EXPLORATORY ANALYSIS OF HUMAN SLEEP DATA

by

Parameshvyas Laxminarayan

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Masters of Science

in

Computer Science

January, 2004

APPROVED:

_____

Professor Carolina Ruiz, Thesis Advisor

_____

Dr. Majaz Moonis, Thesis Co-Advisor,
Director, Stroke Prevention Clinic, UMass Memorial Medical Center

_____

Professor Sergio A. Alvarez, Thesis Co-Advisor, Boston College

_____

Professor Matthew Ward, Thesis Reader

_____

Professor Michael A. Gennert, Head of Department

Abstract

In this thesis we develop data mining techniques to analyze sleep irregularities in humans. We investigate the effects of several demographic, behavioral and emotional factors on sleep progression and on patient's susceptibility to sleep-related and other disorders. Mining is performed over subjective and objective data collected from patients visiting the UMass Medical Center and the Day Kimball Hospital for treatment. Subjective data are obtained from patient responses to questions posed in a sleep questionnaire. Objective data comprise observations and clinical measurements recorded by sleep technicians using a suite of instruments together called polysomnogram. We create suitable filters to capture significant events within sleep epochs. We propose and employ a Window-based Association Rule Mining Algorithm to discover associations among sleep progression, pathology, demographics and other factors. This algorithm is a modified and extended version of the Set-and-Sequences Association Rule Mining Algorithm developed at WPI to support the mining of association rules from complex data types. We analyze both the medical as well as the statistical significance of the associations discovered by our algorithm. We also develop predictive classification models using logistic regression and compare the results with those obtained through association rule mining.

Acknowledgement

The opportunity to write these few lines would not have materialized without the blessings of Amma, Appa, Mrs. Sampada Kaji and the guidance and support of my advisors and friends. Words I put down here cannot truly express my gratitude to Prof. Carolina Ruiz for taking me under her wing and guiding me in every possible way. Many thanks professor, for taking time out week after week to discuss my work and review the endless drafts of my reports and presentations. Several neurologists and doctors shunned from providing us sensitive data on patient's visiting a sleep laboratory. I would like to express my deepest appreciation to Dr. Moonis for pushing to get the data released and agreeing to discuss with me issues at times in-between clinic appointments. My experimental results would not have been the same without the help and guidance of Prof. Sergio Alvarez. Many thanks Prof. Sergio for all your advice in logistic regression and obtaining of statistically significant association rules. I would also like to thank all my friends at WPI for their support and encouragement during my thesis. Special thanks to Abhishek, Mitesh, Aditya, Roshan, Nitin, Kalyan and Sidharth for literally going out of your way in helping me reach UMass Medical Center or Day Kimball Hospital on very short notices. Thanks guys!!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Domain overview

Though it is a very common behavior, sleep is an extremely complicated domain to study because of the large number and diversity of factors influencing it. Sleep is a dynamic phenomenon with the interplay between several factors continuously affecting it. This interplay has a direct influence on the mental and physical well- being of the individual and exhibits a high correlation in inducing sleep related disorders. A full understanding of sleep, therefore, requires disentangling the complex psychological, demographic and behavioral factors that have a direct bearing on an individual's sleep. Determination of the factors affecting sleep has come about after years of study by several behavioral scientists, psychologists, medical professionals and biomedical engineers. Sleep experts with the aid of the factors identified as important have successfully provided valid explanations to several of the phenomena noticed during sleep. However, there are several other interesting facets that if explored, will prove to be interesting and possibly useful to the sleep research community.

1.2 Need for computational analysis on collected data

Exploratory analysis in any domain requires accumulation of large amounts of data. The domain of sleep analysis is no exception to this. With the increase in the amount of data for analysis, use of computational techniques becomes possible. Today, due to the advent of powerful computational tools and models, the sleep domain is turning out to be a prime area of investigation in computational research in biological systems.

The initial sleep studies concentrated on finding the various emotional and physical traits of individuals that bear an influence on sleep. However, in the last three to four decades, due to rapid advances in electronic instrumentation and technology, reliable and accurate measurement of electronic signals along with storage and retrieval of large volumes of data has become particularly easy. This development has encouraged scientists to utilize the recordings made by several electronic devices to obtain a better understanding of the complex processes and transformations in vital body processes that occur during sleep. This procedure inevitably leads to the collection of large amounts of data that needs to be analyzed to provide any kind of informed judgment. This is where, computational techniques and algorithms used primarily in data mining and machine learning come into play.

1.3 Standardized procedure for sleep scoring

As mentioned earlier, the improved accuracy and interpretability of electronic instrument readings led to their use in numerous areas. Hence, several instruments for measuring physiological behavior were used by researchers to help interpret the various intricacies of sleep. Researchers were using recordings from several instruments to classify sleep into different stages. However, since no strict standard was being followed or maintained on the devices and the recordings that need to be analyzed, it became difficult to interpret the results produced by different researchers. A change was brought about after the Association for the Psychophysiological Study of Sleep (APSS) appointed a committee of sleep researchers to develop a standardized system to score sleep into different stages. Two sleep researchers, viz., Rechtschaffen and Kales (henceforth R&K) brought forth a model that took the help of the measured readings from a few electronic instruments as well as visual interpretation of features by technologists during the night under observation to classify sleep into different stages [RK68]. The model also called R&K model divides a normal sleep into six stages: five stages of sleep (sleep stages 1-4 and Rapid Eye Movement, or REM sleep) plus periods of wakefulness. Variations in the sleep stages lead to different sleep patterns. The suite of instruments whose combined tracings help in sleep stage classification are together called Polysomnogram.

Polysomnogram comprises instruments used to record the brain activity by measuring electric potentials in the brain using an Electroencephalograph (EEG), electric potential output caused by movement of body parts using an Electromyograph (EMG), and the electric potentials resulting from eye movements through an Electro-oculograph (EOG). In most cases, additional equipment to keep track of breathing and the amount of oxygen in the patient's bloodstream is also used. Every thirty-second of polysomnogram recording is classified into one of the six sleep stages based on the R&K classification model. The thirty second window of sleep that is classified is called 'epoch'.

1.4 Knowledge discovery using data mining

In this thesis we concentrate on applying a computational technique that is widely used in the field of Data mining to analyze sleep. Data mining is a branch of Knowledge Discovery in Databases (KDD), which elicits knowledge or information from data stored in large data repositories. It uses techniques that can help in extracting interesting and novel patterns from the data.

Data mining is a multidisciplinary field greatly influenced by a number of fields including artificial intelligence, machine learning, statistics, databases, visualization, and pattern recognition. The algorithms implemented in data mining utilize principles, proofs and deductions from all these fields. They are of great help in exploratory data analysis and in predictive classification type of problems. Deciding as to which data mining technique needs to be applied depends largely on the nature of the dataset, the type of problem being attempted and the interpretability of the results generated.

A technique often used in the field of data mining to perform exploratory data analysis is association rule mining. The idea behind the use of association rule mining is to be able to state with suitably high accuracy that presence of certain attribute value conditions lays the seed for presence or observance of certain other attribute conditions. For instance, moderate to high depression in patients complaining of excessive daytime sleepiness is indicative of very low sleep efficiency. In this case, the values for attributes,

viz., depression and excessive daytime sleepiness lay the seed for observing low sleep efficiency levels. Another motivating factor in using this approach was the fact that, there is very little research material available on past work done in mining associations over sleep data.

1.5 Context of this research

As part of the treatment of people suffering from sleep disorders, physicians evaluate their mental and physical state of health. Towards this end patients are required to fill in a questionnaire when they visit the sleep laboratory (refer to Appendix-A for the sample questionnaire). The questionnaire requires the patient to reveal among other information their demographic details, the nature of sleep disorders they suffer from, other disorders or diseases they have or are suffering from, their food habits, behavioral habits etc. Following this, the patient undertakes a sleep test.

Based on the initial assessment, there are different sleep studies that the physician may recommend. The more common laboratory studies are the all-night polysomnogram study and split-night study. Normally, the all-night polysomnogram study is recommended for patients who have a suspected sleep disorder. For patients suffering from or suspected of suffering from sleep apnea (most common sleep disorder), a Split-night study is generally recommended. During the night of the test, the technicians monitor the patient's progress into sleep. In the event of severe apneaic episodes, they put the patient under CPAP (continuous positive airway pressure). Following the night under observation, sleep stage classification as per the R&K model is performed using the polysomnogram readings and the technician's interpretations of the patient's sleep. The technician also prepares a report summarizing the patient's sleep trends during the night of study. This summary report aids the physician in making a diagnosis.

The dataset used for analysis comprises subjective and objective information collected from the epoch-by-epoch polysomnogram recordings, questionnaire responses and the summary report created by technicians for every patient.

1.6 Problem statement

Our motivation to perform this analysis is dual fold. Firstly, we aim at capturing relationships of significance among the several factors considered. We are more interested in finding if a particular demographic attribute or factor correlates highly with a certain sleep behavior or disorder. For example, patients having high Body-mass index (i.e., patients who are overweight or obese) are more prone to snoring while asleep. This is pretty obvious pattern noticed in patients whose Body-mass index exceeds the normal range. A more interesting result though is ascertaining that increasing levels of depression do not influence Periodic Leg Movements (PLMS), a very common sleep disorder. In other words, the degree of depression a patient suffers from cannot be used as a marker for diagnosing PLMS. Secondly, we are also interested in obtaining relationships linking sleep pattern variations with susceptibility of patients to suffer from disorders or diseases. To achieve these twin objectives we rely on association rule mining.

1.7 Thesis contribution

Currently, the most commonly used approach for analyzing data from medical domains is the use of statistical techniques. One of the major contributions of this thesis is to introduce and demonstrate by way of experimental results that techniques from the field of data mining are richer in the quantum of information that they provide following data analysis in comparison with statistical techniques that are currently in vogue. Also we highlight the advantages of association rule mining over statistical techniques like logistic regression when it comes to exploratory data analysis. This thesis also demonstrates different ways of data organization to counter the non-homogeneity of data and perform analysis to extract useful information.

The traditional association rule mining algorithm [AS94] can only handle data that can be represented in categorical or numeric forms. Former and current students at WPI have worked on extending this system to support more complex data formats such

as sets [Shoe01, Stoe02] and time sequences [Pray02]. In this thesis, we conduct experiments by representing data as both time sequences and sets by writing special filters to capture events of interest that occur during the different epochs of the patient's sleep. We also describe the window-based method of data transformation and organization that transforms the raw time-sequence attribute values to a form where attribute values with real-time information and occurring within the same window are placed in the same set. This system helps in providing more information as part of the rules. This aids the domain expert to obtain greater insight into the patient's sleep progression, which in turn helps in making a more accurate or reliable prognosis.

1.8 Application areas

The initial impetus to initiating this project was to develop systems that would find use in the area of Untethered healthcare and Patient monitoring. We believe that by including a diverse set of patients suffering from other disorders or pathological conditions and with more research it is possible to develop early detection systems purely on the basis of monitoring the patient's sleep and responses to questionnaires that elicit the necessary demographic and behavioral details. This is especially significant in the light of the fact that today pathology detection is mostly symptom-driven. Our models for detection of pathology are not symptom-driven. Instead they are based on subjective and objective information gathered from patients visiting the sleep laboratory. This will facilitate screening of patients visiting physicians with any complaints and short-listing likelihood of pathologies they might be suffering from. Thus, in addition to early detection, it can probably even serve to pre-empt occurrence of other pathologies. The models could also be embedded into devices designed to monitor patients during the night of sleep and relay the information to a central repository or emergency units, thus providing untethered healthcare services.

1.9 Summary

In this chapter we provide the overview of how research in human sleep has evolved from subjective behavioral assessments initially to introduction of objective measures with the advancement of technology over the past several decades. The suite of electronic instruments used to collect data from observing the patient during sleep is called Polysomnogram.

There are plenty of interesting associations and relations, which would enormously benefit the researchers in this field. Past attempts have involved abundant use of statistical techniques to dig out these relations. Our emphasis will be on use of association rule mining to extract novel and interesting patterns from the data collected. We are particularly interested in patterns relating sleep variations with degrees of pathology. Also, domain experts view with interest rules that can establish relationships among demographic, behavioral and clinical factors considered in the analysis.

The dataset for this thesis comprises the questionnaire responses of patients visiting a sleep laboratory in addition to the measurements made by polysomnogram instruments and the summary report compiled by technicians. Data collection was not easy since information was spread over two locations. As a result, we were unable to obtain all information that was originally intended for most patients. We attend to this problem by modifying our datasets for analysis accordingly.

This thesis shows that medical experts can rely on techniques used in data mining instead of purely statistical techniques for data analysis. It demonstrates that association rules are more informative than logistic regression, a technique used widely for medical data analysis. The thesis also introduces the concept of mining for association rules within user-specified windows in real-time to obtain rules with more information content.

With further research, results obtained from experiments performed during this thesis will find applications in the area of untethered healthcare and remote patient

monitoring. Medically relevant and strong rules can perhaps in the near future also lead to pre-emptive treatment for several pathological disorders.

Chapter 2

Background

2.1 Instrumentation and signal processing

Amongst all the polysomnogram instruments in use today, the EEG was the first to be used to analyze human sleep. The earliest description of the use of EEG to analyze sleep appears in the paper chronicling the work done by Loomis et al. [LHH36]. With the exception of REM, all other stages of sleep were distinguishable with the help of measurement logs obtained from EEG records. It was not until the addition of the EOG as another device to analyze sleep that REM stage identification became possible [AK53]. EOG, which detects the electro-potential difference between the front and back of the eye, is able to identify the REM stage, which is characterized by repeated bursts of rapid eye movement. Later, researchers found that EMG provided a more reliable marker for detection of REM than just the presence of bursts of rapid eye movements [CR94]. EMG is a device used to measure muscle tension using the surface electrodes placed over the chin. Also EMG electrodes placed over the skin on either leg pick up movement of legs during sleep. R&K conceived their sleep stage classification model on the basis of the collective experience and results of numerous researchers targeting similar goals.

Polysomnography was first used to detect sleep irregularities at Stanford University during the 1970s. Polysomnography combines tracings from respiratory and cardiac sensors together with EEG, EMG and eye movement recordings to partition sleep into different stages [But96]. EEG is a multi-dimensional signal recorded using an array of electrodes placed at different locations on the scalp. The EEG electrodes are placed over the central and occipital areas of the brain in accordance with the international 10-20 system (C3 or C4 and O1 or O2) [Jas58]. EOG electrodes are placed in the vicinity of the right and left outer canthus (canthus- the angular junction of the eyelids at either corner

of the eyes). EMG electrode is fixed over the chin to trace muscular variations. The EMG electrode placed over the skin surface on the leg captures leg movements. The raw signals derived from the sleep patient have extremely low voltages and are highly susceptible to electrical interference. The electrical background activity of the human brain is in the range of 1-200uV, while the Evoked Potentials (EP) have an amplitude of only 1-30uV[BS90]. Many techniques have been adopted to keep the amount of electrical interference to a minimum. One of the measures for instance is selecting high quality electrodes. However not all artifacts are easy to be contained or minimized. Artifacts resulting due to movement of the patient while asleep make signal processing complicated.

2.2 Sleep stage scoring and their interpretation

Every thirty-second piece of a sleep recording constitutes an epoch. However, for specific research experiments epochs with smaller magnitude have also been used. Sleep stage scoring involves classifying the polysomnogram signals on an epoch-to-epoch basis. Epochs are scored into sleep stages in accordance with the R&K model. Sleep staging is guided by the frequency of the waves recorded by the polysomnogram instruments. Among the various polysomnogram instruments, R&K relies most on the EEG recordings to aid in staging. The EEG frequency bands used by the R&K model for staging sleep are mentioned below:

Alpha Rhythm: 8 to 13 cycles per second or cps

Beta Rhythm: more than 13 cps

Delta Rhythm: less than 4 cps

Theta Rhythm: 4 to 7 cps

Table 2.1 below gives the outline of the R&K model for sleep staging using polysomnogram data.

Stage/Stage	EEG	EOG	EMG
Relaxed Wakefulness	Eyes closed: alpha (8-13cps) Eyes open: lower voltage, mixed frequency	Voluntary control; REMs or none; blinks; SEMs* when drowsy	Voluntary movement; tonic activity relatively high
Stage 1	Relatively low voltage, mixed frequency Maybe theta (3-7 cps) activity with greater amplitude	SEMs	Tonic activity, maybe slight decrease from waking
Stage 2	Relatively low voltage, mixed frequency. Presence of sinusoidal waves (sleep spindles) 12-14 cps. Presence of negative sharp wave followed by slow positive component (K-complex) for time ≥ 0.5 sec	Occasional SEMs	Tonic activity, low level
Stage 3	≥20 ≤ 50% high amplitude (>75 μV), slow frequency (≤ 2cps)	None, picks up EEG	Tonic activity, low level
Stage 4	>50% high amplitude, slow frequency	None, picks up EEG	Tonic activity, low level
REM	Relatively low voltage, mixed frequency, sawtooth waves, theta activity; slow alpha	Phasic REMs	Tonic suppression

Table 2.1: Simplified outline of the R&K sleep scoring criteria based on polysomnogram signals (Modified from Chapter 15, Page 1202, [CR94]).

* SEMs – Slow Eye Movement

2.3 Sleep disorders

The more common of the numerous sleep disorders afflicting patients include narcolepsy, insomnia, excessive daytime sleepiness (EDS), snoring, periodic leg movement syndrome and sleep apnea. Sleep apnea is perhaps the most commonly diagnosed sleep disorder. This disorder has quite a few variants, which have been studied closely due to their common occurrence. Sleep apnea is a disorder where the patient experiences difficulties in breathing for several seconds during sleep. Based on the degree of inconvenience, we find physicians diagnosing patients with the different variants of sleep apnea. Hypopnea, for instance, is the condition where patient can breathe with some amount of effort and difficulty (can vary from person to person from mild effort to concerted effort) for several seconds during sleep. Obstructive sleep apnea on the other hand, is the case when the patient completely fails to inhale for several seconds during sleep. Of the different variants, Obstructive Sleep Apnea or OSA is found to be the most predominant form of sleep apnea affecting people.

2.4 Waikato environment for knowledge analysis – WEKA

For the mining task we seek the services of an open source machine learning suite called WEKA (stands for Waikato Environment for Knowledge Analysis). WEKA comprises of a collection of machine learning algorithms catering to different data mining applications and domains [FW00]. Being open source software, not only is it possible for users to utilize the various machine learning algorithms to perform different experiments but also suitable for developers to extend the algorithm functionality to build different applications. At WPI, students have worked on optimizing the execution of the association rule mining algorithm and extended it to handle other more complex data formats such as set-values [Sho01, Sto02] and time sequences [Pra02].

2.5 Mining association rules over transactional data

Association rule mining is used to extract interesting correlations from datasets having several data items. The discovery of interesting associations and relationships among data items in large sets of data helps analysts and managers enormously during the decision making process [AS94]. The original motivation for and the implementation of association mining algorithms were more applicable to providing solutions in business domains. In this thesis, we attempt to gain its benefits by applying it over the medical domain.

We now explain briefly a few important points relating to the functioning of association rule mining. Every variable or attribute-value pair that forms part of the dataset is called an item. A set of items is referred to as an itemset. Association rule mining employs two basic measures to extract interesting relationships between itemsets (set of one or more items) in a dataset. The two measures are called support and confidence.

To explain the terms and measures used in association rule mining, we introduce an example. Consider the dataset to represent the transactions of individual customers visiting a store. Each row of the dataset comprises any combination of items ('A', 'B', 'C', 'D' and 'E') indicating the set of items purchased by each customer during a single transaction. Each row is referred to as an instance of the dataset.

Consider the dataset comprising ten instances as shown in Table 2.2. Each instance indicates the set of items or itemset purchased by the customer from the store. Consider also the rule A & B ➔ E. This rule states that the presence of A and B in an instance makes it likely for E to also occur in the instance. This likelihood is measured by the *Confidence* of the rule. Confidence informs us as to how frequently the consequent of the rule (in this case, itemset Y = {E}) appears along side the antecedent (in this case, the itemset X = {A,B}) among the instances forming the dataset. The *Support* for a rule is the percentage of instances in the dataset that contain all items appearing in the rule.

