of the rows of data that meet the condition on both sides of the equation (X and Y). The higher the lift the stronger the relationship between the lhs and rhs. A lift of 1 means the antecedent and consequent are independent of one another.

Rules Generation

The key steps used to generate rules were:

Step 1: Run the Apriori algorithm with the default arguments (Figure 10).

Default parameters are confidence of 80%, support of 10%. Minimum length of 1 and maximum length of 10. Target association mined were rules. Appearance, by default shows all itemsets. The default algorithm resulted in 113 rules.

```
> # Rules and Model Generation
> # Run the method with default parameters
> rules<-apriori(eliminated)
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       5     0.1      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 46

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[36 item(s), 463 transaction(s)] done [0.00s].
sorting and recoding items ... [20 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [113 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Figure 10: Apriori algorithm with default values

Step 2: Inspect the rules generated from the algorithm (Figure 11). Inspecting the rules it can be seen that the first rule has a maximum length of 1. As we are looking for associations, we need to prune this parameter. Furthermore, decisions on how to reduce the number of rules to make sense of the output in relationship to the objective of providing meaning associations to medical professionals is needed.

```
> inspect(rules[1:15])
     lhs                   rhs                 support   confidence coverage  lift      count
[1]  {}                 => {Smoking}         0.8272138 0.8272138  1.0000000 1.0000000 383
[2]  {FVC=-2SD}         => {Diagnosis=DGN3} 0.1209503 0.8235294  0.1468683 1.1052003  56
[3]  {FVC=+2SD}         => {Smoking}         0.1231102 0.8382353  0.1468683 1.0133236  57
[4]  {FEV1=-2SD}        => {Smoking}         0.1317495 0.8591549  0.1533477 1.0386129  61
[5]  {Risk1Yr}          => {Smoking}         0.1317495 0.8970588  0.1468683 1.0844340  61
[6]  {Weakness}         => {Cough}           0.1490281 0.9078947  0.1641469 1.3136102  69
[7]  {Weakness}         => {Smoking}         0.1511879 0.9210526  0.1641469 1.1134396  70
[8]  {AGE=70_79}        => {Cough}           0.1555076 0.8089888  0.1922246 1.1705056  72
[9]  {AGE=70_79}        => {Smoking}         0.1598272 0.8314607  0.1922246 1.0051339  74
[10] {FVC=+1SD}         => {Smoking}         0.2159827 0.8264463  0.2613391 0.9990721 100
[11] {FEV1=+1SD}        => {Smoking}         0.2246220 0.8125000  0.2764579 0.9822128 104
[12] {Tumor_Size=OC11} => {Smoking}         0.3066955 0.8208092  0.3736501 0.9922577 142
[13] {FEV1=-1SD}        => {Smoking}         0.3066955 0.8304094  0.3693305 1.0038630 142
[14] {AGE=50_59}        => {Diagnosis=DGN3} 0.3023758 0.8000000  0.3779698 1.0736232 140
[15] {AGE=50_59}        => {Smoking}         0.3174946 0.8400000  0.3779698 1.0154569 147
> |
```

Figure 11: First inspection of the data from default Apriori parameters

Step 3: Show a summary of the rules (Figure 12). As the dataset is medical in nature, accuracy is an important (highly weighted) metric. Therefore, in viewing the summary data we can see the maximum confidence is 94%. This can guide further iterative steps, specifically by starting with the highest confidence and iterating at 1% increments below the maximum threshold. This approach was taken iteratively with steps 4-9, until a threshold of 90% confidence was settled upon.