Transaction number	Itemset
1	A, C, E
2	A, B, C
3	D, C, E
4	A, B, E
5	C, D, E
6	D, E
7	A
8	C, E
9	A, E
10	B, C, D

Table 2.2. Dataset with instances representing customer transactions

Mathematically these measures may be expressed in terms of a probability relationship as,

$$\text{support}\,(X \longrightarrow Y) = P(X \cup Y) = \frac{\text{number of instances with items } \{A, B, E\}}{\text{number of instances in the dataset}} \qquad (1)$$

$$= \frac{1}{10} = 0.1 = 10\,\%$$

$$\text{confidence}\,(X \longrightarrow Y) = P(Y|X) = \frac{\text{number of instances with items } \{A, B, E\}}{\text{number of instances with items } \{A, B\}} \qquad (2)$$

$$= \frac{1}{2} = 0.5 = 50\,\%$$

where $P(X \cup Y)$ = probability of the set $\{X \cup Y\}$ in the dataset,

$P(Y|X)$ = conditional probability of Y given X

The Apriori algorithm [AS94] is the standard algorithm to mine association rules. The apriori algorithm is a two-stage process.
Stage1 consists of finding all frequent itemsets and
Stage2 consists of generating rules from the frequent itemsets.

The input to the association rule mining algorithm is the dataset to be mined together with thresholds of minimum confidence and minimum support that the rules output must satisfy. In other words, the rules output by the system will all have confidence and support values equal to or above the user-specified values of confidence and support.

2.5.1 Generating frequent itemsets

The apriori algorithm is based on the Apriori principle [AS94] for generating frequent itemsets. An itemset is considered frequent if the ratio of the number of instances in which it appears to the total number of instances in the dataset (i.e., the support of the itemset) is equal to or greater than the minimum support threshold value specified by the user. The Apriori principle states that all subsets of a frequent itemset must also be frequent.

The Apriori algorithm constructs the frequent itemsets level by level. On level1, it constructs all frequent itemsets of cardinality 1. On level2, it constructs all frequent itemsets of cardinality 2 and so on. Apriori assumes that the itemsets are arranged in lexicographic order. Joining of two itemsets at one level to generate the itemset of the next higher level takes place only if all the items save for the last item in the two itemsets are alike. For instance, consider two itemsets at level-3, C_1 and C_2 that have only the last item in the itemsets different.

$C_1 = \{c_{11}, c_{12}, c_{13}\}$;

$C_2 = \{c_{21}, c_{22}, c_{23}\}$

where $c_{11} = c_{21}$; $c_{12} = c_{22}$ and $c_{13} < c_{23}$

The condition $c_{13} < c_{23}$ ensures that the two itemsets being joined are not duplicates and helps maintain the lexicographic order while performing the join [HK00]. The new itemset obtained is one level higher. In this example, the itemset generated is of level-4. Let us name it itemset M. We observe that,

$M = \{c_{11}, c_{12}, c_{13}, c_{23}\}$ is the itemset resulting from the join between C_1 and C_2.

On the new itemset generated by the join, the prune step is employed. This step verifies whether the Apriori condition holds true for the new itemset. It checks if all the subsets of the newly formed itemset are frequent. If this is not found to be true, the new itemset generated during the join is pruned. In the example under consideration, the level-4 itemset M has 4 level-3 subsets.

$M_{11} = \{c_{11}, c_{12}, c_{13}\}$; $M_{12} = \{c_{11}, c_{12}, c_{23}\}$; $M_{13} = \{c_{11}, c_{13}, c_{23}\}$; $M_{14} = \{c_{12}, c_{13}, c_{23}\}$

Each of these subsets, viz., M_{11}, M_{12}, M_{13} and M_{14} must be frequent for itemset M to be possibly frequent. Whether the itemset M is frequent or not is determined by counting the actual support of the itemset M by going over every instance of the dataset.

2.5.2 Rule generation process

Following the generation of frequent itemsets, we enter the second stage of association rule mining, which is rule generation. Our objective is to generate strong association rules that satisfy the thresholds of minimum support and minimum confidence specified by the user. As mentioned earlier, confidence for a rule $X \rightarrow Y$ is the ratio of the number of instances in the dataset containing X that also contain Y. The same can be expressed as,

$$\text{confidence}\,(X \longrightarrow Y) = P\,(Y \mid X) = \frac{P\,(X \cup Y)}{P(X)} = \frac{\text{support}\,(X \cup Y)}{\text{support}\,(X)} \qquad (3)$$

We use equation (3) over the generated frequent itemsets to obtain rules with acceptable levels of confidence as follows.

For each frequent itemset M

For each possible subset, S of M

We calculate the confidence of the rule, (M-S) → S using equation (3).

$$r = \frac{\text{support (M)}}{\text{support (M-S)}} \qquad (4)$$

If the value of 'r' is greater than or equal to the minimum confidence threshold set, then the rule, (M-S) → S is output together with its confidence and support values.

2.6 Mining association rules over complex data

The traditional association rule mining algorithm supports only a limited number of data formats that are more common in market-basket datasets and similar transaction-related domains. The data items normally are numeric or categorical in nature. For complex datasets having attributes that hold a set of items or time-series data, a modified version of the Apriori algorithm is required to perform the mining process. We explain in detail the modifications performed to extend the system to mine with more complicated data formats in section 4.1.

2.7 Additional statistical measures for association rules

One of the strengths of association rule mining is that it extracts all associations existing in the dataset that satisfy the user-specified minimum support and minimum confidence levels [LHM99]. This strength could also turn out to be a disadvantage since more often than not, the resulting set of rules is very large and analyzing them can become difficult. This is especially true in the case of data sets with several highly correlated attributes. Another problem that is often noticed is that we obtain rules with very high confidence values even though the attributes forming part of these rules are actually independent. To identify and thereby prune such faulty associations another

measure called 'lift' is used. For an association rule of the form X → Y, lift can be mathematically expressed as,

$$\text{Lift}\,(X \longrightarrow Y) = \frac{P(Y|X)}{P(Y)} \qquad (5)$$

The lift provides a value that helps in understanding the extent of association or dependence the consequent attribute-values have on the antecedent attribute-values that form part of a particular rule. This in turn helps in evaluating the interestingness of a rule. Rules indicating high dependence of the consequent attribute-values on the attribute-values that make up the antecedents are considered particularly interesting. High values of lift are indicative of highly dependent and hence interesting rules. Lift values closer to zero are characteristic of weak associations between the antecedent and consequent attributes with low confidence. Value of lift close to 1 is indicative of the consequent occurring virtually independent of the antecedent. Rules with high confidence and lift values hovering around 1 are indicative of the fact that the association between the antecedents and consequents may have occurred by pure chance, since the consequent occurs roughly to the same proportion when considered in isolation.

Using values of confidence, support and lift it is possible to estimate the statistical significance of the generated rules by performing the chi-square test for independence and correlation [Alv03]. Pearson's chi-square test is based on the comparison between the observed and the expected frequencies (counts). The chi-square statistic is the sum of the squares of the differences between the observed and the expected frequency. The statistic can also be calculated using observed proportions provided the total number of subjects (and hence frequencies) is known. In our case, since we know the total number of instances in the dataset, we make use of the three proportional measures confidence, support and lift to determine the chi-square value as mentioned earlier.

Below, we illustrate how the method introduced in [Alv03] of calculating the chi-square statistic using the three measures of confidence, support and lift helps in determining the interestingness of a rule. Consider, for instance the association rule,

X *(yes)* ➔ Y *(yes) [confidence = 0.82, support = 0.23, lift = 3.2]*

The above rule indicates that there is a correlation between the attributes X and Y. We are trying to ascertain the statistical validity of the association given by this rule using the chi-square test. Using the technique specified in [Alv03] and the values of confidence, support and lift specified for the rule above, we arrive at a chi-square value of 84.6. Consulting the chi-square look-up chart for 2x2 table with 1 degree of freedom, we find that the value corresponds to a significance level that is better than 0.001. In other words, the value indicates that there is 1 in 1000 chance that the correlation exhibited by X and Y in the association rule occurs by chance. This indicates that the rule is represents a true association between X and Y and that it is statistically robust.

2.8 Logistic regression

Logistic regression is a statistical technique that is used widely for analyzing results from clinical and laboratory tests or trials in the medical domain. An English scientist, Francis Galton developed the technique of regression analysis in the 19th century. Logistic regression is a specialized version of simple regression analysis.

Regression analysis is used in predicting value of one variable (also called dependent variable) using values from one or more other variables (also called independent variables). It helps in determining the accuracy with which the independent variables predict the dependent variable. Further, regression analysis can also help in assessing whether a particular relationship is statistically significant. In other words, whether the relation obtained in a small sample is also seen in an entire population. The dependent and independent variables may hold numeric or categorical data values. Ordinary regression analysis, however, has an important limitation. It requires the dependent variable to hold continuous values. Logistic regression, on the other hand, can handle discrete-valued dependent variables.

Generally, in many applications where logistic regression is applied, the dependent variable takes only two values. The two values represent either presence/ absence of a condition or true/false type of responses. For these applications the logistic

regression applied is referred to as dichotomous or binomial logistic regression. There may also be applications where the dependent variable may take more than two values. Logistic regression applied on such datasets is referred to as multinomial logistic regression.

The independent variables in logistic regression may have numeric or categorical values. The logistic regression makes no assumption of the distribution of these variable values. Another advantage with logistic regression over regression is that the fitted regression co-efficients maintain probabilities within (0, 1) range rather than values, which could stretch between ($-\infty$, ∞) in case of regression. This is achieved by log-odds transformation leading to a model given by the expression as described in [ED96],

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_n x_n \qquad (6)$$

p = probability of the dependent variable assuming a particular value,
n = number of independent variables.
$\beta_0, \beta_1, .., \beta_n$ = regression co-efficients,
$x_1, x_2, ..., x_n$ = independent variables.

The regression co-efficients are estimated using the method of "maximum likelihood". The response variable is assumed to have a binomial distribution.

$$p = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n)}} \qquad (7)$$

The problem of establishing a relationship between the dependent variable and the set of independent variables ultimately boils down to identifying the subset of the latter having the greatest influence in determining the former. In logistic regression, deciding

whether a variable must enter the model or be removed from it is based on the chi-square statistic.

2.9 SAS – statistical software package

A number of statistical software packages are available today to perform analysis over the data. We make use of the UNIX SAS version 8.1 available at WPI to perform the logistic regression experiments. This version has a graphical environment for performing all necessary tasks over the data. Most of the manipulation, analysis of the data and plotting of the graphical output can be performed by using the pull-down menus from the windows that appear when SAS is activated. We could alternately write SAS programs to perform the different tasks. Typical SAS programs comprise of DATA step and several PROC steps for analysis. DATA step is used to load all the observations that form part of the dataset. Once the data has been read into the SAS dataset, SAS procedures are activated by the PROC step to analyze the data [ED96].

The weighted independent variables along with the regression co-efficients are used to estimate the response or dependent variable. SAS uses statistical measures to estimate the fitness of the model. The Akaike's information criterion (AIC) [Aka74] and the Schwartz's criterion (SC) are used to assess competing models. The two measures strive to select the most parsimonious model by balancing the fit against other factors [ED96]. Lower values of AIC and SC indicate more preferred models.

There is a third term, $-2\log L = -2\Sigma\log(p_j)$, where p_j are the predicted values obtained by replacing the β terms in equation (7) with estimated values.

The statistical measures to gauge model fitness lead us to test the validity of the hypothesis that says that all regression co-efficients of the model have a value that equals zero. We make use of chi-square to test if any of the independent variables has an influence on predicting the dependent variable.

The ‘Maximum Likelihood Estimates’ section of the SAS output provides us with calculated statistical values of the standard error, chi-square results and probability values of the explanatory variables in order to determine whether the hypothesis is true (i.e., whether the regression co-efficients take a value other than zero).

SAS also provides us with four measures to assess the predictive ability of the model selected. These are Somer’s D, Gamma, Tau-a and c. Higher the values of c, D, Gamma and Tau-a better is the predictive power of the model.

2.10 Summary

Polysomnogram is a suite of instruments used to measure physiological parameters of different body functions during patients sleep. It includes the Electro-encephalogram (EEG), Electro-oculogram (EOG), Electro-myogram (EMG) besides other instruments like the one to measure the oxygen potential in the patient’s bloodstream.

Staging of sleep is performed on the basis of the R&K model of sleep stage classification. It primarily uses the frequency of the EEG signals and the subjective interpretation of the technician’s observations during the night under observation to classify the patient’s sleep.

Some of the common sleep disorders are narcolepsy, obstructive sleep apnea, hypopnea, insomnia and periodic leg movement syndrome. The majority of patients reporting sleep disorders suffer from obstructive sleep apnea (OSA) in which the patient struggles to breathe for several seconds or minutes during sleep.

Association rule mining was designed and developed to be more suitable for data analysis over business domains. However, since they extract correlations between the various attribute-values considered, it is widely used in exploratory data analysis. Association rule mining generates rules from frequent itemsets based on two measures

called support and confidence. At every level frequent itemsets are generated in accordance with the Apriori principle. The principle states that all subsets of a frequent itemset need to be frequent. Rules output by the mining algorithm have support and confidence values equal to or above the user-specified minimum threshold values. We use the WPI-WEKA system for mining association rules. We determine the strength of the resulting associations by calculating the lift for the rules. To prune associations occurring due to pure chance, statistical robustness of rules is evaluated using the chi-square test for statistical significance.

Regression is a technique used to predict a dependent variable based on values taken by several participating independent variables. Ordinary regression analysis imposes the constraint that the dependent variable carries only continuous values. Logistic regression on the other hand supports cases where the dependent variable carries discrete values. Another advantage of logistic regression is that the regression coefficients are fitted to have probabilities within the (0,1) range. Logistic regression uses chi-square testing to determine correlations between the independent variables and the dependent one. Logistic regression experiments in this thesis have been carried out using the SAS statistical package.

Chapter 3

Review of Related Literature

3.1 Mining over questionnaire data

Mining over questionnaire data is not always straightforward. There are several reasons for this. One of the most common factors, though, is the fuzzy nature of the data obtained from patient responses. Consequently, it is not surprising that we repeatedly encounter occasions during the manual entry of data where subjective interpretation of questionnaire responses becomes the only recourse. We provide a sample set of questions from the sleep questionnaire in Appendix-A of this report. The selection of the features to represent the data has a tremendous effect on the consequence of a mining process irrespective of the domain [BP01].

There have been very few papers investigating mining over data obtained from sleep questionnaire. The most relevant work relates to screening patients visiting a sleep laboratory for sleep apnea by mining over the fuzzy knowledge representations in the questionnaire responses using intelligent aggregation techniques [NH99]. The paper refers to exploiting the prior knowledge of the domain that is available (via clinical knowledge of sleep apnea syndrome and questionnaire responses) and the derived knowledge (calculating membership grades for patient responses to questionnaires) to obtain an insightful result. Each attribute-value (question-response) pair is represented by a relevant weight. Groups of variables (attribute-value pairs) are aggregated by one of the three aggregation methods used to provide a membership grade. The three data aggregation methods are Principal Component Analysis, Ordered Weighted Average (OWA) and Weighted Ordered Weighted Average (WOWA). These aggregation techniques use weighted vectors to bias the data with respect to relevance and reliability. They use three weighted vectors based on the relevance of each variable, reliability of the

data values for each variable and the membership grade for each data value of each variable. The medical expert having the necessary domain knowledge decides all the weighted vectors.

3.2 Advantage of Association rule mining over Classification

Generally medical data is analyzed using classification techniques, clustering or regression. Association rule mining is rarely used in this domain. However, there are numerous disadvantages with the classification approach. An excellent argument in favor of using association rule mining instead of classification is presented in [OOBS+01]. Since association rules are combinatorial in nature, a large number of patterns are obtained from the dataset. Some of the associations obtained from the mining process may contain redundant information, may be irrelevant or may describe trivial knowledge that is not of interest. The paper discusses an algorithm aiming to generate only rules that are interesting. The paper argues that medically interesting events are those that occur rarely thus helping identify abnormal circumstances. Hence, medically significant rules were bound to have low support. Incorporating constraints helps in preventing some of the trivial rules from being generated. [OOBS+01] describes an algorithm to obtain interesting rules by specifying antecedent, consequent and group constraints for each attribute of the dataset.

3.3 Inductive learning for sleep stage classification

Inductive learning algorithms have also been used to aid sleep signal classification. To determine the usefulness of sleep stage scoring, nine inductive learning algorithms were tested on sleep data obtained from 161 patients [BR93]. The best overall performance among the several algorithms studied was shown by C4 (precursor of C4.5 classification algorithm [Qui93]) and MDL (Minimum Description Length [Ris83]).

3.4 Influence of papers reviewed in developing our approach

The related work that we review helped us understand drawbacks and advantages in the different approaches. This learning influenced us in selecting the approach to tackle the problem being considered in this thesis.

The first paper [NH99] explains the technique adopted to screen patients suffering from sleep apnea. However, our focus in this thesis is more exploratory in nature and we do not restrict ourselves to just patients suffering with sleep apnea. Besides the paper focuses on determining the relevance and reliability of the questionnaire responses to screen patients. Our approach instead, is to assume that the responses to questions posed in the sleep questionnaire are reliable and thereafter use medically known thresholds to transform some of the patient responses to a more standard response based on suggestions made by the domain expert.

The second paper [OOBS+01] that we review stresses the importance of association rule mining over classification-based algorithms. This paper serves as a validation for choosing association rule mining over other algorithms in pursuing exploratory data analysis. The paper also makes a strong case on identifying rules with low support. It considers these rules as the only ones that can claim to be 'interesting' since they might identify abnormal behavior or rare events. However we believe that this hypothesis may not always be true. Whether a rule is 'interesting' or not is heavily dependent on the extent to which the domain has been explored. Following the hypothesis on which the paper is based for a domain like sleep where there is still some room to explore the inter-play between the different sets of attributes (behavioral, demographic, clinical) may result in several interesting associations being discarded. Therefore, we do not introduce any constraints to capture only 'interesting' rules.

The final paper that we review [BR93] analyzes the performance of several inductive learning algorithms for classifying sleep stages. The input data for these algorithms is the

unstaged raw signals obtained from the polysomnogram instruments. On the other hand, input dataset for this thesis has all these signals neatly staged to their respective sleep stages. However, it was interesting to find out which machine learning algorithms performed the sleep staging operation with high accuracy.

3.5 *Summary*

Mining over questionnaire responses is challenging. There are many reasons for this. Most problems are caused due to the fuzzy nature of the questions and patient responses. At times subjective interpretation of the responses needs to be done which also makes analysis of the responses difficult. In this context, the [BP01] paper extols the importance of feature-selection in representing data for effective analysis.

We discuss three papers having close association with the association rule mining technique that we adopt and the domain of sleep where analysis using several techniques has been performed. The techniques used where designed with different objectives. While the [NH99] paper describes the use of statistical techniques like Ordered Weighted average to determine the significance and reliability of patient responses to a questionnaire, the [BR93] paper is more interested in determining, which inductive learning technique has the best performance in sleep stage classification.

Chapter 4

Our Data Analysis Approach

In this chapter we describe the data mining approach that enables us to mine rules with relative as well as absolute time information. We discuss the pros and cons of the two mining approaches. Since adoption of any data mining technique is data dependent, we start by pointing out relevant features of the dataset under consideration. The data, which is collected from two different locations, has patient information maintained in three records.

(a) Questionnaires (see Appendix A),
(b) Sleep stage information and observation from a few polysomnogram instruments, and
(c) Summary reports compiled by physicians and technicians.

However, for a majority of the patients, information from some but not all of the three sources noted above were available. The content in the different data sources also varied in its depth or detail they provided. In other words, while questionnaires and summary reports were more broad and focused on the 'big picture' or overall assessment of the patient, polysomnogram observations recorded microscopic behavior of the patient during every epoch (30 second duration) of the night's sleep. The questionnaires and summary report information was obtained from UMass Memorial Medical Center (Worcester, MA), while the polysomnogram recordings made during every epoch of sleep study was obtained from Day Kimball Hospital, (Putnam, CT).

To provide a more incisive analysis we organize the data into three levels of detail. We call the three datasets macro-level dataset, micro-level dataset and mixed-level dataset. The analysis is also correspondingly called the macro-level analysis, micro-level

analysis and mixed-level analysis. The constitution of the different datasets is given below.