```
> #Run the summary command on rules
> summary(rules)
set of 113 rules

rule length distribution (lhs + rhs):sizes
 1  2  3  4  5 
 1 19 50 35  8 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   3.000   3.265   4.000   5.000 

summary of quality measures:
    support         confidence        coverage           lift            count       
 Min.   :0.1015   Min.   :0.8000   Min.   :0.1102   Min.   :0.9769   Min.   : 47.00  
 1st Qu.:0.1188   1st Qu.:0.8289   1st Qu.:0.1404   1st Qu.:1.0224   1st Qu.: 55.00  
 Median :0.1404   Median :0.8507   Median :0.1641   Median :1.0667   Median : 65.00  
 Mean   :0.1836   Mean   :0.8598   Mean   :0.2151   Mean   :1.1002   Mean   : 85.01  
 3rd Qu.:0.2030   3rd Qu.:0.8868   3rd Qu.:0.2462   3rd Qu.:1.1134   3rd Qu.: 94.00  
 Max.   :0.8272   Max.   :0.9444   Max.   :1.0000   Max.   :1.9861   Max.   :383.00  

mining info:
       data ntransactions support confidence
 eliminated           463     0.1        0.8
> 
```

Figure 12: Summary of rules from default Apriori parameters

Step 4: Prune the rules by only showing rules where there is an antecedent and consequence as we are looking for associations between data sets, that is run the minimum length option. Inspect the rules again.

Step 5: Sort the rules by lift, in descending order, for ease of inspection

Step 6: Iterate over the support and confidence values based on the question being asked. Inspect results. Iterate further. Slice and dice the data by inspecting subsets such as changing the support parameters.

Step 7: Pull the rules with support over 10%

Step 8: Remove any redundant rules

Step 9: Pick a target variable to inspect. Risk1Yr. Zero rules resulted.

Step 10: Generate rules for a specific itemset on the rhs. Smoking True or False was generated. All rules with smoking generated smoking=TRUE

Step 11: Run addition metrics on the rules. These included chi square, conviction, cosine, coverage, leverage, and odds ratio (Figure 13). This will help with interpretation of the rules (see Appendix C).

```
> interestMeasure(rules, c("chiSquare", "conviction", "cosine", "coverage", "leverage", "oddsRatio"), eliminated)
  chiSquared conviction    cosine  coverage   leverage oddsRatio
1   5.601725   2.188625 0.4102909 0.1641469 0.01540335  2.758253
2   7.888282   1.752546 0.5488661 0.3066955 0.02275516  2.366387
3   4.513655   1.766259 0.4421654 0.1987041 0.01489488  2.182593
4   4.317540   1.814255 0.4237157 0.1814255 0.01406920  2.228013
5   5.123488   1.999383 0.4201359 0.1749460 0.01510946  2.497457
```

Figure 13: Further exploration of rules with addition metrics

Step 12: Do a final summary and inspection of rules. Specifically view rules in relation to the objective of the analysis and within the perspective of the (medical) dataset (see Appendix B for all final output). Five rules were found at the end of the rules generation process (see Figure 14).

```
> inspect(rules[1:5])
    lhs                                   rhs       support   confidence coverage  lift     count
[1] {Weakness}                         => {Smoking} 0.1511879 0.9210526  0.1641469 1.113440  70
[2] {FVC=-1SD,Cough}                   => {Smoking} 0.2764579 0.9014085  0.3066955 1.089692 128
[3] {Diagnosis=DGN3,FEV1=-1SD,Cough}   => {Smoking} 0.1792657 0.9021739  0.1987041 1.090618  83
[4] {Diagnosis=DGN3,Cough,AGE=60_69}   => {Smoking} 0.1641469 0.9047619  0.1814255 1.093746  76
[5] {FVC=-1SD,Cough,Tumor_Size=OC12}   => {Smoking} 0.1598272 0.9135802  0.1749460 1.104406  74
```

Figure 14: Final rules generated

Results

Output

The default parameters from the Apriori algorithm resulted in 113 rules. After initial pruning to ensure minimum length was obtained for each rule, 112 rules remained. After much iteration, confidence values were set to 0.90 and support to 0.15. Target variable Risk1Yr was set and inspected, resulting in zero rules. Rules for specific itemset of smoking on the lhs was generated. Rules were pruned for redundancies. Final results revealed five rules.

The objective of the analysis was to determine which symptoms to be co-morbid to improve patient care and medicine prescription (https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html).

Results revealed the following patterns:

Rule 1: weakness is an antecedent for smoking. These item sets occurred together 70 times with a confidence level of 92%, support was 15% and lift was 1.11.

Rule 2: Patients who fell one standard deviation below the FVC average and had a cough were an antecedent to smoking. These item sets occurred together 128 times with a confidence level of 90%, support was 27% and lift was 1.09.

Rule 3: Patients who had a diagnosis level of 3, who fell one standard deviation below the FVC average and had a cough were an antecedent to smoking. These item sets occurred together 83 times with a confidence level of 90%, support was 18% and lift was 1.09.

Rule 4: Patients who had a diagnosis level of 3 and had a cough and were aged between 60-69 years were an antecedent to smoking. These item sets occurred together 76 times with a confidence level of 91%, support was 16% and lift was 1.09.

Rule 5: Patients who fell one standard deviation below the FVC average and had a cough and a tumor the size of 12 were an antecedent to smoking. These item sets occurred together 74 times with a confidence level of 91%, support was 16% and lift was 1.10.

In summary, smoking was frequently associated with several poor health antecedents. This is important for medical educators as it shows an association between smoking and poor health in a population of people who needed lung re-sectioning from lung cancer. This information could provide a deterrent for young people who are considering smoking and for those who are smoking to consider giving up the habit. Therefore, the objective of understanding patterns, specifically co-morbid symptoms for lung re-sectioning was met.

Rules Properties

Functions used to summarize the rule properties were length and summary command (Figure 15).

Number of rules

Length

Count

```
> summary(rule
set of 5 rules

rule length distribution (lhs + rhs):sizes
2 3 4
1 1 3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     3.0     4.0     3.4     4.0     4.0

summary of quality measures:
    support          confidence        coverage           lift            count
 Min.   :0.1512   Min.   :0.9014   Min.   :0.1641   Min.   :1.090   Min.   : 70.0
 1st Qu.:0.1598   1st Qu.:0.9022   1st Qu.:0.1749   1st Qu.:1.091   1st Qu.: 74.0
 Median :0.1641   Median :0.9048   Median :0.1814   Median :1.094   Median : 76.0
 Mean   :0.1862   Mean   :0.9086   Mean   :0.2052   Mean   :1.098   Mean   : 86.2
 3rd Qu.:0.1793   3rd Qu.:0.9136   3rd Qu.:0.1987   3rd Qu.:1.104   3rd Qu.: 83.0
 Max.   :0.2765   Max.   :0.9211   Max.   :0.3067   Max.   :1.113   Max.   :128.0