Macro-level dataset = questionnaire responses + summary report details,

Micro-level dataset = sleep stage information + information from a few polysomnogram instruments

Mixed-level dataset = questionnaire responses + summary report details + polysomnogram observations.

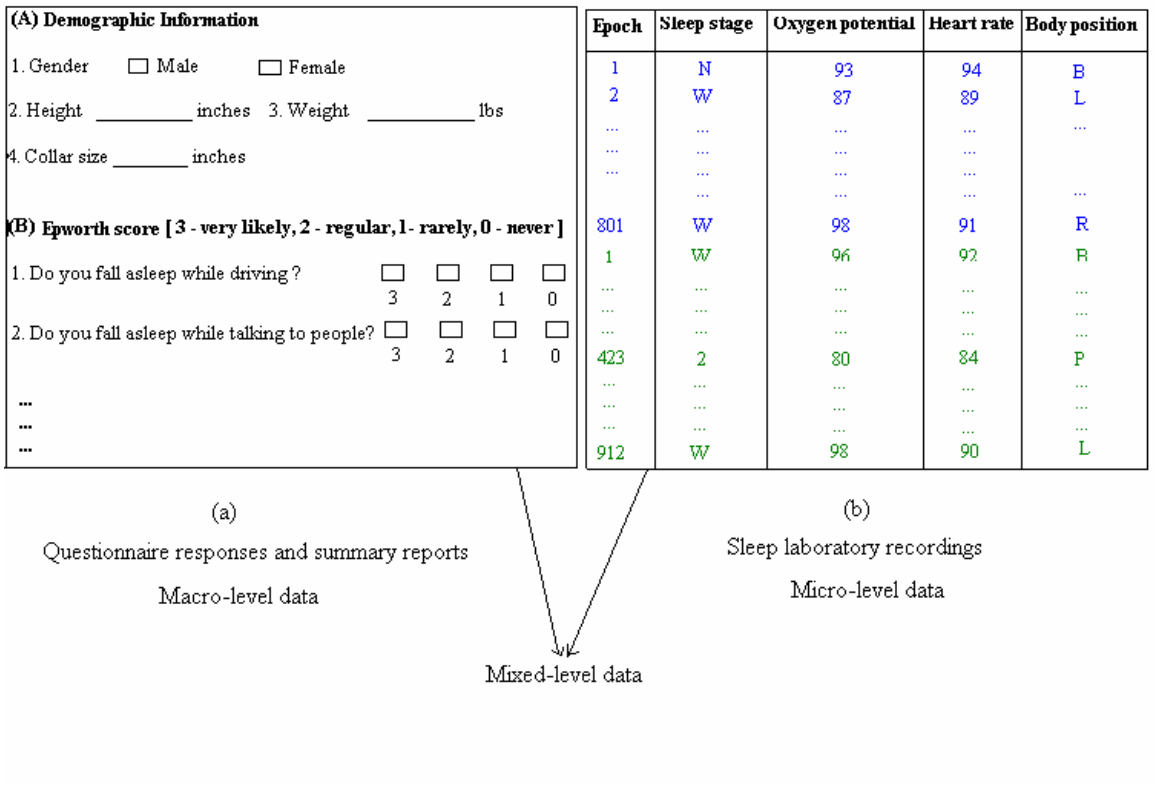


Fig. 4.1 Organization of data into macro, micro and mixed-level datasets

The questionnaire responses and summary reports providing overall statistics regarding pathological conditions, demographic and behavioral assessment of patients and clinical interpretations made by the physicians are combined together to form the macro-level dataset. Micro-level dataset includes epoch-by-epoch information of the patient's sleep stages, oxygen potential, Electro-cardiogram readings of average heart rates every epoch, technician's record of the patient's body position during the epoch and recording of any

significant event of interest during the epoch. Mixed-level dataset is a compilation of all information used in macro and micro-level analysis. Only patients having both macro and micro-level information are considered while generating the mixed-level dataset. This dataset as a result has the least number of instances. In the remainder of this chapter we describe the data mining technique used to mine rules with relative time and evolution of a technique to mine rules with absolute time information over sleep data. In the chapters marked 5, 6 and 7 we describe in detail with experimental support how the technique has been used to perform macro, micro and mixed-level data analysis.

4.1 Association rule mining with time sequences

The sleep laboratory recordings collect data during every epoch of a patient's sleep. In other words, the recordings are attribute-values measured at different times during the patient's sleep. Hence, the data recorded for every patient during the entire night of sleep represents a time series. For mining association rules over time sequence attributes, the concept of events was introduced [Pra02]. Events are templates or patterns in data that capture time-related occurrences of interest. The temporal association rules are capable of representing all possible temporal relations that may be present among the set of events. In the domain of sleep analysis, a number of events can be identified depending on the objective of the analysis. For example, events of interest may include various stages in sleep or frequency of arousals that occur in a typical recording. The existing system captures the relationships between the events in relative time. For instance consider the following rule obtained from the rule-mining system.

sleepstage2 = 0:1 *AND* sleepstage2 = 4:5 ➔ sleepstageREM = 2:3

[Confidence = 79.49%; Support = 90.91%]

This rule captures the temporal relationship between two events, viz., stage-2 of sleep and stage REM. The interpretation of the rule is as follows:

If a sleep stage-2 event is observed and later on another sleep stage-2 event is observed, then there is 79.49% likelihood that a REM stage event has occurred sometime in between the two stage-2 events. In other words, this rule with very high support and

confidence measures states that REM stage of sleep occurs in-between two occurrences of stage-2 of sleep. This is a well-known phenomenon of sleep progression known to sleep experts. The same rule can be represented diagrammatically in Figure 4.2. In this figure t_0, t_1, t_2, t_3, t_4 and t_5 represent arbitrary time indices that help sort the event occurrences in time.

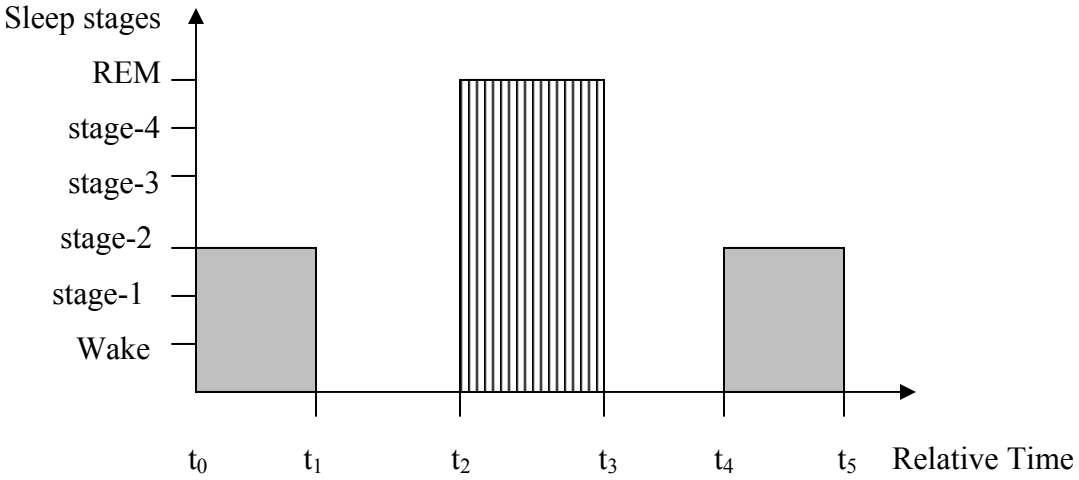


Fig. 4.2. Relative time variation of stage-2 and REM.

The extended association rule mining system helps understand the timing dependencies between the observed events. Thus, we obtain more information by mining over the time sequence variable than was possible using traditional association rule mining.

4.2 Event identification using filters

We wrote special filters to identify and isolate events of interest existing in the different datasets. For instance, sleep experts consider baseline oxygen potential to be an important marker or event. Baseline oxygen potential is the measurement recorded by the pulse oxymeter during the epoch when the person enters stage 1 of sleep with no event being observed for the first time since he/she drifted into sleep. Therefore, we wrote a special filter to identify and isolate epochs during the whole night sleep when the pulse oxymeter readings indicated values equal to or above baseline oxygen potential and

epochs during which the drops in oxygen potential measured more than three percent below the baseline potential. This process of converting the epoch-by-epoch data sequence into an event-based sequence leads to reduction in the size of the data. There are universally accepted ranges for deciding whether heart rates, oxygen potentials, body-mass index values are within normal or acceptable ranges. There are also additional ranges used in the field of medicine to identify the severity of the disorders if readings indicating the performance of the different physiological processes do not fall in the normal range. We wrote filters transforming the raw dataset to hold these range details.

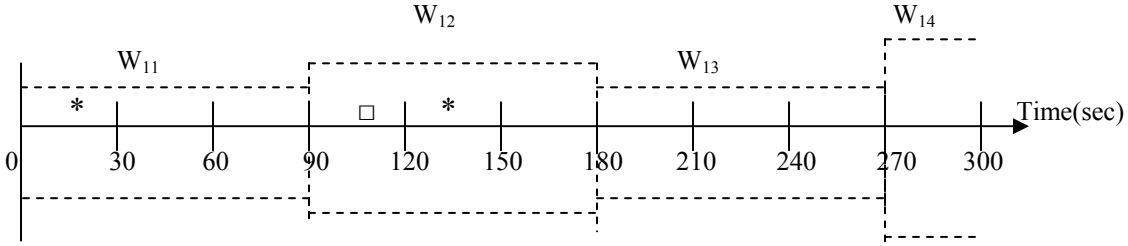
4.3 Limitations in association rule mining with relative time

Mining of association rules with relative time information contained in them though sufficient to capture some interesting patterns in the sleep data does not uncover all the interesting patterns. For instance, just relative time is not sufficient when the time span between the events of interest or the length of time for which a particular event of interest was witnessed is considered important. Also in the domain of sleep analysis, the epoch periods in real-time during which certain events occurred or the length of time for which an event occurred is considered important since they can co-relate with pathology in many cases. They also help physicians to zero-in on a smaller subset of data or information to make their assessments. In order to address these limitations, we have extended the temporal association rules described above with time windows.

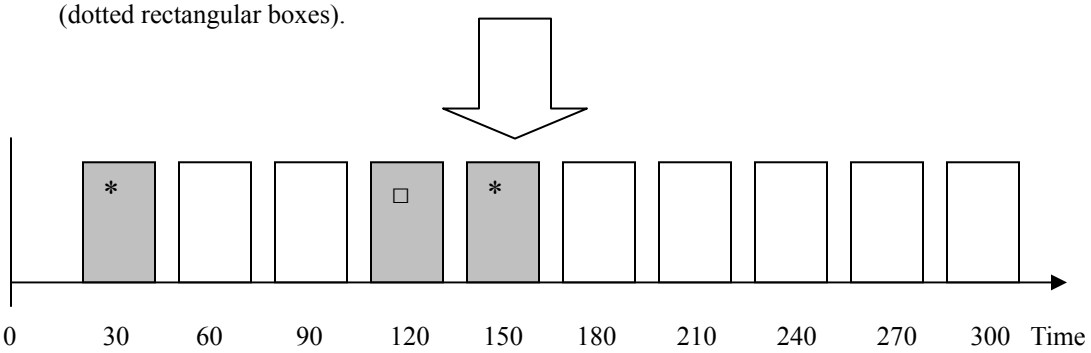
4.4 Temporal association rules using time windows

The WPI-WEKA system [Pra02] was designed to mine for rules over time sequences using relative time. To mine for rules with real-time information using the existing system we introduce the concept of windows. We divide the whole night sleep into equal-sized blocks called windows. Within these windows we identify the presence of different events of interest and the frequency with which the events occur within the windows.

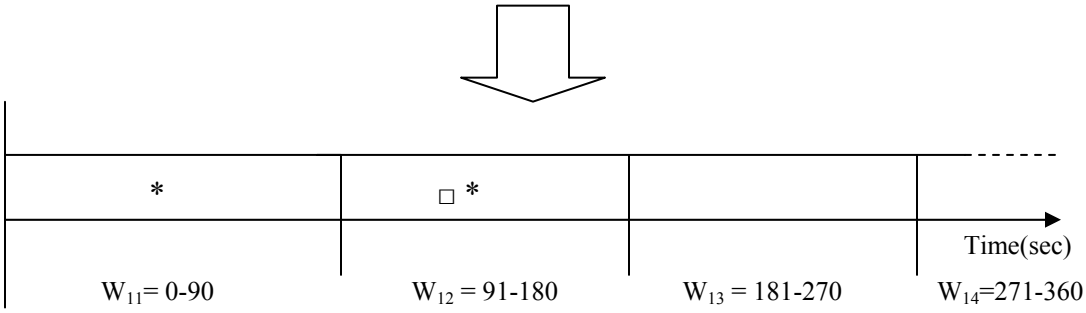
Figures 4.3 and 4.4 illustrate the aggregation of epochs into larger blocks of fixed size that we refer to as 'windows'. Figure 4.3 shows a window formed by aggregating 3 contiguous epochs while figure 4.4 illustrates a window organized by aggregating 5 contiguous epochs together. Identical events may be noticed in contiguous epochs or they may recur with breaks or fragments (i.e., some epochs in-between may not have or experience the concerned event).



(a) Illustration of event occurrences during sleep and intended grouping of epochs into windows (dotted rectangular boxes).



(b) Representation of events of interest within epochs in the dataset

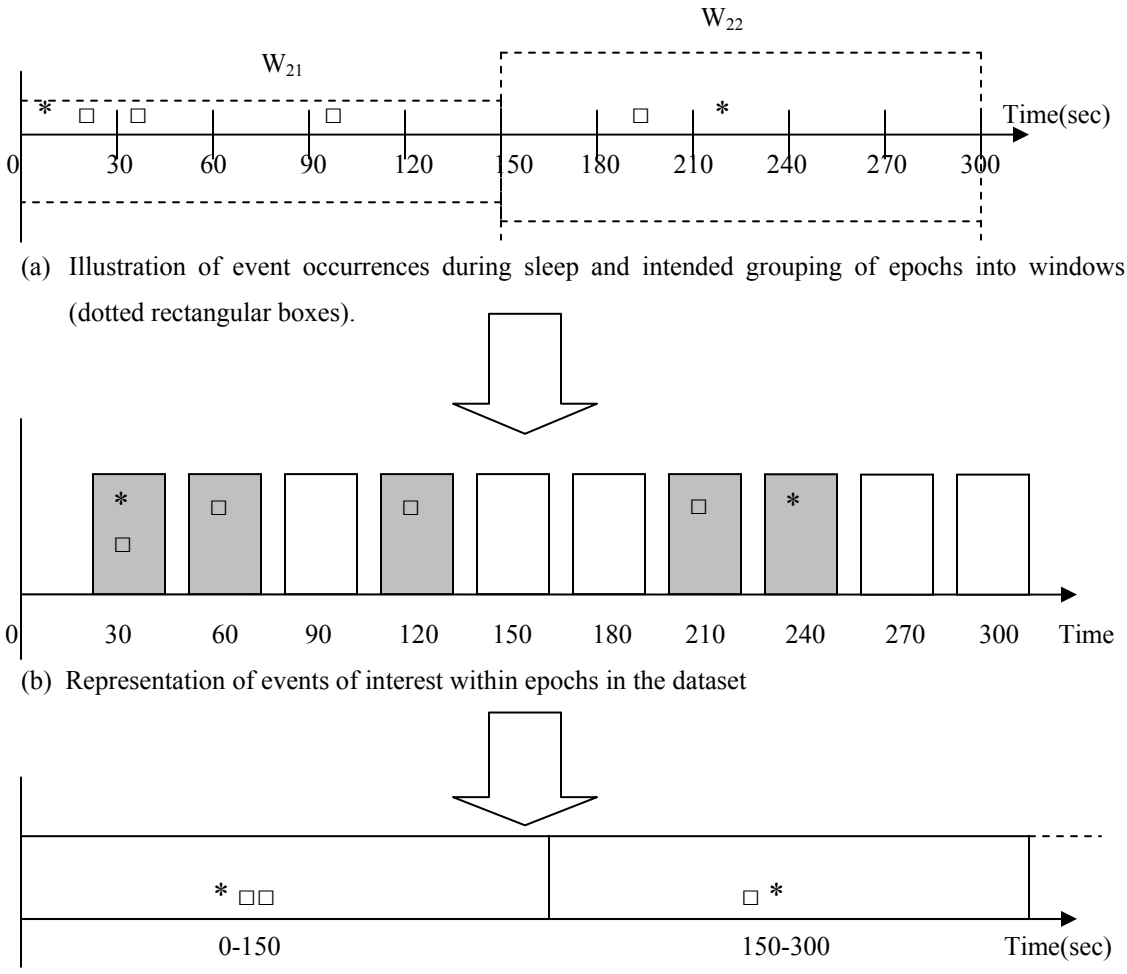


(c) Grouping of epochs into windows and depicting of events noticed within them.

Fig. 4.3 Building windows from epochs (Window = 3 epochs = 90 seconds)

In figures 4.3 and 4.4, the * and □ represent events that occur during the patient's sleep. The shaded areas in those figures represent the epochs when the events * and □ events are recorded.

Another example of converting the dataset into one recording events of interest within windows is shown below.



(a) Illustration of event occurrences during sleep and intended grouping of epochs into windows (dotted rectangular boxes).

(b) Representation of events of interest within epochs in the dataset

0-150
150-300
Time(sec)

(c) Grouping of epochs into windows and depicting of events noticed within them.

Fig. 4.4. Building Windows from epochs (Window = 5 epochs = 150 seconds)

For simplicity in analyzing fragments detected within windows and to keep the input data format for our system simple, we make certain concessions. If the same event occurs during consecutive epochs within the same window, then they are represented as though a single event of the same type takes place within the window (no fragmentation witnessed). For instance, in Figure 4.4.b, we find that the event represented by symbol □ appears in two contiguous epochs (epoch 1 = 0-30 seconds and epoch 2 = 31-60

seconds). However, since it is the same event occurring in contiguous epochs (where all the contiguous epochs are contained within the same window), we consider only one occurrence of the event within the window. There is another □ symbol in the same window, which represents the fragmented event that occurs during epoch 4 (120 seconds). Hence Figure 4.4.c shows only 2 occurrences of □ event within the window. This process of crunching identical events occurring in contiguous epochs into one leads to the identification of the number of fragments of the particular event noticed within the window. This parameter is of medical significance. It becomes very interesting whenever we are able to detect multiple fragmentations of the same event within a single window.

This technique is implemented in two steps with the help of the developed filters. The first stage involves the conversion of the sequence information in the raw dataset into a set-valued format. The set-values have the window-level information of the events that occur within them. The second stage is implemented by another filter, which accepts the transformed set-valued dataset as input to identify event-fragmentations within the windows. Pictorially the two-stage conversion process can be represented as shown in Figure 4.5.

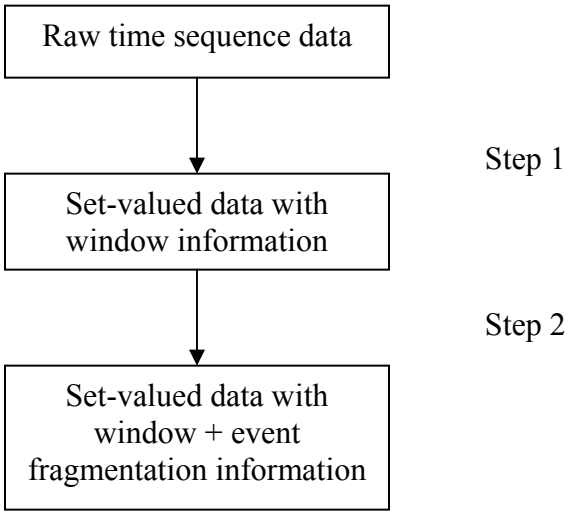


Fig. 4.5 Filters for transforming raw sequence data into set-valued with information on the windows within which events are witnessed along with the frequency of event-fragmentation

The difference in the representation of the events between the technique, which mines for rules with relative time information and our system which mines for rules in real-time can be explained with the help of Figure 4.4.b. In the figure we observe event □ occurs during the 1st, 2nd, 4th and 7th epochs of sleep. The system that mines rules in relative time would represent this event as a time sequence as follows.

Event □ = {1 : 2 ; 4 : 4 ; 7 : 7}[1].