mining info:
       data ntransactions support confidence
 eliminated           463    0.15        0.9
> |
```

Figure 15: Rules properties summarized

The Apriori method generated five rules. Rule length distribution shows that number of rules with each length. One rule has a length of two, another rule has a length of three and three rules have a length of four.

The summary of quality measures displays various statistical outputs of the five rules generated. Included are minimum, maximum (from which can be derived range). The 1st and 3rd quartiles are included as measures of variability. Measures of central tendency include mean and median. As confidence intervals were set to 90% the range is high. Confidence in the results is an important outcome for medical datasets.

The mining information at the end of the summary command output show the data set used, number of instances in the dataset, and the minimum parameters set for support (0.15) and confidence (0.90).

Evaluation

The highest lift (1.11), which also has the highest confidence value (92%) is visualized in Figure 16, at the top left corner of the graph. This is rule 1; weakness -> cough. Therefore, as this rule has two parameters that are high, it is considered the strongest rule and we would want to explore this further with additional metrics (seen in Figure 13 are interpreted in Appendix C).

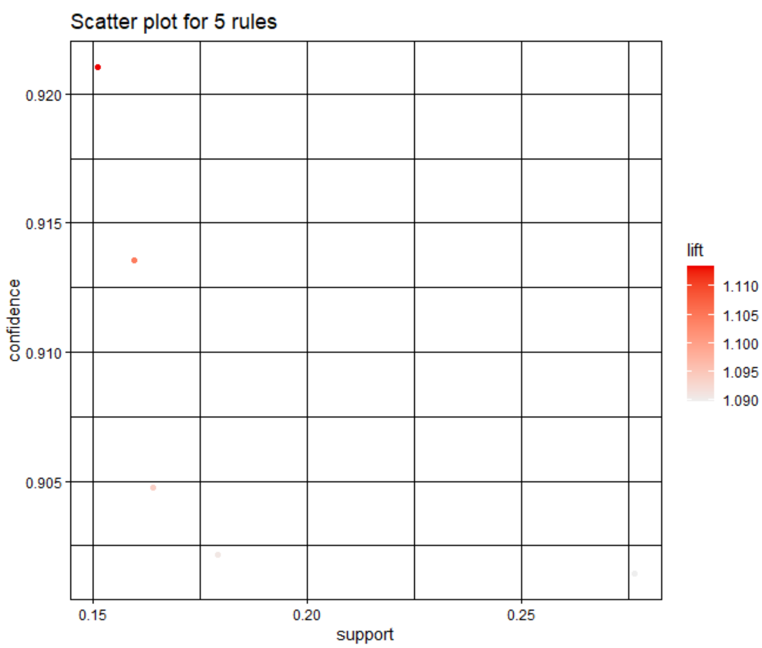


Figure 16: Scatterplot for rules generated to include three parameters, lift, confidence, and support.

Figure 17 shows the relationships in relation to one another and in what direction they occur. For instance, weakness is the antecedent (X) to smoking (Y). Also, FEV1 value in first standard deviation below the mean and a cough and a Diagnosis of DGN3 are antecedents to smoking. The red dot shows a high lift value. Black dot shows support.

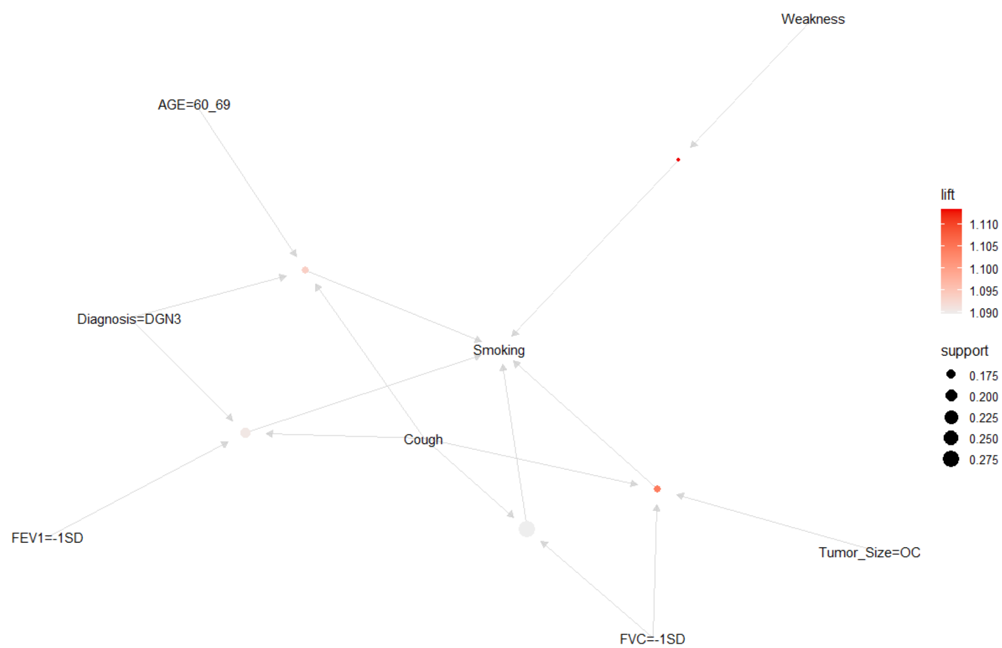


Figure 17: Directional relationship between variables and lift and support values

Figure 18 is a different way to show the rules by group rather than individual item (variable). This visualization reveals the two different outcomes (confidence and support) for each rule, via a matrix that is plotted against their strength on the lhs (i.e. the consequent of smoking).

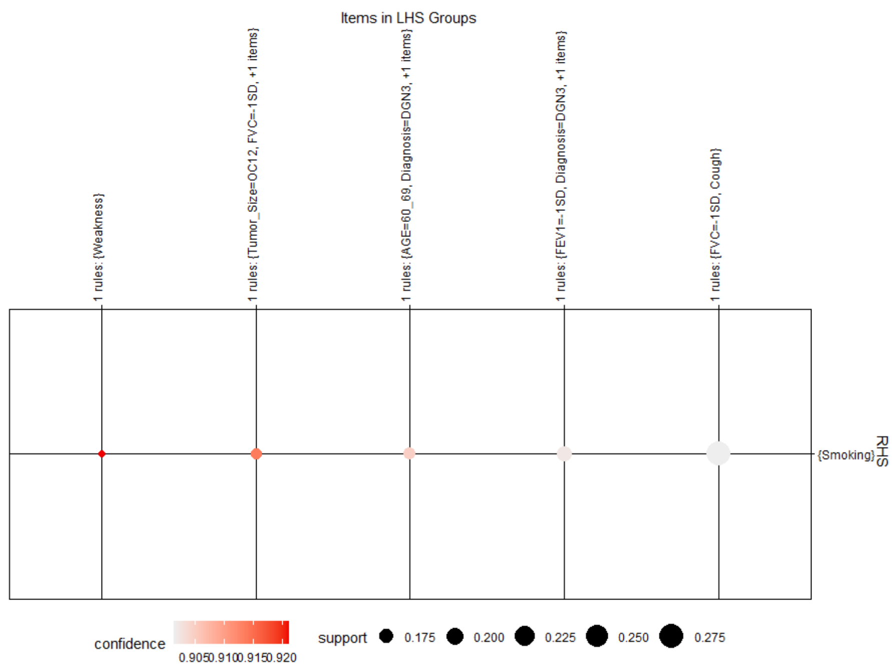


Figure 18: Grouped matrix showing rules by confidence and support

Additional metrics seen in Figure 15 are interpreted in Appendix D.

Conclusion

Summary

In summary, the main findings show five association rules. The consequent of which are smoking. The rules suggest that cough, weakness a diagnosis of 3, and FEV one standard deviation below the mean, an age of 60-69 and a tumor size of 12 have a frequent occurrence with smoking.

This is a strong finding that can assist in further follow-up analysis to determine if smoking is a cause of lung cancer. Most importantly, this may be used as a way for doctors and medical educators to speak with patients about such associations. More specifically, for doctors to be aware of the co-morbid symptoms of lung cancer patients in order to help with diagnosis, medications and medical awareness.