The same dataset after passing through the filters in our system will transform the events identified into a set-valued dataset as shown below.

Event □ = {event-□-0-150#2 ; event-□-150-300#1}

This is illustrated in Figure 4.4.c. Events occurring in epochs falling under a particular window are placed directly within that window. Thus, we are able to transform a time-series dataset with events into a set-valued dataset with the event information together with the frequency of the occurrence of the event within a window recorded. As can be clearly seen from the two representations, more information is obtained using our approach due to more detailed information. The size of the window is a user-specified value provided to the filters at the time of transformation.

We give an example of the nature of rules obtained using this technique below.

stage-2 (750-800#1) && stage-R (750-800#1) && low-oxy (0-50#1) ➔ stage-W (300-350#1) && low-hrate (0-50#1)

[Confidence = 76.47%; Support = 10.74%; Lift = 2.1518; Chi-square = 14.4598; P<0.001]

The rule states that when stage-2 and REM are noticed between the 750th to 800th epochs with low oxygen event in the first 50 epochs, wake state during the 300th and 350th epoch with low heart rate in the first 50 epochs of sleep is witnessed. The high chi-square value is indicative of the statistical robustness of the rule. The same rule can be diagrammatically represented as in Figure 4.6.

[1] The actual input data file uses '^' character instead of ';' to separate epoch periods between which the event of interest was witnessed within a set.

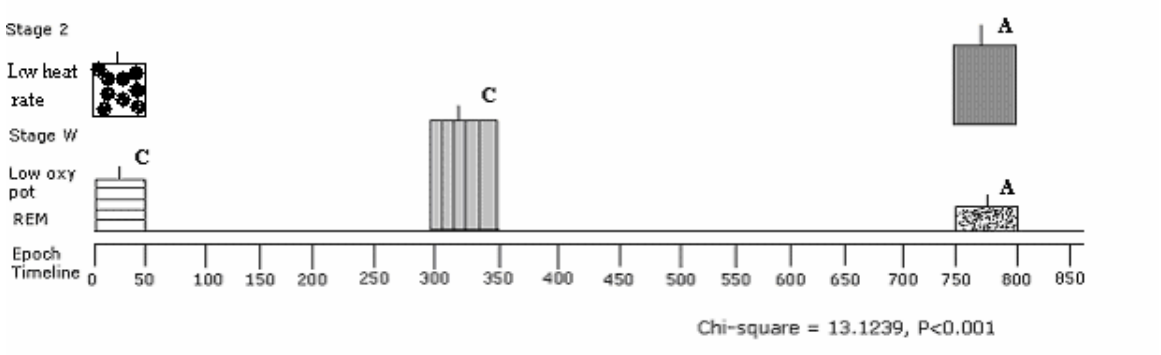


Fig 4.6. Statistically robust rule obtained from micro-level dataset

4.5 Contributions of our approach

Normally there is always some event of interest that occurs within every epoch. It is hence convenient to represent events of interest as time sequences. However, the WPI-WEKA system to mine for time sequences uses relative time to indicate the interrelationships between the participating attribute-values. Rules with relative time are not very insightful for this domain. Another disadvantage is because of the length of sequences (on an average a sequence has 800 data points; one for each epoch) the time taken to mine for rules is fairly long. Keeping this in mind, we developed filters to transform the sequence attributes into set-values. The set-valued representation of events reduced the size of the dataset making mining significantly faster. Also, the set-valued representation of events had time information contained in them making the rules more informative and easily interpretable. This made the task of analyzing the results easier for the domain expert.

4.6 Summary

For countering the non-availability of all information regarding patients and to analyze data at different levels of detail, the data collected is organized into macro-level dataset, micro-level dataset and mixed-level dataset.

The Waikato Environment for Knowledge Analysis (WEKA) is an open-source suite of several machine-learning algorithms.

Time-sequence attributes can be mined for association rules with relative time information using the WPI's version of association rule mining system that is built into WEKA. These rules provide details regarding the temporal interrelation among events occurring in the attributes. For the domain of sleep these rules are not always helpful. It is necessary also to mine for rules with real-time information present in them.

Towards this end, we developed the window-based association rule mining technique. This method mines rules by identifying events of interest within the real-time windows considered. Aggregating a fixed number of contiguous epochs of sleep forms these windows. Additional information in the form of the number of fragmentations that events of sleep exhibit within the windows is also provided. All this makes the rules more intuitive, easy to understand and hence analyze.

Chapter 5

Macro-level: Data Description and Analysis

The macro-level dataset as mentioned in Chapter 5 is generated from the questionnaire responses and clinical summary report of the patient on the night of sleep study. The information for producing this dataset was collected from paper documents and needed manual entry into the database. A total of 242 patient records were entered. The initial raw dataset had 70 attributes. Pre-processing over the dataset reduced the number of attributes to 63.

5.1 Macro-level dataset details

Here we provide a few examples on the type of attributes to gain an insight into the constitution of the dataset. It is mandatory for patients visiting a sleep laboratory to fill in a questionnaire. See Appendix A for a sample of the questionnaire provided to patients in the laboratory. The first part of the questionnaire solicits demographic information such as the patient's age, sex, height, weight, neck collar size etc. The second part, probes the patient on his/her general habits (e.g., alcohol intake) and sleep-specific habits (e.g., gasping for breath while asleep). Additional information from groups of variables aggregated together by index scales such as Epworth and Depression scale is also obtained. Epworth score is calculated on the basis of patient responses to a set of eight questions, while Beth's depression scale is a value calculated on the basis of the responses to a set of twelve questions. There are four optional choices for answering every question. The responses have points assigned to it, ranging between 0 and 3. The values for the two indices (epworth and depression indices) are obtained by adding the points assigned to the responses provided by the patients. The two indices are important benchmarks used by physicians to aid diagnosis. Epworth scale provides an index value (minimum value is 0, while maximum value possible is 24) indicative of the severity of

the sleep disorder. Likewise, Beth's depression scale also provides an index value (minimum value is 0 and maximum value possible is 48) indicative of depression being a contributing factor to sleep disorders. The third and final part of the questionnaire focuses on the sleep-related disorders and other medical problems that the patient suffers from.

Clinical information comprises the summary of observations recorded during the patient's study. The clinical data holds important information such as the percentage time spent by patients in the different sleep stages, their mean and minimum oxygen potential levels in the blood, latency time to stage1 and REM. A block diagrammatic representation of the Macro-level dataset is illustrated in the Fig. 6.1 below.

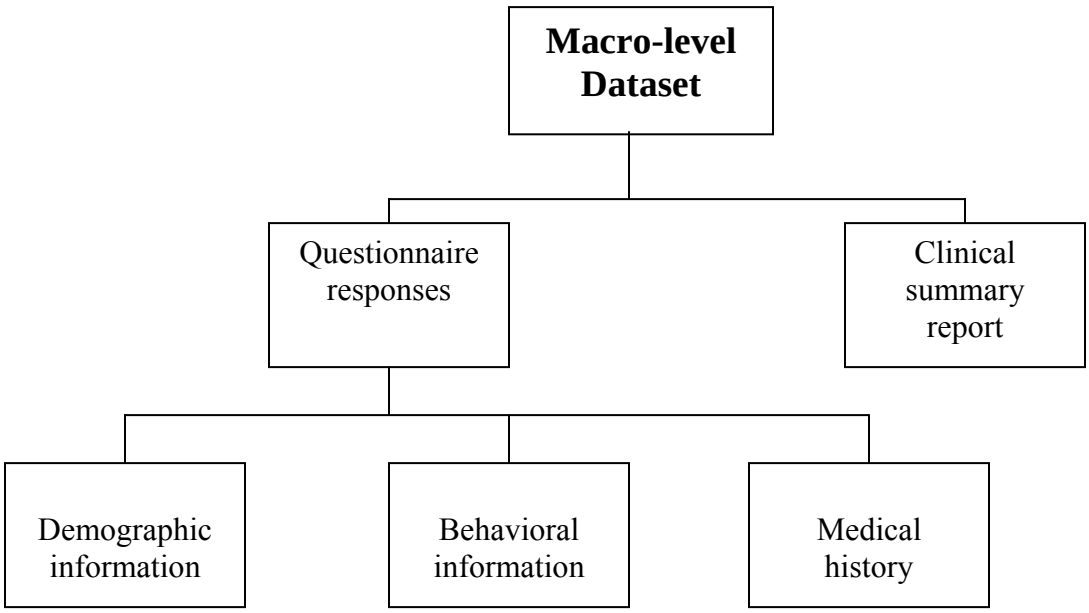


Fig 5.1: Block diagram of Macro-level dataset structure.

5.2 Pre-processing of macro-level dataset

Data extraction from questionnaires posed numerous tricky situations. Most questions required patients to pick the most suitable or applicable option from the list provided. A few questions though were descriptive in nature. For example, replies to queries probing problems other than sleep that patients suffered from or the list of medicines being taken by the patients at the time of visiting the physician.

A bulk of the pre-processing was necessary primarily because patient responses to questions posed in the questionnaire did not follow the expected pattern. For example, to questions where the expected responses were either 'yes' or 'no', there were cases where patients expressed the view that either both the options were to be considered true or that they were not very sure as to the right answer. They would generally indicate this by a question-mark symbol or by stating in words that they were not sure of the answer. In other cases, they would select one of the two valid options and would provide an explanation as basis to support their selection. This posed a tricky situation while trying to represent the information in the dataset. Further, deciding on how best to represent the responses expressed by the patient was not just important but also time consuming. One could argue that the additional pieces of information need to be retained since it might prove to be medically significant in making an accurate diagnosis. However, for this project we decided in consultation with the domain expert to drop all such data entries and identify them as missing values. Responses to questions based on the patient's behavioral habits and health also needed pre-processing in order to standardize the input data format. For instance, there were several answers to the questions probing the amount of alcohol consumption or the intake of caffeinated drinks etc. Based on the nature of responses to the questions, we came up with a discretized range of values that would help record information in a more standardized and consistent way. These values were also arrived at in consultation with the neurologist. Clinical record pre-processing also involved discretization of attribute values. Most of the clinical records were numerical attributes. We discretized the numeric attributes using standard cut-off values used in the medical domain. For instance, the average heart rate (measured in beats/minute) is discretized into three ranges – normal (average heart rates above 92 bpm, mildly-reduced (average heart rate between 86 and 92 bpm) and low (average heart rates less than 86 bpm).

A complete list of the 63 attributes and their actual values and values stored in the database following pre-processing is indicated below.

Attribute Number	*Attribute Name*	*Attribute-values (before preprocessing)*	*Attribute-values (post-preprocessing) with distribution*
1	gender	male, female	male = 136, female = 95
2	Body Mass Index class (BMI-class)	numeric value of height (ft/ inches) and weight (pounds)	underweight = 4, normal = 33, marginally-overweight = 12, overweight = 61, obese = 108
3	pain	descriptive explanation	yes = 17, no = 225, sometimes = 0
4	fibromyalgia	descriptive explanation	yes = 2, no = 240, sometimes = 0
5	depression	descriptive explanation	yes = 7, no = 235, sometimes = 0
6	hypertension	descriptive explanation	yes = 0, no = 242, sometimes = 0
7	heart-fail	descriptive explanation	yes = 0, no = 242, sometimes = 0
8	stroke	descriptive explanation	yes = 1, no = 241, sometimes = 0
9	stress	descriptive explanation	yes = 15, no = 227, sometimes = 0
10	allergies	descriptive explanation	yes = 2, no = 240, sometimes = 0
11	other problems	descriptive explanation	no = 209, fatigue = 22, no-communication = 3, relationship problems = 5, worry = 1, anxiety = 1, seizure = 1
12	age	numeric	Numeric (minimum = 7 years, maximum = 89 years, mean = 50.47 years, standard deviation = 16.06)
13	collar-size	numeric	0-14 inches = 6, 14-16 inches = 18, 16-18 inches = 80, 18-20 inches = 24, 20-above = 5
14	tired	descriptive explanation	yes = 224, no = 18, sometimes = 0
15	doze in meeting/ watching TV	yes, no	yes = 178, no =57, sometimes = 7
16	doze while talking	yes, no	yes = 40, no = 198, sometimes = 2
17	gasp for breath	yes, no	yes = 88, no = 140, sometimes = 3
18	snore while asleep	yes, no	yes = 201, no =38, sometimes = 1

Attribute Number	*Attribute Name*	*Attribute-values (before preprocessing)*	*Attribute-values (post-preprocessing) with distribution*
19	stop breathing	yes, no	yes = 105, no = 131, sometimes = 1
20	cramps in legs	yes, no	yes = 118, no = 116, sometimes =2
21	twitching of legs	yes, no	yes = 84, no = 152, sometimes = 0
22	sleep sickness	yes, no	yes = 187, no = 30, sometimes = 4
23	weakness	yes, no	yes = 40, no = 192, sometimes = 0
24	acting out in sleep	yes, no	yes = 54, no = 171, sometimes =2
25	unable to move	yes, no	yes = 38, no = 196, not-known = 1, sometimes = 1
26	Epworth scale	numeric (answers to 8 questions having 0-3 weights)	mild = 82, moderate = 86, severe = 73 (mild =>0-5, moderate => 5-15, severe => 15-24)
27	headache	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=77, 1=54, 2=58, 3=51]
28	walking in sleep	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=182, 1=20, 2=19, 3=12]
29	suffer from nightmares	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=72, 1=60, 2=53, 3=56]
30	paralysis	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=198, 1=25, 2=8, 3=7]
31	crawly feeling in legs	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=142, 1=33, 2=29, 3=29]
32	pain in legs	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=119, 1=49, 2=38, 3=31]
33	Habit of working from bed	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=206, 1=14, 2=12, 3=7]
34	Habit of eating in bed	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=197, 1=24, 2=10, 3=10]
35	Worry a lot in bed	nominal values (0,1,2,3)	nominal values (0,1,2,3) [0=57, 1=69, 2=60, 3=54]
36	poor sleeping habits	yes, no	yes = 129, no =3
37	poor eating habits	yes, no	yes = 75, no = 4
38	exercise/ week	descriptive explanation	0 = 139, 1-2 days = 23, 3-5 days = 51, 6-7 days = 21, stopped exercising recently = 2

Attribute Number	*Attribute Name*	*Attribute-values (before preprocessing)*	*Attribute-values (post-preprocessing)*
39	caffeine intake	descriptive explanation	0 = 27, occasional = 7, 1-3 cups/day = 128, 3-6 cups/day = 62, 6 or more cups/day = 15, yes = 1
40	alcohol intake	descriptive explanation	0 = 131, moderate – 2 to 4 glasses/day = 8, normal – less than 2 glasses/day = 95, high – more than 4 glasses/day = 2, yes = 2
41	Smoking (no. of packs/day)	descriptive explanation	quit smoking long back = 1, quit recently = 3, no = 188, 1-2 packs/day = 30, less than 1 pack/day = 15, 2-4 packs/day = 1, attempting to quit = 1, yes = 1
42	depression	numeric (answers to 8 questions having 0-3 weights)	mild (0 to 15) = 116, moderate (15 to 35) = 119, severe (35 or more) = 7
43	study-type	allnight , allnight-for followup CPAP, splitnight, allnight BiPAP, allnight-CPAP, allnight followed by MSLT, splitnight with CPAP & BiPAP, splitnight-no CPAP, splitnight with CPAP, allnight with CPAP & BiPAP, followup CPAP & BiPAP, CPAP, followup CPAP	allnight study = 156, allnight study for followup CPAP = 3, splitnight = 44, allnight with BiPAP from beginning = 1, allnight with CPAP = 22, allnight followed by MSLT = 1, splitnight with CPAP & BiPAP = 1, splitnight ordered however no CPAP = 1, splitnight with CPAP = 2, allnight with CPAP & BiPAP = 2, followup CPAP & BiPAP = 1, CPAP, followup CPAP = 3
44	Total hours in bed	numeric	0-5 hours = 3, 5-8 hours = 226, 8 or more hours = 11
45	Total hours in sleep	numeric	0-5 hours = 50, 5-8 hours = 186, 8 or more hours = 3
46	sleep efficiency	numeric	low (less than 70%) = 39, mild (80 to 90%) = 91, moderate (70-80%) = 23, normal (90 to 100%) = 86
47	arousal-index	numeric	mild (less than 10 arousals/hour) = 114, moderate (10-20 arousals/hour) = 71, severe (more than 20 arousals/hour) = 35
48	sleep latency till stage 1	numeric	short (less than 5 minutes) = 61, normal (5 to 18 minutes) = 90, long (18 or more minutes) = 64
49	sleep latency till stage REM	numeric	low (less than 60 minutes) = 24, mildly-reduced (60 to 70 minutes) = 11, normal (70 to 100 minutes) = 30, prolonged (100 or more minutes) = 95

Attribute Number	*Attribute Name*	*Attribute-values (before preprocessing)*	*Attribute-values (post-preprocessing)*
50	percent time in stage 1	numeric	reduced (less than 7%) = 50, normal (7 to 10%) = 47, prolonged (greater than 10%) = 55
51	percent time in stage 2	numeric	reduced (less than 50%) = 121, normal (50-60%) = 31, prolonged (60% or more) = 16
52	percent time in stage REM	numeric	reduced (less than 22%) = 173, normal (22 to 27%) = 11, prolonged (27% or more) = 4
53	percent time in slow-wave sleep	numeric	reduced (less than 7%) = 30, normal (7 to 22%) = 60, prolonged (22% or more) = 56
54	mean oxygen potential in blood	numeric	low (less than 85%) = 1, mild-decrease (86% to 92%) = 50, normal (92% or more) = 121
55	average heart rate in beats/minute	numeric	very-low (less than 40 beats/minute) = 0, mildly- reduced (40 to 60 beats/minute) = 22, normal (60 to 100 beats/minute) = 154
56	heart disorder	bradycardia, occasional premature ventricular contractions, sinus tachycardia, intermittent tachycardia	badycardia = 5, occasional premature ventricular contractions = 1, sinus tachycardia = 2, intermittent tachycardia = 1
57	apnea-hypopnea index	numeric	mild (less than 10) = 94, moderate (10 to 20) = 46, severe (20 or greater) = 47
58	apnea-hypopnea index while lying on back	numeric	mild (less than 10) = 9, moderate (10 to 20) = 19, severe (20 or greater) = 36
59	Apnea-hypopnea index while lying on sides	numeric	mild (less than 10) = 27, moderate (10 to 20) = 13, severe (20 or greater) = 6
60	obstructive sleep apnea	no, insignificant, mild, mild to moderate, moderately-severe, severe, apnea removed by CPAP	no = 2, insignificant = 2, mild = 47, mild to moderate = 14, moderately-severe = 50, severe = 23, apnea removed by CPAP = 1
61	clinical evaluation of snoring	no, mild, mild to moderate, moderate, heavy, snoring removed by CPAP	no = 2, mild = 5, mild to moderate = 1, moderate = 11, heavy = 42, snoring removed by CPAP = 2
62	central sleep apnea	no, insignificant, mild, severe	no = 2, insignificant = 1, mild = 1, severe = 2
63	periodic leg movement syndrome	no, infrequent, moderately-severe, severe	no = 1, infrequent = 146, moderately-severe = 48, severe = 31

Table 5.1: Macro-level dataset before and after pre-processing

The following table shows a miniature snapshot of the macro-level dataset, each row corresponds to a patient. For the sake of illustration we show only six attributes of the 63 that form part of the macro-level dataset. From the original set of 70 attributes pre-processing reduced the number of attributes to 63. The dropped attributes include the duration of bradycardia condition, percentage wake time before sleep, minimum oxygen potential, height and weight of patients and a listing of medications. The attributes were dropped either because their distribution was very sparse as in the case of bradycardia or because of data transformation, which made some of the attributes redundant. For instance, the attributes of height and weight were replaced after being used to generate a new, more informative attribute called Body-mass index (BMI).

Instance	**Gender**	**Body-mass index**	**Epworth**	**Exercise/week**	**Sleep-efficiency**	**Obstructive sleep apnea**	…
1	male	underweight	20	1-2 days	mild	insignificant	…
2	female	normal	13	3-5 days	moderate	moderate	…
3	female	obese	22	0	low	severe	…
..	….	…	…	…	…	…	…
…	…	…	…	…	…	…	…
N	male	overweight	17	6-7 days	normal	no	…

Table 5.2: Miniature snapshot of macro-level dataset.

5.3 Analysis of macro-level data

In this section, we analyze the results of experiments performed on the macro-level dataset.

Experiment-1:

Objective:

Determine association between the percent time spent by patients in the different sleep stages.

Input parameters:

Instances:	242 instances (Each instance corresponds to a patient)
Total number of attributes:	8
Attribute names:	sleep-eff (sleep efficiency),
	arousal-index (arousal-index),

sleep-latency-till-1 (sleep latency till stage 1),
sleep-latency-till-REM (sleep latency till stage REM),
stage1 (percentage time in stage1),
stage2 (percentage time in stage2),
REM (percentage time in REM),
slow-wave (percentage time in slow-wave sleep).

Minimum number of rules sought: 10
Minimum confidence: 70%
Upper bound for support: 100%
Lower bound for support: 10%

All rules obtaining from Association rule mining over the full training set are provided below:

1. sleep-latency-till-REM = prolonged *AND* stage2 = reduced ➔ REM = reduced
   [Confidence = 96.07%; Support = 20.25%; Lift = 1.3439; Chi-square = 19.1683; P< 0.001]
2. stage1 = prolonged ➔ REM = reduced
   [Confidence = 92.73%; Support = 21.07%; Lift = 1.2971; Chi-square = 15.7493; P<0.001]
3. slow-wave = prolonged ➔ REM = reduced
   [Confidence= 87.5% ; Support = 20.25%; Lift = 1.2239; Chi-square = 9.1614; 0.001<P<0.01]
4. stage2 = reduced ➔ REM = reduced
   [Confidence = 86.78% ; Support = 43.39%; Lift =1.2139; Chi-square = 27.7623; P<0.001]
5. slow-wave = normal ➔ REM = reduced
   [Confidence=86.67% ; Support = 21.49%; Lift = 1.2123; Chi-square = 9.0185; 0.001<P<0.01]
6. arousal-index = mild *AND* stage2 = reduced ➔ REM = reduced
   [Confidence=84.06% ; Support = 23.97%; Lift = 1.1758; Chi-square = 7.4818; 0.001<P<0.01]
7. sleep-latency-till-REM = prolonged ➔ REM = reduced
   [Confidence=81.05% ; Support = 31.82%; Lift = 1.1338; Chi-square = 7.0201; 0.001<P<0.01]
8. sleep-latency-till-1 = long ➔ REM = reduced
   [Confidence = 79.69% ; Support = 21.07%; Lift = 1.1147; Chi-square = 2.8695; P>0.05]
9. arousal-index = mild *AND* REM = reduced ➔ stage2 = reduced
   [Confidence = 74.36% ; Support = 23.97%; Lift = 1.4872; Chi-square = 27.3246; P<0.001]
10. arousal-index = moderate ➔ REM = reduced
    [Confidence = 73.24% ; Support = 21.49%; Lift = 1.0245; Chi-square = 0.1512; P>0.05]

Analysis of Rules:

Rules 1 and 4 clearly indicate that suppression in stage 2 sleep causes a corresponding reduction in the percentage time spent by patients in the REM stage. Also a symmetric association is indicated by rule 9.

Rules 1 and 7 lead us to deduce that increase in sleep latency till REM brings about a decrease in percentage time spent in REM stage. Similarly, rule 2 indicates that prolonged duration of stage 1 of sleep leads to suppression of stage REM. Rules 1, 2, 4 and 9 pass the test of statistic robustness since there is a less than 1 in 1000^{th} probability that the correlations they describe occur by chance. In the case of rule 7, there is almost a 10-fold increase in the likelihood that the correlation it describes occurs by chance. Hence it is less statistically robust. However, we still consider the rule to be statistically significant.

All the rules have good confidence values above 73%. Their support count averages around 20%, which makes these rules interesting.

From a medical perspective, the above set of rules validates the fundamental principles used for diagnosis and treatment of sleep disorders.

Experiment-2:

Objective:

For the same set of attributes as those considered in Experiment-1, we try to analyze the effect that sleep stages and related parameters have on sleep- efficiency.

Experiment-2a:

Input Parameters:

Instances:	242 instances (Each instance corresponds to a patient)
Total number of attributes:	8
Attribute names:	sleep-eff (sleep efficiency), arousal-index (arousal-index), sleep-latency-till-1 (sleep latency till stage 1), sleep-latency-till-REM (sleep latency till stage REM), stage1 (percentage time in stage1), stage2 (percentage time in stage2), REM (percentage time in REM), slow-wave (percentage time in slow-wave sleep).

Minimum number of rules sought: 20
Minimum confidence: 70%
Upper bound for support: 100%
Lower bound for support: 1%
Required attributes in the antecedents: None
Required attributes in the consequents: sleep-eff

All rules obtaining from Association rule mining over the full training set are provided below:

1. arousal-index = mild *AND* sleep-latency-till-1 = short ➔ sleep-eff = normal

[Confidence = 73.08% ; Support = 7.85; Lift: 2.0479; Chi-square = 17.9276; P<0.001]

2. sleep-latency-till-1 = short ➔ sleep-eff = normal

[Confidence = 70.49% ; Support = 17.77; Lift = 1.9754; Chi-square = 43.5034; P<0.001]

Analysis of Rules:

In this experiment, we forced the consequent of the rules to have the attribute-sleep efficiency. However at higher confidence levels (above 70%), we do not find rules indicating the effect of sleep stage variations on sleep efficiency. Instead, we obtain rules suggesting that shortened sleep latency till stage 1 does not affect sleep efficiency. Similarly, shortened latency to stage 1 of sleep and mild arousal index also do not seem to affect sleep efficiency. The sleep efficiency in both cases remains within the normal range. Since no rules with sleep stage variations were found we lower the threshold confidence to 40%. Experiment 2b, with the modified input parameter settings is indicated below. Both the rules represent true associations since they are found to be statistically robust.

Experiment-2b:

Input Parameters:

Instances: 242 instances (Each instance corresponds to a patient)
Total number of attributes: 8
Attribute names: sleep-eff (sleep efficiency),
arousal-index (arousal-index),

sleep-latency-till-1 (sleep latency till stage 1),
sleep-latency-till-REM (sleep latency till stage REM),
stage1 (percentage time in stage1),
stage2 (percentage time in stage2),
REM (percentage time in REM),
slow-wave (percentage time in slow-wave sleep).

Minimum number of rules sought: 50
Minimum confidence: 40%
Upper bound for support: 100%
Lower bound for support: 1%
Required attributes in the antecedents: None
Required attributes in the consequents: sleep-eff

Rules obtaining from Association rule mining over the full training set are provided below: (We provide a collection of statistically robust rules, with P-values less than 0.001).

1. sleep-latency-till-1 = long *AND* stage2 = reduced ➔ sleep-eff = low
   [Confidence = 50% ; Support = 7.4%; Lift= 3.1026; Chi-square = 35.93; P<0.001]
2. arousal-index = mild *AND* sleep-latency-till-1 = short ➔ sleep-eff = normal
   [Confidence = 73.08%; Support = 7.85%; Lift= 2.0564; Chi-square = 17.9176; P<0.001]
3. sleep-latency-till-1 = short ➔ sleep-eff = normal
   [Confidence = 70.49%; Support = 17.77%; Lift = 1.9836; Chi-square = 43.5034; P<0.001]
4. sleep-latency-till-1 = short *AND* REM = reduced ➔ sleep-eff = normal
   [Confidence = 63.41%; Support = 10.74%; Lift = 1.7845; Chi-square = 16.7398; P<0.001]
5. arousal-index = mild *AND* sleep-latency-till-REM = prolonged *AND* stage2 = reduced ➔ sleep-eff = mild *AND* REM = reduced
   [Confidence = 53.57%; Support = 6.2%; Lift = 2.0578; Chi-square = 12.469; P<0.001]
6. sleep-latency-till-1 = long *AND* stage2 = reduced *AND* REM = reduced ➔ sleep-eff = low
   [Confidence = 50%; Support = 7.02%; Lift = 3.1026; Chi-square = 33.5708; P<0.001]
7. slow-wave = reduced ➔ sleep-eff = normal *AND* REM = reduced
   [Confidence = 50%; Support = 6.2%; Lift = 2.283; Chi-square = 15.8068; P<0.001]
8. sleep-latency-till-1 = long *AND* stage2 = reduced ➔ sleep-eff = low *AND* REM =reduced
   [Confidence = 47.22%; Support = 7.02%; Lift = 3.463; Chi-square = 40.4757; P<0.001]
9. arousal-index = mild *AND* sleep-latency-till-1 = long ➔ sleep-eff = low
   [Confidence = 46.43%; Support = 5.37%; Lift = 2.881; Chi-square = 21.5232; P<0.001]
10. arousal-index = mild *AND* sleep-latency-till-1 = normal *AND* REM = reduced ➔ sleep-eff = mild

AND stage2 = reduced

[Confidence = 43.33%; Support = 5.37%; Lift = 2.33; Chi-square = 13.8397; P<0.001]

11. sleep-latency-till-1 = long *AND* REM = reduced ➔ sleep-eff = low

[Confidence = 43.14%; Support = 9.09%; Lift = 2.6767; Chi-square = 34.8962; P<0.001]

12. sleep-latency-till-1 = short ➔ sleep-eff = normal *AND* REM = reduced

[Confidence = 42.62%; Support = 10.74%; Lift = 1.9462; Chi-square = 20.4662; P<0.001]

13. sleep-latency-till-1 = long ➔ sleep-eff = low

[Confidence = 42.19%; Support = 11.16%; Lift = 2.6178; Chi-square = 43.7668; P<0.001]

Anaysis of Rules:

By going through the set of 13 most significant rules (P-values less than 0.001), the following trends can be noticed. If the latency to stage-1 of sleep exceeds the normal range in time, sleep efficiency gets greatly hampered. However, when stage-1 is reached earlier than the normal time (short latency), sleep efficiency does not get affected. It is interesting to note the difference in the sleep efficiency perception when stage-REM is reached well past its expected duration of presence (i.e., latency is prolonged). Prolonged latency to stage-REM only causes a mild reduction in sleep efficiency. This indicates that prolonged latency to stage-1 has a comparatively more severe effect on sleep efficiency as compared to prolonged latency to stage-REM.

From the rules obtained in Experiments 2a and 2b the following correlations stand out as being medically interesting besides being statistically significant,

Latency-till-1 = reduced ➔ sleep –eff = normal,
Latency-till-1 = prolonged ➔ sleep-eff = low
Latency-till-REM = prolonged ➔ sleep-eff = mildly-reduced

Experiment-3:

Objective:

Influence of heart rate and oxygen potential variations on sleep stages.

Input Parameters:

Instances:	242 instances (Each instance corresponds to a patient)
Total number of attributes:	10
Attribute names:	sleep-eff (sleep efficiency),

arousal-index (arousal-index),
sleep-latency-till-1 (sleep latency till stage 1),
sleep-latency-till-REM (sleep latency till stage REM),
stage1 (percentage time in stage1),
stage2 (percentage time in stage2),
REM (percentage time in REM),
slow-wave (percentage time in slow-wave sleep),
oxy-mean (mean oxygen potential in blood),
heart-rate (mean heart rate in beats/minute).

Minimum number of rules sought: 20
Minimum confidence: 70%
Upper bound for support: 100%
Lower bound for support: 10%
Required attributes in the antecedents: none
Required attributes in the consequents: none

Rules obtaining from Association rule mining over the full training set are provided below: (We provide a collection of statistically robust rules, instead of all 20).

1. sleep-latency-till-REM = prolonged *AND* stage-2 = reduced ➔ REM = reduced
   [Confidence = 96.08%; Support = 20.25%; Lift = 1.3439; Chi-square = 19.1683; P<0.001]
2. stage-1 = prolonged ➔ REM = reduced
   [Confidence = 92.73%; Support = 21.07%; Lift = 1.2971; Chi-square = 15.7493; P<0.001]
3. stage-2 = reduced *AND* oxy-mean = normal ➔ REM =reduced
   [Confidence = 89.06%; Support = 23.55%; Lift = 1.2458; Chi-square = 13.1787; P<0.001]
4. stage-2 = reduced *AND* heart-rate = normal ➔ REM = reduced
   [Confidence = 87.01%; Support = 27.69%; Lift = 1.2172; Chi-square = 13.3589; P<0.001]
5. stage-2 = reduced ➔ REM = reduced
   [Confidence = 86.78%; Support = 43.49%; Lift = 1.2139; Chi-square = 27.7623; P<0.001]
6. REM = reduced *AND* oxy-mean = normal ➔ heart-rate = normal
   [Confidence = 81.61%; Support = 29.34%; Lift = 1.2824; Chi-square = 18.9597; P<0.001]
7. stage-2 = reduced *AND* oxy-mean = normal ➔ heart-rate = normal
   [Confidence = 81.25%; Support = 21.49%; Lift = 1.2768; Chi-square = 11.688; P<0.001]
8. oxy-mean = normal ➔ heart-rate = normal
   [Confidence = 79.34%; Support = 39.67%; Lift = 1.2468; Chi-square = 25.7939; P<0.001]
9. arousal-index = mild *AND* heart-rate = normal ➔ oxy-mean = normal

[Confidence = 78.13%; Support = 20.66%; Lift = 1.5625; Chi-square = 27.5277; P<0.001]

10. arousal-index = mild *AND* REM = reduced ➔ stage-2 = reduced

[Confidence = 74.36%; Support = 23.97%; Lift = 1.4872; Chi-square = 27.3256; P<0.001]

Analysis of Rules:

The rules above indicate that variation in the time spent in the different sleep stages has no affect on mean heart rate or oxygen potential in the blood. In other words, the oxygen potential or heart rate remain within the normal range despite variations in the time spent in the different sleep stages. We ran further experiments with minimum confidence as low as 30% and lower bound for support as low as 1%. However, we were unable to obtain associations with heart rate or oxygen potential deflecting out of the normal zone. This leads us to the conclusion that oxygen potential and heart rate are independent of sleep stage variations.

Further, we notice very interesting rules that can potentially be applied to predicting reduced presence of stage-REM in sleep. We notice that reduced stage-2 of sleep has a corresponding influence on stage-REM. In other words, reduced stage-2 of sleep causes a reduction in percentage time spent in stage-REM during sleep. Also, the prolonged presence of stage-1 (beyond the normal duration) leads to a much-reduced presence of stage-REM in sleep. Mathematically, this can be represented as,

time(stage-REM) α *time*(stage-2), ----------(1)

time(stage-REM) α 1 / *time*(stage-1)

Experiment-4:

Objective:

To determine whether there is a relation between Obstructive Sleep Apnea (OSA) and Periodic Leg Movement Syndrome (PLMS).

Input Parameters:

Instances:	242 instances (Each instance corresponds to a patient)
Total number of attributes:	3
Attribute names:	epworth(Epworth index reading),
	OSA (Obstructive sleep apnea diagnosis),

PLMS (Periodic leg movement syndrome diagnosis).

Minimum number of rules sought: 10

Minimum confidence: 50%

Upper bound for support: 100%

Lower bound for support: 1%

Required attributes in the antecedents: OSA

Required attributes in the consequents: PLMS

All rules obtaining from Association rule mining over the full training set are provided below:

1. OSA = severe ➔ PLMS = infrequent

[Confidence = 86.96%; Support = 8.26%; Lift = 1.4413; Chi-square = 7.5279; 0.001<P<0.01]

2. OSA = moderately-severe ➔ PLMS = infrequent

[Confidence = 58% ; Support = 11.98%; Lift = 0.9614; Chi-square = 0.1427; P>0.05]

3. OSA = mild ➔ PLMS =infrequent

[Confidence = 57.45%; Support = 11.16%; Lift = 0.9522; Chi-square = 0.2028; P>0.05]

Analysis of Rules:

The three rules clearly indicate that presence of variable degrees of OSA is independent of PLMS severity. However, we generate no rules when we try to investigate presence of OSA under variable degrees of PLMS. Also, we notice that the rules 2 and 3 are not statistically significant owing to very low chi-square values.

Experiment-5:

Objective:

Relation among depression, Obstructive Sleep Apnea (OSA) and Periodic Leg Movement Syndrome (PLMS).

Input Parameters:

Instances: 242 instances (Each instance corresponds to a patient)

Total number of attributes: 3

Attribute names: depression(Depression index reading),
OSA (Obstructive sleep apnea diagnosis),
PLMS (Periodic leg movement syndrome diagnosis).

Minimum number of rules sought: 10

Minimum confidence: 50%

Upper bound for support: 100%

Lower bound for support: 1%

Required attributes in the antecedents: depression

Required attributes in the consequents: none

Rules obtaining from Association rule mining over the full training set are provided below:

1. depression = moderate *AND* OSA= moderately-severe ➔ PLMS =infrequent
   [Confidence = 65.22%; Support = 6.2%; Lift = 1.081; Chi-square = 0.2536; P>0.05]
2. depression = moderate ➔ PLMS = infrequent
   [Confidence = 62.18%; Support = 30.58%; Lift = 1.0307; Chi-square = 0.3356; P>0.05]
3. depression = mild ➔ PLMS = infrequent
   [Confidence = 58.62%; Support = 28.1%; Lift = 0.9717; Chi-square = 0.2714; P>0.05]
4. depression = mild *AND* OSA = mild ➔ PLMS= infrequent
   [Confidence = 58.33%; Support = 5.78%; Lift = 0.9669; Chi-square = 0.0444; P>0.05]
5. depression = mild *AND* OSA = moderately-severe ➔ PLMS = infrequent
   [Confidence = 54.17%; Support = 5.37%; Lift = 0.8978; Chi-square = 0.4231; P>0.05]

Analysis of Rules:

The above set of rules indicates that depression is not a cause for PLMS. Mild to moderate depression does not result in increase in PLMS. However, these set of rules do not pass the chi-square test for validating their statistical significance.

Experiment –6:

Objective:

Relation among Epworth and Obstructive Sleep Apnea (OSA), Periodic Leg Movement Systems (PLMS)

Input Parameters:

Instances:	242 instances (Each instance corresponds to a patient)
Total number of attributes:	3
Attribute names:	epworth(Epworth index reading), OSA (Obstructive sleep apnea diagnosis), PLMS (Periodic leg movement syndrome diagnosis).
Minimum number of rules sought:	10
Minimum confidence:	50%

Upper bound for support: 100%

Lower bound for support: 1%

Required attributes in the antecedents: epworth

Required attributes in the consequents: none

All rules obtaining from Association rule mining over the full training set are provided below:

1. epworth = severe ➔ PLMS = infrequent
   [Confidence = 63.01%; Support = 19.01%; Lift = 1.0445; Chi-square = 0.3146; P>0.05]
2. epworth = moderate ➔ PLMS = infrequent
   [Confidence = 62.79%; Support = 22.31%; Lift = 1.0408; Chi-square = 0.3376; P>0.05]
3. epworth = mild ➔ PLMS = infrequent
   [Confidence = 54.88%; Support = 18.60%; Lift = 0.9096; Chi-square = 1.541; P>0.05]

Analysis of Rules:

The above rules indicate that Epworth score provides no indication of PLMS. We do not obtain rules providing relationship between Epworth scale and OSA for confidence levels as low as 50% and lower bound minimum support of 1%. These set of rules too are not statistically robust owing to low chi-square values.

Experiment-7:

Objective:

Determine relation among BMI, collar-size, snoring and other factors

Input Parameters:

Instances: 242 instances (Each instance corresponds to a patient)

Total number of attributes: 5

Attribute names: bmi-class (discretized BMI),
collar-size (discretized collar-size),
gasping (gasping for breath),
snore,
depression (Depression index score).

Minimum number of rules sought: 20

Minimum confidence: 70%

Upper bound for support: 100%

Lower bound for support: 5%

Required attributes in the antecedents: none

Required attributes in the consequents: none

Rules obtaining from association rule mining over the full training set are provided below: (We provide a collection of the more relevant rules, instead of all 20).