Limitations

The main limitation of this analysis is the type of analysis for the data set. In medical research the common questions are usually "What caused the illness?" and "How can it be prevented?" Neither of those questions can be answered by an Aprior algorithm (unfortunately). Nevertheless, the algorithm gives us "hints" on what to do next.

Other limitations include the data itself. For example, several variables were eliminated because they all had the same results (e.g. the Zubrod scale was eliminated as a variable because the majority of the results were PRZ1).

Categorical variables within the dataset could have been categorized in a more meaningful manner. For instance, size of the tumor can be categorized as small (OC11 and 12) and large (OC 13 and 14). This can help further reduce the levels within a factor and potentially provide more insights into the data. A similar case can be made for the different diagnosis levels whereby 1 tumor (DGN1), 2 tumors (DGN2) and multiple tumors (DGN3-8) could have been classified together.

Finally, the dataset had many pre-operative variables and only one post-operative variable (Risk1Yr). To make better use of such data it would have been interesting to see the same variables as post-operative measures. For instance, was cough and weakness still frequently occurring in people who were alive one-year post surgery?

Improvement Areas

Although the algorithm and principle of associative learning can be useful in some circumstances, I think it is only a very first step in this type of dataset. Future improvements would include additional causal and predictive analytics.

Improvement areas in the dataset include additional post-operative variables within the dataset, more meaningful classification of factor variables.

Appendix A

Variables Explained

Original Variable Name	Transformed Variable Name	Variable Definition	Variable Type	Details
DGN	Diagnosis	Diagnosis: specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any	Factor	7 levels DGN1, DGN3, DGN2, DGN4, DGN6, DGN5, DGN8
PRE4	FVC	Forced Vital Capacity: air that can be forcibly exhaled from the lungs	num	Number with two decimal places e.g. 2.88, 3.40
PRE5	FEV1	Volume that has been exhaled at the end of the first second of forced expiration	num	Number with two decimal places e.g. 2.16, 1.88
PRE6	Zubrod	Performance status - Zubrod scale	Factor	3 levels PRZ0, PRZ1, PRZ2
PRE7	Pain	Pain before surgery	logi	(T, F)
PRE8	Haemoptysis	Haemoptysis before surgery	logi	(T, F)
PRE9	Dyspnoea	Dyspnoea before surgery	logi	(T, F)
PRE10	Cough	Cough before surgery	logi	(T, F)
PRE11	Weakness	Weakness before surgery	logi	(T, F)
PRE14	Tumor_Size	T in clinical TNM. Size of the original tumor.11 = smallest. 14 = largest	Factor	4 levels OC11, OC14, OC12, OC13
PRE17	Diabetes	Type 2 DM: Diabetes Mellitus	logi	(T, F)
PRE19	Heart_A	Myocardial infarction (heart attack)	logi	(T, F)
PRE25	PAD	Peripheral Arterial Diseases	logi	(T, F)
PRE30	Smoking	Smoking	logi	(T, F)
PRE32	Asthma	Asthma	logi	(T, F)
AGE	Age	Age at surgery	int	Measured in whole numbers
Risk1Yr	Risk1Yr	1 year survival period	logi	(T)rue value if died (T, F)

Appendix B

Final Rule Output