1. bmi-class = obese *AND* collar-size = 16-18 ➔ snore = yes
   [Confidence = 100%; Support = 12.81%; Lift = 1.2040; Chi-square = 7.2532; 0.001<P<0.01]
2. bmi-class = obese *AND* gasping = yes ➔ snore = yes
   [Confidence = 95.92%; Support = 19.42%; Lift = 1.1548; Chi-square = 7.2191; 0.001<P<0.01]
3. gasping = yes ➔ snore = yes
   [Confidence = 93.18%; Support = 33.88%; Lift = 1.1219; Chi-square = 10.0705; 0.001<P<0.01]
4. bmi-class = obese ➔ snore =yes
   [Confidence = 91.67%; Support = 40.91%; Lift = 1.1036; Chi-square = 10.2674; 0.001<P<0.01]

Analysis of Rules:

From the rules, we are able to determine a very strong association (high confidence) between body-mass index and snoring. These set of rules validate the fact known in the medical community that patients falling under the overweight and obese categories of body-mass index complain of moderate to heavy snoring. Their statistical robustness is also good since the probability that the correlations take place by chance is less than 1 in 100.

5.4 Summary of results

Macro-level dataset analysis produced several interesting rules. Some of these rules confirmed medically known facts. For instance, we obtained rules stating that prolonged latency to stage-1 of sleep coupled with reduction in stage-2 would result in low sleep efficiency. It is interesting however, that while the time spent in stage-2 of sleep is proportional to time spent in REM, the time spent in stage-1 of sleep is inversely proportional to time spent in REM. Also interesting is the effect that latency to reach stage-1 and REM of sleep has on sleep-efficiency. While reduction in latency to stage-1 of sleep does not affect sleep efficiency, prolongation of this latency can significantly lower sleep efficiency. This observation validates a well-known medical fact. However,

what is interesting is that a similar increase in the latency time to reach stage-REM does not cripple the sleep efficiency in the same way as latency to reach stage-1 does. Another interesting set of rules was obtained when we attempted to find correlations between the different sleep disorders. We found through experimental results that Obstructive sleep apnea and Periodic leg movements are highly independent disorders with no correlation between them. We also confirmed a few other medically known facts by obtaining rules indicating that people suffering from overweight problems and having trouble breathing are more susceptible to snoring.

Chapter 6

Micro-level: Dataset Description and Analysis

As the name of the chapter suggests, this dataset aims at providing a better insight into the microstructure of sleep by analyzing some of the physiological changes occurring in the human body. Real-time measurement of the physiological changes is captured every 30 seconds during the patient's night under observation during the sleep study. This 30 second window of measurement is called an *epoch*. The micro-level dataset includes primarily, details on the stage of sleep, the oxygen saturation in the blood, observation of events like an arousal, leg movement or variation in body position during an epoch and miscellaneous notes made by the technician during the patient's period under observation. On an average 700-900 epoch measurements are recorded for every patient constituting their whole night sleep. Thus, this turns out to be a large time sequence data set from which a lot of knowledge can potentially be unearthed. Data collected from 120 patients who underwent a sleep study at Day Kimball Hospital, Putnam, CT constitutes this micro-level dataset.

6.1 Dataset details

The sleep laboratory at Day Kimball Hospital uses the product developed by SensorMedics® to record and save the observations from the sleep study in a suitable storage device (normally the system's hard drive). The product is called the Alpha Sleep system. The data saved in the storage device constitutes the actual, raw, non-staged polysomnogram measurements with noise and artifact present. Following the storage of the raw data, technicians attempt to classify them into different sleep stages using the R&K model of sleep stage classification. Since R&K model is a system for visually scoring stages of sleep, the interpretation as to which sleep stage the patient belongs to in a given epoch may vary from technician to technician. The different polysomnogram

readings and observations made by technicians during the night of the study are analyzed together in order to decide the stage of sleep during an epoch. The process of sleep stage classification therefore becomes tedious and time consuming. The software provides the technicians with tools to visualize the multichannel output of polysomnogram data that helps in sleep stage classification. The raw data along with the sleep stage information forms the micro-level dataset that is analyzed as part of this project.

The micro-level dataset prior to pre-processing comprised of six time sequence attributes that were recorded or deduced for every epoch. The six attributes were,

a) Sleep stage,
b) The pressure generated by the water column during ventilation in CPAP (Continuous Positive Airway Pressure) and BiPAP (Bi-level Positive Airway Pressure),
c) Oxygen potential in the blood,
d) Heart rate in beats/minute,
e) Observation of an event such as an apneaic episode, desaturation or others during an epoch, and
f) Body position of the patient (i.e., whether the patient is sleeping on his/her back, on his/her side or in a prone position).

The diagram on the following page, illustrates the variations in polysomnogram recordings.

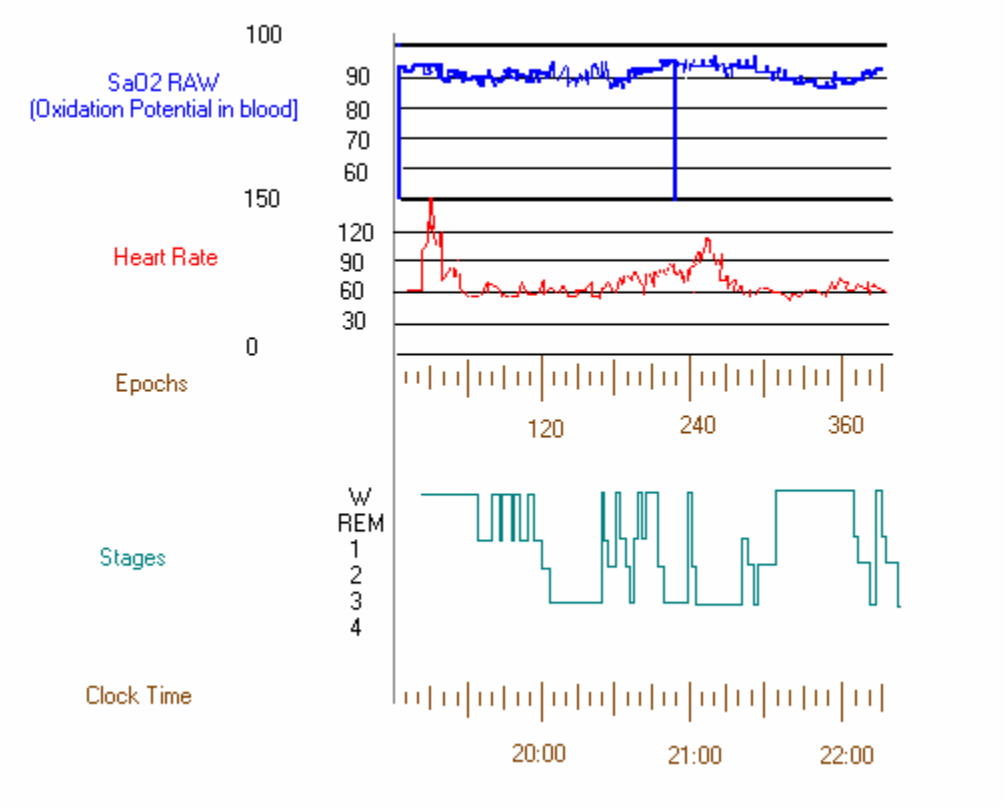


Fig. 6.1: Variations in recordings from the polysomnogram instruments.

6.2 Pre-processing of micro-level dataset

The main objective of pre-processing is to identify events of interest in the long sequences that form part of the micro-level dataset. In order to identify the events of interest we write suitable filters. The filters are responsible for identifying the various events and modifying the epoch-by-epoch time sequence micro-level dataset into a dataset containing event-related sequences. The numeric values of a few time-sequence attributes are discretized into nominal-valued blocks. Thus, the nominal-valued blocks represent the range of numeric values.

Pre-processing, in this case, leads to data compression or data reduction. The reason being, the pre-processed data is no longer represented by the 700-900 data points

with the measured values during every epoch, but instead by the sequences made up of events of interest.

Every attribute forming part of the dataset has events of interest. The different events of interest for the attributes are as listed in the table below.

Attribute Names	***Events of interest***
Sleep stages	Wake stage, Stage1, Stage2, Stage3, Stage4, REM, no sleep stage
Oxygen potential in blood	Normal oxygen potential, mildly-reduced potential and low oxygen potential
Heart rate in beats/minute	Normal heart rate, mildly-reduced heart rate and low heart rate
Observed events during patient's sleep	Any event leading to sleep fragmentation e.g., occurrence of apnea, hypopnea event etc.
Body position during sleep	Back, prone, right side or left side

Table 6.1: Attribute values and their events of interest

The normal, mildly reduced and low threshold values for oxygen potential and heart rate are on the basis of standard cut-offs followed in the medical community. The ranges for these values are given in the following table.

Oxygen Potential ranges	***Event values***
92-100 %	normal
86-92%	Mildly-reduced
Less than 86%	low

Table 6.2: Events based on oxygen potential value ranges

Heart rate ranges	***Event values***
60-100 beats/minute	Normal
40-60 beats/minute	Mildly-reduced
less than 40 beats/minute	Low

Table 6.3: Events based on heart rate value ranges

The filters go over the dataset to identify the events of interest as listed in the tables 6.1, 6.2 and 6.3. The epoch times during which the event was observed is recorded in the modified dataset.

6.3 Dataset organizations

For this level we organize the dataset in two different ways as shown below.

6.3.1 Epoch-based data organization

Since the polysomnogram signals that are analyzed in this level are recorded every epoch, we organize the dataset in this fashion. Every instance represents an epoch of recording during the patient's night of sleep in the laboratory. A snapshot of the dataset is as shown below.

Epoch	**Sleep stage**	**Heart rate**	**Oxygen potential**	**Event of interest**	**Body position**
1	W	normal	mildly-reduced	NA	B
2	1	low	low	OH	L
…	…	…	…	…	…
…	…	…	…	…	…
817	W	normal	normal	Awake	R
1	W	mildly-reduced	mildly-reduced	Desat	P
2	2	normal	normal	OHa	L
…	…	…	…	…	…

Table 6.4: Epoch-based micro-level data organization

As is obvious from the nature of data organization, this dataset has a large number of instances (121,944 instances). On an average around 800 instances represent information from a single patient.

6.3.2 Patient-based data organization

This model of data organization is, loosely speaking, the transpose of the Epoch-based data organization model. In this dataset, every instance represents a single patient. The number of instances in this dataset will be consequently small. In fact, this dataset has only 121 instances. The following table provides a snapshot of this data organization.

Patient-ID	Sleep stage	Heart rate	Oxygen potential	Events	Body position
1	{W,W,..,2,3,..,R,..,W}	{low,...,normal,...}	{low,…,low,..}	{Wake,...,Fluct,..}	{B,B,..,L,..,B,..}
2	{W,..,1,2,2,..,R,…}	{low,..,low,…}	{normal,…,low,..}	{OH,…,CA,…}	{B,B,…,R,..}
….	….	…	…	…	…

Table 6.5: Patient-based data organization

6.4 Analysis of micro-level data

Micro-level analysis is performed over both the datasets defined in the previous section viz., dataset obtained from epoch-based data organization and patient-based data organization.

6.4.1 Analysis over Epoch-based data

As mentioned in the earlier sections of this chapter, the numeric values for heart rate and oxygen potential are discretized into ranges acceptable to the medical community. We conducted preliminary experiments on the raw micro-level datset organized as shown in the Table 6.3. Since the dataset comprised only nominal attributes, the basic association rule mining algorithm was employed to generate rules. Below we discuss the results obtained from some of the experiments.

Experiment-1:

Objective:

Determine variation in sleep stage, heart rate, oxygen potential, events and body position in the micro-level arrangement.

Input Parameters:

Instances:	121944 instances (Every instance corresponds to an epoch)
Total number of attributes:	5
Attribute names:	sleepstage, oxy-pot (oxygen potential), heart-rate, event, body-position.
Minimum number of rules sought:	10
Minimum confidence:	70%
Upper bound for support:	100%
Lower bound for support:	10%

Required attributes in the antecedents: none

Required attributes in the consequents: none

All rules obtaining from Association rule mining over the full training set are provided below:

1. body-pos = B ➔ heart-rate = normal [Confidence = 87%]
2. event = NA ➔ heart-rate = normal [Confidence = 85%]
3. oxy-pot = normal *AND* event = no-event-seen ➔ heart-rate = normal [Confidence = 83%]
4. oxy-pot = normal ➔ heart-rate = normal [Confidence = 83%]
5. oxy-pot = normal *AND* heart-rate = normal ➔ event = no-event-seen [Confidence = 79%]
6. oxy-pot = normal ➔ event = no-event-seen [Confidence = 79%]
7. event = no-event-seen ➔ oxy-pot = normal [Confidence = 75%]
8. body-pos = B ➔ event = no-event-seen [Confidence = 73%]
9. heart-rate = normal AND event = no-event-seen ➔ oxy-pot = normal [Confidence = 73%]
10. heart-rate = normal ➔ event = no-event-seen [Confidence = 71%]

Analysis of Results:

The above rules indicate that patients lying on their back have normal heart rate. Also there are rules that indicate that no events occur when the heart rate and oxygen

potential are normal. Since this experiment does not produce rules with the different sleep stages, we try another experiment by reducing the bound for minimum support.

Experiment-2:

Objective:

Determine variation in sleep stage, heart rate, oxygen potential, events of interest and body position in the micro-level arrangement.

Input Parameters:

Instances: 121944 instances (Every instance corresponds to an epoch)

Total number of attributes: 5

Attribute names: sleepstage,
oxy-pot (oxygen potential),
heart-rate,
event,
body-position.

Minimum number of rules sought: 20

Minimum confidence: 90%

Upper bound for support: 100%

Lower bound for support: 1%

Required attributes in the antecedents: none

Required attributes in the consequents: none

Rules obtaining from Association rule mining over the full training set are provided below: (We provide a collection of the more relevant rules, instead of all 20).

1. sleepstage = 1 *AND* oxy-pot = mild-decrease ➔ heart-rate = normal [Confidence = 95%]
2. oxy-pot = mild-decrease *AND* body-pos = B ➔ heart-rate = normal [Confidence = 94%]
3. oxy-pot = low ➔ heart-rate = normal [Confidence = 94%]
4. sleepstage = W ➔ heart-rate = normal [Confidence = 93%]
5. sleepstage = W *AND* oxy-pot = normal ➔ heart-rate = normal [Confidence = 93%]
6. sleepstage = W *AND* body-pos = B ➔ heart-rate = normal [Confidence = 91%]
7. sleepstage = W *AND* oxy-pot = normal *AND* body-pos = B ➔ event = no-event-seen [Confidence = 91%]
8. sleepstage = 2 *AND* oxy-pot = mild-decrease ➔ heart-rate = normal [Confidence = 91%]

Analysis of Results:

The results show that variation in oxygen potential does not affect heart rate. In fact, even when the oxygen potential is much below the normal levels the heart rate does not suffer. The results also indicate that whether in wake state or in stage 1 or 2 the heart rate is normal with a very high degree of confidence associated with those rules.

6.4.2 Analysis over Patient-based data

Our real purpose in performing micro-level analysis was in determining events of interest in the data recorded for each patient. Therefore, we transformed the dataset from one where every instance represented an epoch to a dataset where every instance represented all the epoch records for an individual patient as shown in Table 6.4. The transformation was achieved by writing a suitable filter. The filter also treated every event of interest detected as a unique time sequence attribute. The converted dataset indicated the actual times during which the event of interest was witnessed. The following analyses are based on experiments performed on this dataset.

Experiment-3:

Objective:

Determine associations among oxygen potential, heart rate and body position during sleep.

Input Parameters:

Instances: 121 instances (Every instance represents a single patient in the dataset)

Total number of attributes: 11

Attribute names: low-oxy-pot (low oxygen potential),
mild-reduce-oxy-pot (mildly reduced oxygen potential),
normal-oxy-pot (normal oxygen potential),
very-low-heart-rate (very low heart rate),
mild-reduce-heart-rate (mildly reduced heart rate),
normal-heart-rate (normal heart rate),
body-position-N (non-classified body position),
body-position-B (lying on the back side),
body-position-P (lying on the prone side),

body-position-L (lying on the left side),
body-position-R (lying on the right side),

Minimum number of rules sought: 10

Minimum confidence: 70%

Upper bound for support: 100%

Lower bound for support: 10%

Required attributes in the antecedents: none

Required attributes in the consequents: none

Rules obtaining from Association rule mining over the full training set are provided below: (Few of the 10 rules generated are presented here).

1. body-position-B = 0:1 *AND* body-position-R = 2:3 ➔ body-position-B = 4:5
   [Confidence = 82.85%; Support = 75.21%]
2. low-oxy-pot = 0:1 *AND* body-position-L = 2:3 ➔ body-position-B = 4:5
   [Confidence = 83.25%; Support = 70.25%]
3. body-position-B = 4:5 ➔ mild-reduce-oxy-pot = 0:1 *AND* normal-oxy-pot = 2:3
   [Confidence = 72.85%; Support = 70.25%]

Analysis of Rules:

Rule 2 is interesting since it indicates that when a patient experiences a decrease in oxygen potential, he/she flips over generally to the left side of the body before lying on the back. The following reasoning can medically validate this rule. Sleep apnea is particularly severe for people sleeping on their back. Also, a drop in oxygen potential characterizes sleep apnea. Therefore, following a drop in oxygen potential the patient does not sleep on his back until possibly the optimum oxygen potential levels get restored. Most of the other rules, though, are difficult to validate medically. All that they indicate is a behavior that gets regularly noticed.

Experiment-4:

Objective:

Determine the most frequently occurring variations in sleep stages along with occurrences of low oxygen potential or low heart rate.

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	6
Attribute names:	sleepstage2, sleepstageR (sleep stage REM), low-oxy-pot (low oxygen potential), mild-reduce-oxy-pot (mildly reduced oxygen potential), very-low-heart-rate (very low heart rate), mild-reduce-heart-rate (mildly reduced heart rate).
Minimum number of rules sought:	20
Minimum confidence:	60%
Upper bound for support:	100%
Lower bound for support:	10%

Required attributes in the antecedents: none

Required attributes in the consequents: none

Rules obtaining from Association rule mining over the full training set are provided below: (Few of the 10 rules generated are presented here).

1. sleepstage2 = 0:1 *AND* sleepstageR = 4:5 ➔ sleepstage2 = 2:3
   [Confidence = 88.70%; Support = 93.39%]
2. sleepstage2 = 0:1 *AND* sleepstage2 = 4:5 ➔ sleepstageR = 2:3
   [Confidence = 79.49%; Support = 90.91%]

Analysis of the Rules:

The two rules shown above are a sample of the ten rules that we obtain from this experiment. Rule 2 suggests that stage 2 and stage REM of sleep alternate each other. This alternating behavior is noticed mid-way into whole night sleep. Apart from rules such as Rule 2, it is extremely difficult to validate patterns detected by the rules. Also, since the rules identify patterns in relative time, it is not possible to pin point during which epochs these patterns were noticed. Though the two rules shown above have the same relative times to identify different events in the pattern, the real times when these events were noticed would have been different.

6.4.3 Analysis of micro-level data using window-based association rule mining

To obtain rules with more meaning and insight, we perform further transformations on the dataset. The main bottleneck we identify in the previous set of rules that makes analysis and validation difficult is the presence of relative time. The filters we wrote were therefore intended to preserve the actual epoch durations during which the events of interest were noticed.

We ran fresh experiments to see the benefit of organizing data in the manner described above. Below, we provide some of the results obtained from the experiments.

Experiment-5:

Objective:

Determine associations among the REM, stage-2 and wake sleep stages and periods where below normal oxygen potential or heart rate observed

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	7
Attribute names:	sleepstageW, sleepstage2, sleepstageR (sleep stage REM), low-oxy (low oxygen potential), reduce-oxy (reduced oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate).
Minimum number of rules sought:	10
Minimum confidence:	30%
Upper bound for support:	40%
Lower bound for support:	3%
Required attributes in the antecedents: none	
Required attributes in the consequents: none	
Epoch-window:	50

All rules obtaining from Association rule mining over the full training set are provided below:

1. low-oxy (0:50) ➔ low-hrate (0:50) [Confidence: 100%; Support: 5.79%]

Analysis of Rule:

Only one rule was obtained from the mining process. This rule is informative since it indicates that a reduced oxygen potential and reduced heart rate both seem to occur during the initial 50 epochs (i.e., 25 minutes) of sleep.

Experiment-6:

Objective:

Determine variations in oxygen potentials and heart rates and find associations with the stage of sleep that is not be classified.

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	5
Attribute names:	stageN (non-classified sleep stage), low-oxy (low oxygen potential), reduce-oxy (reduced oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate).
Minimum number of rules sought:	10
Minimum confidence:	30%
Upper bound for support:	40%
Lower bound for support:	3%
Required attributes in the antecedents: none	
Required attributes in the consequents: none	
Epoch-window:	50

All rules obtaining from Association rule mining over the full training set are provided below:

1. low-oxy (0:50) ➔ low-hrate (0:50) [Confidence: 100%; Support: 5.79%]
2. stage-N(0:50) *AND* stage-N (850:900) ➔ low-hrate(0:50)
   [Confidence: 93.3%; Support: 11.57%]

Analysis of Rules:

Rule 1 was also obtained from the previous experiment. Rule 2 indicates, that during the initial and towards the end of sleep (900 epochs = 7.5 hours) the technicians find it hard to classify the sleep into a specific stage. Further during the initial periods of sleep when stage-N sleep is witnessed a decrease in heart rate is also seen.

Experiment-7:

Objective:

Find associations between items of interest in a reduced epoch window size of 10.

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	18
Attribute names:	stageN (non-classified sleep stage), stageW, stage1, stage2, stage3, stage4, stage-R, low-oxy (low oxygen potential), reduce-oxy (reduced oxygen potential), normal-oxy (normal oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate), normal-hrate (normal heart rate), body-N (non-classified body sleep position), body-B (back side sleep position), body-P (front side sleep position), body-L (left side sleep position), body-R (right side sleep position).
Minimum number of rules sought:	10
Minimum confidence:	70%
Upper bound for support:	50%
Lower bound for support:	1%

Required attributes in the antecedents: none
Required attributes in the consequents: none
Epoch-window: 10

Sample of the rules obtaining from Association rule mining over the full training set are provided below:

1. normal-hrate (0:940) ➔ low-hrate (0:10) [Confidence: 100%; Support: 14.28%]

2. stage3(70:150) *AND* stage3(220:260) *AND* stage3(390:440) *AND* stage3(590:640) ➔ stage2(40:70) *AND* stage2(100:220) *AND* stage2(250:290) *AND* stage2(320:340) *AND* stage2(360:410) *AND* stage2(450:600) *AND* stage2(630:720) [Confidence: 100%; Support: 14.28%]

3. body-N (0:30) ➔ low-hrate (0:10) [Confidence: 100%; Support: 17.39%]

4. low-oxy (0:10) ➔ low-hrate (0:10) [Confidence: 100%; Support: 5.88%]

Analysis of Rules:

Rule 1 is interesting since it indicates that if there is a dip in the heart rate below normal then it can occur only during the initial 10 epochs (5 minutes). Rule 2 suggests that stages 2 and 3 of sleep dominate a large part of sleep by occurring regularly (almost every 5 minutes) starting from epoch 40 (20 minutes into sleep) to nearly the end of sleep. Rule 3 suggests that during the initial periods of sleep when it is difficult to suggest in most cases the body position of the person (since he/she is wriggling in the bed before settling into a relaxed position) he/she experiences reduced levels of heartbeats. Rule 4 suggests that low oxygen potential in the initial 10 epochs results in a corresponding drop in heartbeat levels during the same time period.

6.4.4 Analysis of micro-level dataset with window based association rule mining (fragmentation frequency count included)

The above set of rules can be improved upon to include additional information. The additional information would reveal the frequency with which the event of interest (could persist for several epochs together) repeats within the window-size under consideration. A new filter enables us to transform the dataset we used to obtain the results shown above, into one that now captures the frequency of fragmentation of the

events of interest. Experimental results obtained by running over the transformed dataset are provided below.

Experiment-8:

Objective:

Determine the variations in sleep stages, heart rates and oxygen potential in the blood.

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	7
Attribute names:	stageW (wake state of sleep), stage2 (stage-2 of sleep), stageR (REM stage of sleep), low-oxy (low oxygen potential), reduce-oxy (reduced oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate).
Minimum number of rules sought:	10
Minimum confidence:	70%
Upper bound for support:	60%
Lower bound for support:	10%
Required attributes in the antecedents:	none
Required attributes in the consequents:	none
Disallowed attributes in the antecedents:	none
Disallowed attributes in the consequents:	none
Epoch-window:	50

Sample rules obtaining from Association rule mining over the full training set are provided below:

1. stageW (0-50#1) ➔ low-hrate (0-50#1)
   [Confidence: 86.46%; Support: 68.6%; Lift: 0.9963; Chi-square: 0.0417; P>0.05]
2. low-oxy (0-50#1) ➔ low-hrate (0-50#1)
   [Confidence: 88.04%; Support: 66.94%; Lift: 1.0146; Chi-square: 0.5366; P>0.05]
3. stageW (0-50#1) && low-oxy (0-50#1) ➔ low-hrate (0-50#1)
   [Confidence: 87.84%; Support: 53.72%; Lift: 1.0122; Chi-square: 0.1862; P>0.05]

Analyses of Rules:

From the sample list of rules mentioned above, we see that there is association between wake state, lower heart rates and low oxygen potentials observed in the first 50 epochs. Though the confidence levels of the associations are fairly high, their statistical significance is weak and hence not much can be read from the rules generated.

Experiment-9:

Objective:

Determine variations in heart rates and oxygen potential with variations in stage-2 and REM stage of sleep.

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	7
Attribute names:	stageW (wake state of sleep), stage2 (stage-2 of sleep), stageR (REM stage of sleep), low-oxy (low oxygen potential), reduce-oxy (reduced oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate).
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedents:	stage2, stageR
Required attributes in the consequents:	none
Disallowed attributes in the antecedents:	none
Disallowed attributes in the consequents:	stage2, stageR
Epoch-window:	50

Sample rules obtaining from Association rule mining over the full training set are provided below:

1. stage-2 (750-800#1) && stage-R (750-800#1) ➔ stage-W (300-350#1) && low-oxy (0-50#1)
   [Confidence = 72.22%; Support = 10.74%; Lift = 2.0807; Chi-square = 13.1239; P<0.001]

2. stage-2 (750-800#1) && stage-R (750-800#1) ➔ stage-W (300-350#1) && low-hrate (0-50#1)

[Confidence = 72.22%; Support = 10.74%; Lift = 2.0323; Chi-square = 12.4171; P<0.001]

3. stage-2 (750-800#1) && stage-R (750-800#1) ➔ stage-W (300-350#1) && low-hrate (0-50#1) && low-oxy (0-50#1)

[Confidence = 72.22%; Support = 10.74%; Lift = 2.2997; Chi-square = 16.3467; P<0.001]

4. stage-2 (750-800#1) && stage-R (750-800#1) && low-oxy (0-50#1) ➔ stage-W (300-350#1) && low-hrate (0-50#1)

[Confidence = 76.47%; Support = 10.74%; Lift = 2.1518; Chi-square = 14.4598; P<0.001]

5. stage-2 (750-800#1) && stage-R (750-800#1) && low-hrate (0-50#1) ➔ stage-W (300-350#1) && low-oxy (0-50#1)

[Confidence = 76.47%; Support = 10.74%; Lift = 2.2031; Chi-square = 15.2139; P<0.001]

Analyses of Rules:

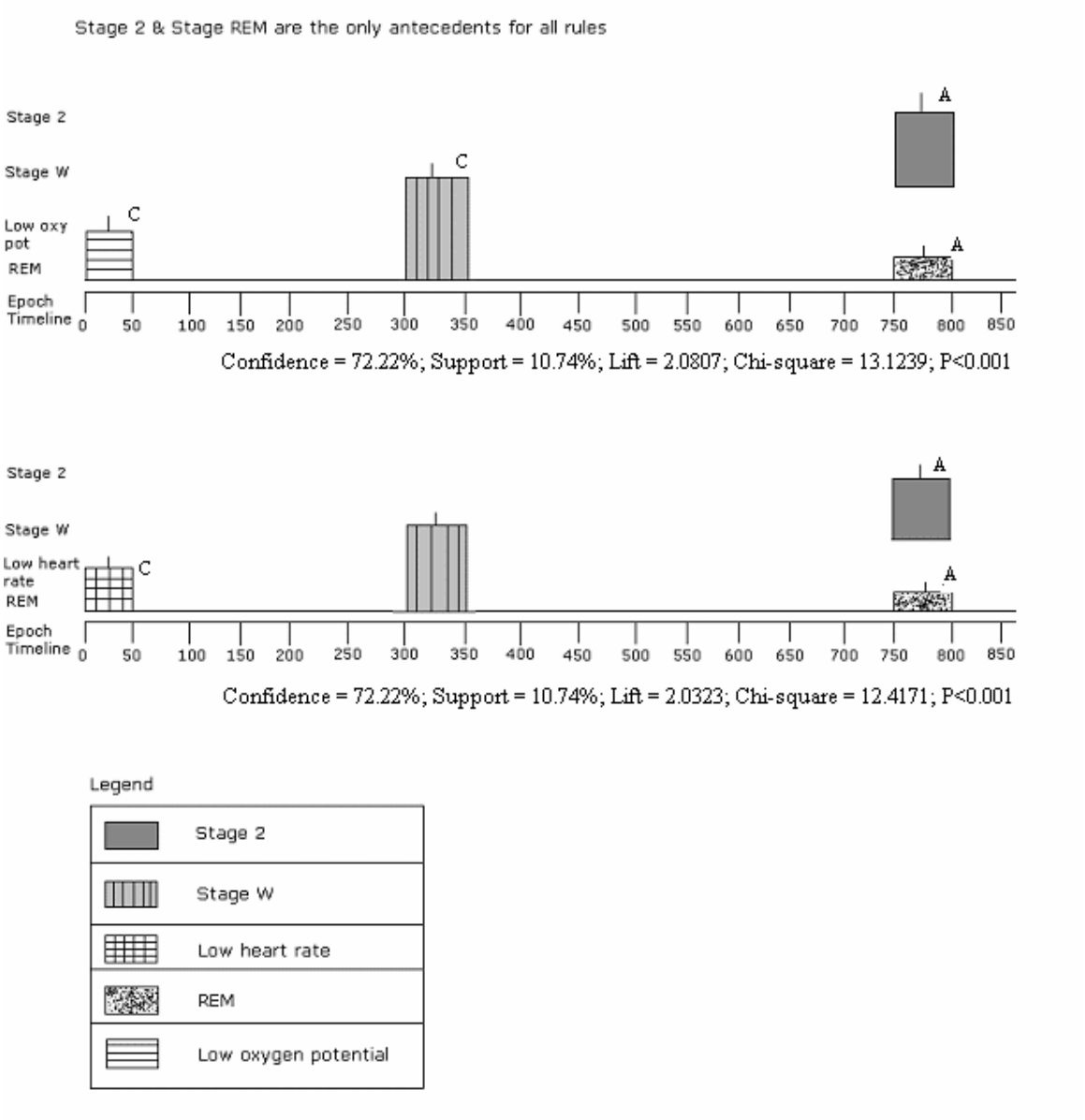


Fig. 6.2. Variation in stage-2 and REM correlating with low heart rates and low oxygen potential

The resulting rules and the above figure illustrate that low oxygen potentials and low heart rates are witnessed during the first 50 epochs of sleep. They show high correlation with stage-2 and REM occurring frequently between the 750th and 800th epochs. It is interesting that wake state between 300th and 350th epochs correlates highly with low oxygen and heart rates too.

Experiment-10:

Objective:

Determine variations in heart rates and oxygen potential with variations in stage 1, 2 and REM stage of sleep.

Input Parameters:

Instances:	121 instances (Every instance represents a single patient in the dataset)
Total number of attributes:	7
Attribute names:	stage1 (stage-1 of sleep), stage2 (stage-2 of sleep), stageR (REM stage of sleep), low-oxy (low oxygen potential), reduce-oxy (reduced oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate).
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedents:	stage1, stage2, stageR
Required attributes in the consequents:	none
Disallowed attributes in the antecedents:	none
Disallowed attributes in the consequents:	stage1, stage2, stageR
Epoch-window:	50

Sample rules obtaining from Association rule mining over the full training set are provided below:

1. stage1 (300-350#1) && stage2 (800-850#1) && stageR (650-700#1) ➔ low-hrate (0-50#1)
   [Confidence: 100%; Support: 5.79%; Lift: 1.1524; Chi-square: 1.1333; P>0.05]
2. stage1 (500-550#1) && stage2 (600-650#1) && stageR (650-700#1) ➔ low-hrate (0-50#1)
   [Confidence: 100%; Support: 7.44%; Lift: 1.1524; Chi-square: 1.4822; P>0.05]
3. stage1 (300-350#1) && stage2 (750-800#1) && stageR (750-800#1) ➔ low-oxy (0-50#1)
   [Confidence: 100%; Support: 6.61%; Lift: 1.1352; Chi-square: 2.6951; P>0.05]
4. stage1 (550-600#1) && stage2 (50-100#1) && stageR (650-700#1) ➔ low-oxy (0-50#1)

[Confidence: 100%; Support: 5.79%; Lift: 1.3152; Chi-square: 2.3483; P>0.05]

5. stage1 (550-600#1) && stage2 (50-100#1) && stageR (650-700#1) ➔ low-oxy (0-50#1) && low-hrate (0-50#1)

[Confidence: 100%; Support: 5.79%; Lift: 1.4938; Chi-square: 2.3483; P>0.05]

Analyses of Rules:

A sample set of all rules generated during the mining process is provided above. All the generated rules have either low heart rate or low oxygen potential detected during the first 50 epochs of sleep as the consequent. Though we obtain rules with high confidence, their lift values and consequently chi-square values are very low resulting in them having very little statistical significance.

6.5 *Summary of results*

With the effective use of filters, we prove on the basis of the results obtained, how window-based association rule mining technique which accounts for the fragmentation of events occurring within the window makes the rules more insightful and easy to analyze.

Initial experiments were performed over the dataset organized on the basis of epochs of sleep. Some of the rules obtained from these experiments were medically known. For instance, the rule, which suggests that patients lying on their back have a normal heart rate, is an expected behavior. We also obtain interesting rules like the one indicating that variations in oxygen potential do not affect normal functioning of the heart.

Further experiments were conducted with data organized differently. This time, every instance represented all records belonging to a single patient. We perform the analysis initially with the mining technique that generates association rules with relative time information embedded in them. Some of the rules generated validate known facts of sleep medicine. For instance, decrease in oxygen potential while the person is asleep disturbs the patient, causing them to first move towards the left before resting on their backs. This is an expected phenomenon since patient suffers from resistance to oxygen

flow more acutely while lying on the back especially when suffering from a condition called sleep apnea. There are many other rules produced which explain the timing interactions of the different events of interest.

More intuitive rules are obtained by using the window-based association rule mining technique. This technique generates rules with real-time information in them. For instance, it is interesting to note, that a drop in oxygen potential accompanies any drop in heart rate. However, this drop is seen only in the first ten epochs of sleep. Thus, we get a real-time idea of when such interesting events occur facilitating easy analysis. Likewise, inclusion of fragmentation frequency of events within the window also leads to significant amount of information being available to make a good analysis.

Chapter 7

Mixed-Level: Dataset Description and Analysis

7.1 Dataset and pre-processing details

Some of the data collected from patient records has both macro and micro-level data information. By going over individual patient records of the macro-level and micro-level dataset, we find 81 patients whose macro as well as micro-level information is available. These 81 patient records are compiled together to constitute the mixed-level dataset. This dataset includes complex data since the attributes hold numeric, nominal, set as well as time sequence values. Analysis of the mixed-level data holds a lot of promise, since it can potentially reveal several interesting associations between macro and micro-level attributes. All the pre-processing mentioned for the macro and micro-level dataset is performed for the mixed-level dataset.

7.2 Analysis of mixed-level data

We begin by describing the set of experiments performed with the aim of identifying physiological parameter variation patterns with baseline oxygen potential, a threshold parameter used widely in sleep-disorder diagnosis. Baseline oxygen potential can be defined as the oxygen potential measured during the epoch when the patient enters stage-1 of sleep for the first time without experiencing any sleep fragmentation events, for instance, apneaic episodes or arousals. With this potential as the new reference or threshold, the oxygen potential values measured during every epoch is discretized into two classes. Epochs during which oxygen potential falls below 3% baseline oxygen

potential are classified into the 'low-oxygen potential' category, while those above this level are classified into 'normal-oxygen potential' category.

Experiment-1a:

Objective:

Determine associations among normal and low baseline oxygen potential, low or reduced heart rate, and body position during the night under observation.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	8
Attribute Names:	low-baseline (low baseline oxygen potential), normal-oxy (normal oxygen potential), low-hrate (low heart rate), reduce-hrate (reduced heart rate), body-B (body position – Back), body-P (body position – Prone), body-L (body position – Left), body-R (body positon – Right)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	10%
Required attributes in the antecedent:	low-baseline, normal-oxy
Required attributes in the consequent:	none
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	low-baseline, normal-oxy
Epoch-window:	100

Sample rules obtained from association rule mining over the full training set are provided below:

1. low-baseline (0-100#1) && normal-oxy (400-500#1) ➔ body-B (700-800#1)
   [Confidence = 82.35%; Support = 17.28%; Lift = 1.2586; Chi-square = 7.226; 0.001<P<0.01]
2. low-baseline (0-100#1) && normal-oxy (0-100#1) && body-R (100-200#1) ➔ low-hrate (0-100#1)
   [Confidence = 100%; Support = 16.05%; Lift = 1.1739; Chi-square = 2.693; P>0.05]

Analyses of Rules:

Above is a sample set of rules generated during the experiment. The integer following the '#' sign in any rule indicates the number of times the spell comprising occurrence of the specific event in contiguous epochs is broken within the window size considered. For this experiment it must however be said that 100 epochs is a fairly large window-size (100 epochs = 50 minutes) since it can help in observing and analyzing only the coarse structure of sleep. Though several rules were obtained from this experiment, only rule-1 was found to be statistically significant. It is interesting to note that the oxygen potential is found to be below the baseline level during the first hour of sleep. Also interesting is the finding of a reasonably strong correlation between patients who experience below baseline oxygen potential during the first hour of sleep and waking up after spending a major portion of the last hour sleeping on his/her back.

Experiment-1b:

Objective:

This experiment is similar to Experiment-1a, with the exception that normal oxygen potential is not considered as an attribute in the mining process.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	7
Attribute Names:	low-baseline (low baseline oxygen potential),
	low-hrate (low heart rate),
	reduce-hrate (reduced heart rate),
	body-B (body position – Back),
	body-P (body position – Prone),
	body-L (body position – Left),
	body-R (body positon – Right)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	10%

Required attributes in the antecedent: low-baseline
Required attributes in the consequent: none
Disallowed attributes in the antecedent: none
Disallowed attributes in the consequent: low-baseline
Epoch-window: 100

Sample rules obtained from association rule mining over the full training set are provided below:

1. low-baseline (200-300#1) ➔ low-hrate (0-100#1)
   [Confidence = 85%; Support = 20.99%; Lift = 0.9978; Chi-square = 0.0007; P>0.05]
2. low-baseline (200-300#1) ➔ body-B(700-800#1)
   [Confidence = 85%; Support = 20.99%; Lift = 1.2991; Chi-square = 4.4565; P>0.05]
3. low-baseline (800-900#1) ➔ low-hrate (0-100#1)
   [Confidence = 84.62%; Support = 27.16%; Lift = 0.9933; Chi-square = 0.0098; P>0.05]
4. low-baseline (0-100#1) ➔ low-hrate (0-100#1)
   [Confidence = 84.62%; Support = 27.16%; Lift = 0.9933; Chi-square = 0.0098; P>0.05]
5. low-baseline (0-100#1) && body-B (0-100#1) ➔ low-hrate (0-100#1)
   [Confidence = 85%; Support = 20.99%; Lift = 0.9978; Chi-square = 0.0007; P>0.05]

Analyses of Rules:

From the above set of rules, it becomes clear that lower heart rates are evident in the initial 100 epochs of patient sleep. However, oxygen potential lower than baseline may be experienced even later (during the 200 and 300th epochs). Also it is interesting to note, that during the window when the heart rate or oxygen potential falls below normal, the characteristic of sleeping on the back is also noticed. Patients suffering from severe sleep apnea show a tendency to sleep on their backsides. In this position, there is a good chance for the oxygen potential to dip below the baseline threshold value. Thus, we can see a link from the results between a dip in oxygen potential and patients sleeping on their backs. In this experiment although the rules can be medically proved to represent a normally observed behavior, their statistical significance is above the acceptable threshold limit of 0.05 thus making them weak.