```
> rules <- apriori(eliminated, parameter= list(minlen=2, supp=0.15, conf=0.90))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.9    0.1    1 none FALSE            TRUE       5    0.15      2     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 69

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[36 item(s), 463 transaction(s)] done [0.00s].
sorting and recoding items ... [14 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [5 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> # preview the rules
> inspect(rules[1:5])
    lhs                                   rhs       support   confidence coverage  lift     count
[1] {Weakness}                         => {Smoking} 0.1511879 0.9210526  0.1641469 1.113440  70
[2] {FVC=-1SD,Cough}                   => {Smoking} 0.2764579 0.9014085  0.3066955 1.089692 128
[3] {Diagnosis=DGN3,FEV1=-1SD,Cough}   => {Smoking} 0.1792657 0.9021739  0.1987041 1.090618  83
[4] {Diagnosis=DGN3,Cough,AGE=60_69}   => {Smoking} 0.1641469 0.9047619  0.1814255 1.093746  76
[5] {FVC=-1SD,Cough,Tumor_Size=OC12}   => {Smoking} 0.1598272 0.9135802  0.1749460 1.104406  74
> summary(rules)
set of 5 rules

rule length distribution (lhs + rhs):sizes
2 3 4
1 1 3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     3.0     4.0     3.4     4.0     4.0

summary of quality measures:
    support          confidence        coverage           lift           count
 Min.   :0.1512   Min.   :0.9014   Min.   :0.1641   Min.   :1.090   Min.   : 70.0
 1st Qu.:0.1598   1st Qu.:0.9022   1st Qu.:0.1749   1st Qu.:1.091   1st Qu.: 74.0
 Median :0.1641   Median :0.9048   Median :0.1814   Median :1.094   Median : 76.0
 Mean   :0.1862   Mean   :0.9086   Mean   :0.2052   Mean   :1.098   Mean   : 86.2
 3rd Qu.:0.1793   3rd Qu.:0.9136   3rd Qu.:0.1987   3rd Qu.:1.104   3rd Qu.: 83.0
 Max.   :0.2765   Max.   :0.9211   Max.   :0.3067   Max.   :1.113   Max.   :128.0

mining info:
       data ntransactions support confidence
 eliminated           463    0.15        0.9
>
```

Appendix C

Additional Metrics

Conviction is defined as the ratio of the expected frequency that the antecedent occurs without the consequent if consequent and antecedent were independent divided by the observed frequency of incorrect predictions. A high value means that the consequent depends strongly on the antecedent (https://rpubs.com/CFernandez/686167). Rule 1, weakness -> cough, reveals a high conviction value of 2.18.

Odds ratio is defined as the likelihood that the antecedent and consequent will occur, expressed as a proportion of the likelihood that they will not occur. Therefore, if X is the probability of subjects affected and Y is the probability of subjects not affected, then odds = X /Y (https://psychscenehub.com/psychpedia/odds-ratio-2/#:~:text=Odds%20of%20an%20event%20happening,then%20odds%20%3D%20A%20%2FB) The highest odds ratio is found in rule 1, weakness -> cough.

Appendix D

References

Aydogmus, U., Cansever, L., Sonmezoglu, Y., Karapinar, K., Kocaturk, C.I., .Bedirhan, M.A. (2010). The impact of the type of resection on survival in patients with n1non-small-cell lung cancers. *European Journal of Cardio-Thoracic Surgery 37*, p. 446–450.

Feferman, S. (1989). *The Number Systems: Foundations of Algebra and Analysis*, AMS Chelsea, ISBN 0-8218-2915-7.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Technique, 3rd Ed.* Ch.3. Elsevier Science. New York, NY.

Icard, P. Heyndrickx, M. Guetti, L. Galateau-Salle, F. Rosat, P. Le Rochais,J.P., Hanouz, J.-L. (2013). Morbidity, mortality and survival after 110 consecutive bilobec-tomies over 12 years, *Interactive Cardiovascular and Thoracic Surgery, 16,* 179–185.

Intermountain Healthcare. (2018). *Lung Resection.* Retrieved from: https://intermountainhealthcare.org/services/respiratory-care/treatment-and-detection-methods/lung-resection/.

Nuggets, K.D. (2021). *Association Rules and Aprior Algorithm.* Retrieved from: https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html.

National Health System (NHS). (2019). *Lung Cancer: Overview.* Retrieved from: https://www.nhs.uk/conditions/lung-cancer/#:~:text=Cancer%20that%20begins%20in%20the,forms%20of%20primary%20lung%20cancer.

Ridge, C. A., McErlean, A. M., & Ginsberg, M. S. (2013). Epidemiology of lung cancer. *Seminars in interventional radiology*, *30*(2), 93–98. https://doi.org/10.1055/s-0033-1342949

Shahian, D. Edwards, F. (2008). Statistical risk modeling and outcomes analysis, *Annals of Thoracic Surgery 86* (2008) 1717–1720.

Shapiro, M., Swanson, S.J., Wright, C.D., Chin, C., Sheng, D., Wisnivesky, J., & Weiser, T.S. (2010). Predictors of major morbidity and mortality after pneumonectomyutilizing the society for thoracic surgeons general thoracic surgery database, *Annals of Thoracic Surgery 90*, p. 927–935.

Very Well Health (2021). *What is Forced Vital Capacity?* Retrieved from: https://www.verywellhealth.com/forced-expiratory-capacity-measurement-914900#:~:text=Forced%20vital%20capacity%20(FVC)%20is,possible%2C%20as%20measured%20by%20spirometry

West, H., Jin, J.O. (2015). Performance Status in Patients With Cancer. *Journal of the American Medical Association: Annals of Oncology*; *1*(7):998. doi:10.1001/jamaoncol.2015.3113

ZiÄ, B.A., M., Tomczak, J. M., Lubicz, M., & ÅšwiÄ…tek, J. (2013). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing. 14*(PART A), 99-108. https://doi.org/10.1016/j.asoc.2013.07.016