Experiment –2a:

Objective:

Determine associations between Periodic leg movement syndrome and lowering of heart rate and oxygen potential.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	3
Attribute Names:	low-baseline (low baseline oxygen potential), low-hrate (low heart rate), PLMS (Periodic Leg Movement Syndrome)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	95%
Lower bound for support:	10%
Required attributes in the antecedent:	none
Required attributes in the consequent:	PLMS
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	low-baseline, low-hrate
Epoch-window:	50

Sample rules obtained from association rule mining over the full training set are provided below:

1. low-baseline (850-900#1) ➔ PLMS (infrequent)
   [Confidence = 72.22%; Support = 16.05%; Lift = 1.2188; Chi-square = 1.6114; P>0.05]
2. low-baseline (750-800#1) ➔ PLMS (infrequent)
   [Confidence = 72.22%; Support = 16.05%; Lift = 1.2188; Chi-square = 1.6114; P>0.05]
3. low-baseline (800-850#1) ➔ PLMS (infrequent)
   [Confidence = 68.18%; Support = 18.52%; Lift = 1.1506; Chi-square = 0.9964; P>0.05]
4. low-baseline (500-550#1) ➔ PLMS (infrequent)
   [Confidence = 65.22%; Support = 18.52%; Lift = 1.1005; Chi-square = 0.4762; P>0.05]
5. low-baseline (800-850#1) && low-hrate (0-50#1) ➔ PLMS (infrequent)
   [Confidence = 65%; Support = 16.05%; Lift = 1.0969; Chi-square = 0.3627; P>0.05]

Analyses of Rules:

From the sample set of rules seen above it is clear that below baseline oxygen potential is not a cause for increased severity of Periodic Leg Movement Syndrome (PLMS). This experiment was performed on a 25-minute window (i.e., with window size of 50 epochs). However, the rules resulting from the experiment had very little statistical significance owing to very low chi-square values.

Experiment-2b:

Objective:

The objective of this experiment is identical to that of Experiment-2a. However the upper bound and lower-bound support is reduced to check if more rules are possible.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	3
Attribute Names:	low-baseline (low baseline oxygen potential), low-hrate (low heart rate), PLMS (Periodic Leg Movement Syndrome)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	20%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	PLMS
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	low-baseline, low-hrate
Epoch-window:	50

Sample rules obtained from association rule mining over the full training set are provided below:

Same as in Experiment-2a.

Analyses of Rules:

The rules that result from the mining process are identical to the ones in Experiment-2a.

Experiment-2c:

Objective:

The objective is the same as that of Experiments-2a and 2b. However, in this case, we reduce the size of the windows further to analyze events of interest more closely.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	3
Attribute Names:	low-baseline (low baseline oxygen potential),
	low-hrate (low heart rate),
	PLMS (Periodic Leg Movement Syndrome)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	95%
Lower bound for support:	10%
Required attributes in the antecedent:	none
Required attributes in the consequent:	PLMS
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	low-baseline, low-hrate
Epoch-window:	10

Sample rules obtained from association rule mining over the full training set are provided below:

1. low-baseline (0-10#1) && low-baseline (540-550#1) ➔ PLMS (infrequent)
   [Confidence = 88.24%; Support = 18.52%; Lift = 1.489; Chi-square = 7.4833; 0.001<P<0.01]
2. low-baseline (0-10#1) && low-baseline (540-550#1) && low-hrate (0-10#1) ➔ PLMS (infrequent)
   [Confidence = 92.86%; Support = 16.05%; Lift = 1.567; Chi-square = 7.9149; 0.001<P<0.01]
3. low-baseline (0-10#1) && low-baseline (580-590#1) && low-hrate (0-10#1) ➔ PLMS (infrequent)
   [Confidence = 86.67%; Support = 16.05%; Lift = 1.4625; Chi-square = 5.7283; 0.01<P<0.05]
4. low-baseline (0-10#1) && low-baseline (130-140#1) ➔ PLMS (infrequent)
   [Confidence = 78.95%; Support = 18.52%; Lift = 1.3322; Chi-square = 3.9853; 0.01<P<0.05]
5. low-baseline (0-10#1) && low-baseline (560-570#1) ➔ PLMS (infrequent)
   [Confidence = 87.5%; Support = 17.28%; Lift = 1.4766; Chi-square = 6.5853; 0.01<P<0.05]
6. low-baseline (560-570#1) && low-hrate (0-10#1) ➔ PLMS (infrequent)

[Confidence = 78.95%; Support = 18.52%; Lift = 1.3322; Chi-square = 3.9853; 0.01<P<0.05]

7. low-baseline (0-10#1) && low-baseline (560-570#1) && low-hrate (0-10#1) ➔ PLMS (infrequent)

[Confidence = 87.5%; Support = 17.28%; Lift = 1.4766; Chi-square = 6.5853; 0.01<P<0.05]

8. low-baseline (540-550#1) && low-hrate (0-10#1) ➔ PLMS (infrequent)

[Confidence = 80%; Support = 19.75%; Lift = 1.35; Chi-square = 4.7311; 0.01<P<0.05]

9. low-baseline (0-10#1) && low-baseline (380-390#1) ➔ PLMS (infrequent)

[Confidence = 81.25%; Support = 16.05%; Lift = 1.3711; Chi-square = 3.994; 0.01<P<0.05]

Analyses of Rules:

The sample set of rules, which meet the statistical significance standards following the mining process, are shown above. The rules reinforce the point we made in the analysis to the earlier experiments stating that lower heart rates are experienced during the initial stages of sleep. From the experiments, we can also observe that the dips in the heart rate and oxygen potential below baseline, which in experiments-1a and 1b were in the first 100 epochs, actually occur in the initial 10 epochs itself (i.e., in the first 5 minutes of sleep). Unlike the results in experiments- 1a and 1b, the rules capturing the information in this experiment are found to be statistically significant. It is interesting to observe, from the sample set of rules shown above, that oxygen potential falls below baseline during the initial 10 epochs as well as during the middle stage of sleep (4.5 to 4.9 hours since beginning of sleep). However, these dips in oxygen potential or heart rate do not seem to induce significant Periodic leg movements (PLMS). The results in this experiment probe the variations between the participating parameters in a microscopic region (epoch window =10) and conform the results obtained from the rules when variations were probed within a macroscopic context (epoch window = 100).

Experiment-3:

Objective:

Establish associations linking PLMS with sleep fragmentation witnessed in stage2, REM, low heart rate and oxygen potential below baseline threshold.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	5
Attribute Names:	stage2
	stageR
	low-baseline (low baseline oxygen potential),
	low-hrate (low heart rate),
	PLMS (Periodic Leg Movement Syndrome)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	20%
Lower bound for support:	1%
Required attributes in the antecedent:	low-baseline
Required attributes in the consequent:	PLMS
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage2, stageR, low-baseline
Epoch-window:	50

Sample rules obtained from association rule mining over the full training set are provided below:

1. low-baseline(850-900#1) ➔ PLMS (infrequent)
   [Confidence = 72.22%; Support = 16.05%; Lift = 1.2186; Chi-square = 1.6091; P>0.05]
2. low-baseline (750-800#1) ➔ PLMS (infrequent)
   [Confidence = 72.22%; Support = 16.05%; Lift = 1.2186; Chi-square = 1.6091; P>0.05]
3. low-baseline (0-50#1) && stage2 (600-650#1) ➔ PLMS (infrequent)
   [Confidence = 68.42%; Support = 16.05%; Lift = 1.1546; Chi-square = 0.863; P>0.05]
4. low-baseline (0-50#1) && stage2 (600-650#1) && low-hrate (0-50#1) ➔ PLMS (infrequent)
   [Confidence = 68.42%; Support = 16.05%; Lift = 1.1546; Chi-square = 0.863; P>0.05]
5. low-baseline (0-50#1) && stage2 (400-450#1) ➔ PLMS (infrequent)
   [Confidence = 68.42%; Support = 16.05%; Lift = 1.1546; Chi-square = 0.863; P>0.05]

Analyses of Rules:

From the rules shown above it is apparent that stage 2 is more prominent during the middle stages of sleep. Also, it is clear that lower heart rates or oxygen potentials below baseline threshold are not experienced around the time, the subject is in stage 2 of sleep. However, not much can be read from these rules since they do not reach the level of statistical significance necessary for them to be considered robust.

Experiment-4a:

Determine correlation between susceptibility to Obstructive Sleep Apnea (OSA) and below baseline oxygen potential and below normal heart rates.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	3
Attribute Names:	low-baseline (low baseline oxygen potential),
	low-hrate (low heart rate),
	OSA (Obstructive Sleep Apnea)
Number of rules sought:	10
Minimum confidence:	20%
Upper bound for support:	25%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	OSA
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	low-baseline, low-hrate
Epoch-window:	50

Sample of statistically significant rules obtained from association rule mining over the full dataset:

1. low-baseline (50-100#1) ➔ OSA (mild to moderate)
   [Confidence = 27.78%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3267; 0.001<P<0.01]
2. low-baseline (250-300#1) ➔ OSA (mild to moderate)
   [Confidence = 27.78%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3267; 0.001<P<0.01]
3. low-baseline (300-350#1) ➔ OSA (mild to moderate)
   [Confidence = 27.78%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3267; 0.001<P<0.01]
4. low-baseline (750-800#1) ➔ OSA (mild to moderate)
   [Confidence = 27.78%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3267; 0.001<P<0.01]

Analysis of Results:

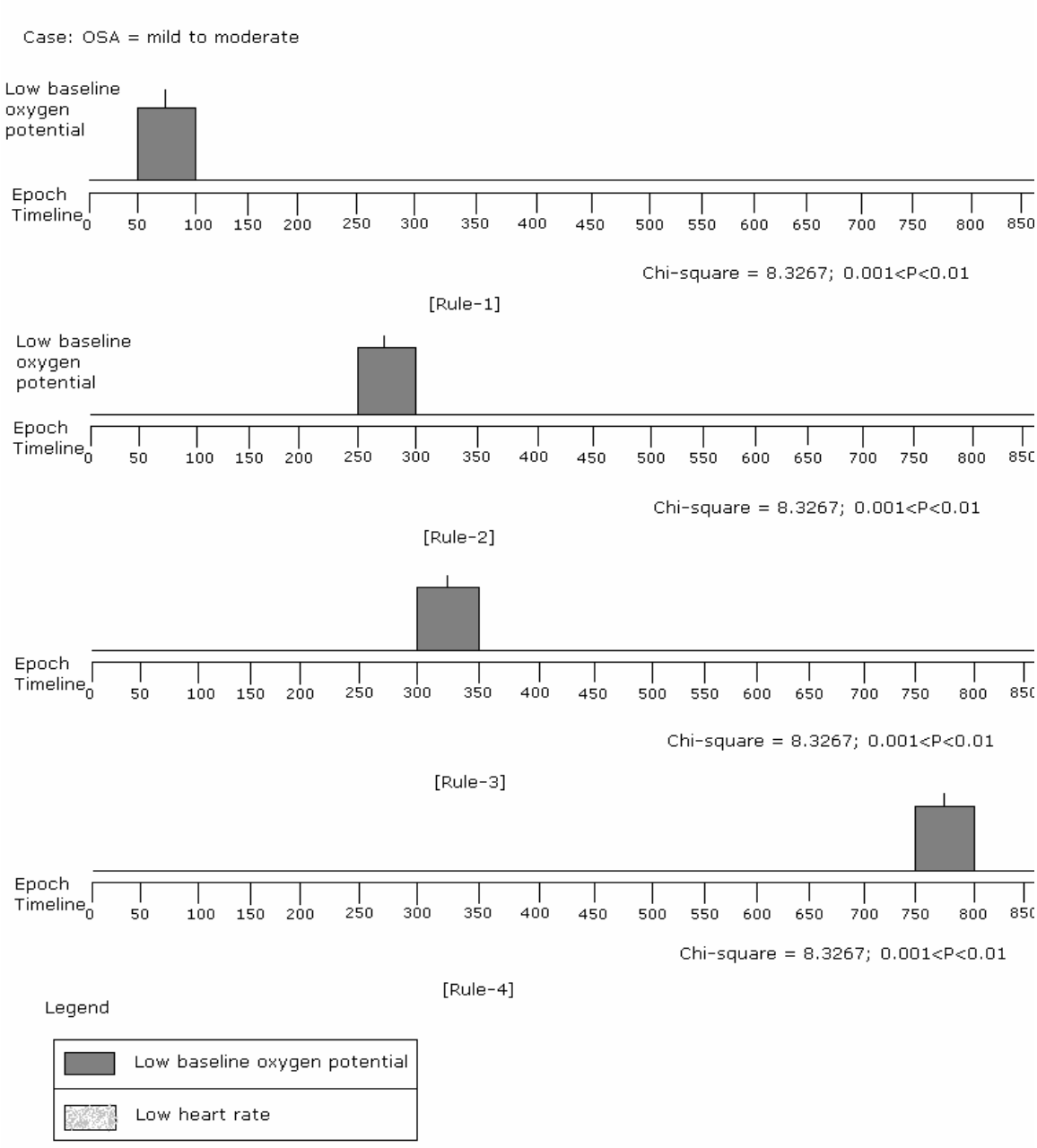


Fig. 7.1: Frequent windows of below baseline oxygen potential that correlates with mild to moderate OSA (Experiment – 4a).

The above figure is another way of looking at the rules generated from the experiment. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. The rules obtained in this window do not identify patterns for patients suffering from acute OSA. The rules validate the medically

well-known fact that patients with mild or moderate OSA do not suffer from recurring or prolonged spells of low oxygen potential.

Experiment-4b:

Determine whether baseline oxygen potential associates with Obstructive Sleep Apnea (OSA) with a window size of 20 epochs.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	3
Attribute Names:	low-baseline (low baseline oxygen potential),
	low-hrate (low heart rate),
	OSA (Obstructive Sleep Apnea)
Number of rules sought:	50
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	OSA
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	low-baseline, low-hrate
Epoch-window:	20

Sample of statistically significant rules obtained from association rule mining over the full dataset:

1. low-baseline (300-320#1) && low-baseline (400-420#1) && low-baseline (440-460#1) ➔ OSA (moderately-severe)
   [Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]
2. low-baseline (300-320#1) && low-baseline (440-460#1) && low-baseline (680-700#1) ➔ OSA (moderately-severe)
   [Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]
3. low-baseline (300-320#1) && low-baseline (400-420#1) && low-baseline (440-460#1) && low-baseline (660-680#1) ➔ OSA (moderately-severe)
   [Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]
4. low-baseline (300-320#1) && low-baseline (440-460#1) && low-baseline (660-680#1) && low-

baseline (680-700#1) ➔ OSA (moderately-severe)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]

5. low-baseline (300-320#1) && low-baseline (400-420#1) && low-baseline (440-460#1) && low-baseline (680-700#1) ➔ OSA (moderately-severe)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]

6. low-baseline (300-320#1) && low- baseline (400-420#1) && low-baseline (440-460#1) && low-hrate (0-20#1) ➔ OSA (moderately-severe)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.81825; Chi-square = 8.3276; 0.001<P<0.01]

7. low-baseline (300-320#1) && low- baseline (440-460#1) && low-baseline (680-700#1) && low-hrate (0-20#1) ➔ OSA (moderately-severe)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.81825; Chi-square = 8.3276; 0.001<P<0.01]

8. low-baseline (80-100#2) ➔ OSA (mild)

[Confidence = 66.67%; Support = 7.41%; Lift = 2.8421; Chi-square = 10.5284; 0.001<P<0.01]

9. low-baseline (80-100#2) && low-hrate (0-20#1) ➔ OSA (mild)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.6645; Chi-square = 7.5328; 0.001<P<0.01]

10. low-baseline (0-20#2) ➔ OSA (mild)

[Confidence = 66.67%; Support = 7.41%; Lift = 2.8421; Chi-square = 10.5284; 0.001<P<0.01]

11. low-baseline (0-20#2) && low-hrate (0-20#1) ➔ OSA (mild)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.6645; Chi-square = 7.5328; 0.001<P<0.01]

12. low-baseline (240-260#1) && low-baseline (320-340#1) ➔ OSA (moderately-severe)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]

13. low-baseline (240-260#1) && low-baseline (320-340#1) && low-hrate (0-20#1) ➔ OSA (moderately-severe)

[Confidence = 62.5%; Support = 6.17%; Lift = 2.8125; Chi-square = 8.3276; 0.001<P<0.01]

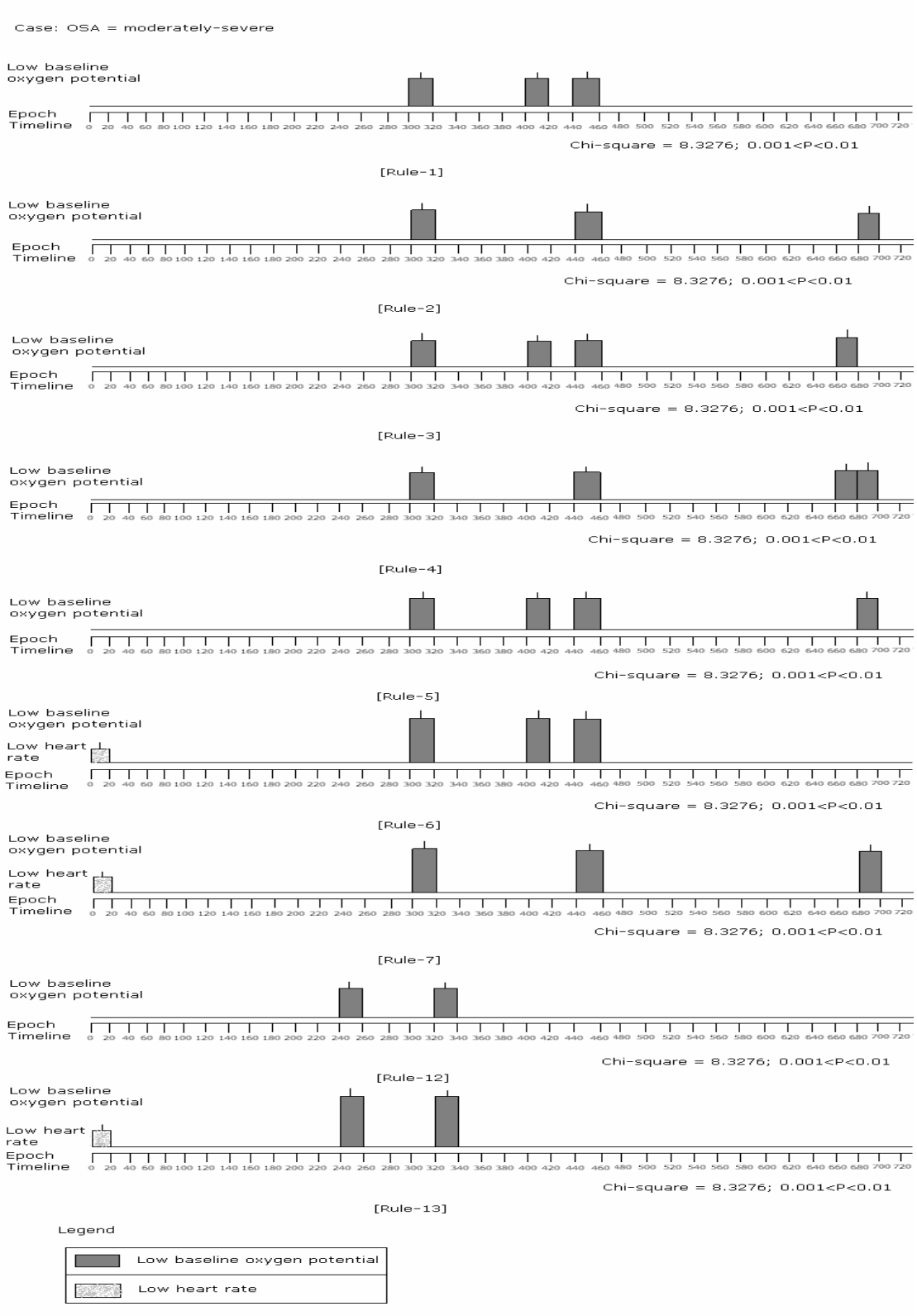


Fig. 7.2: Epoch windows of 20 with low baseline oxygen potential and low heart-rate that correlate with moderately severe OSA (Experiment-4b).

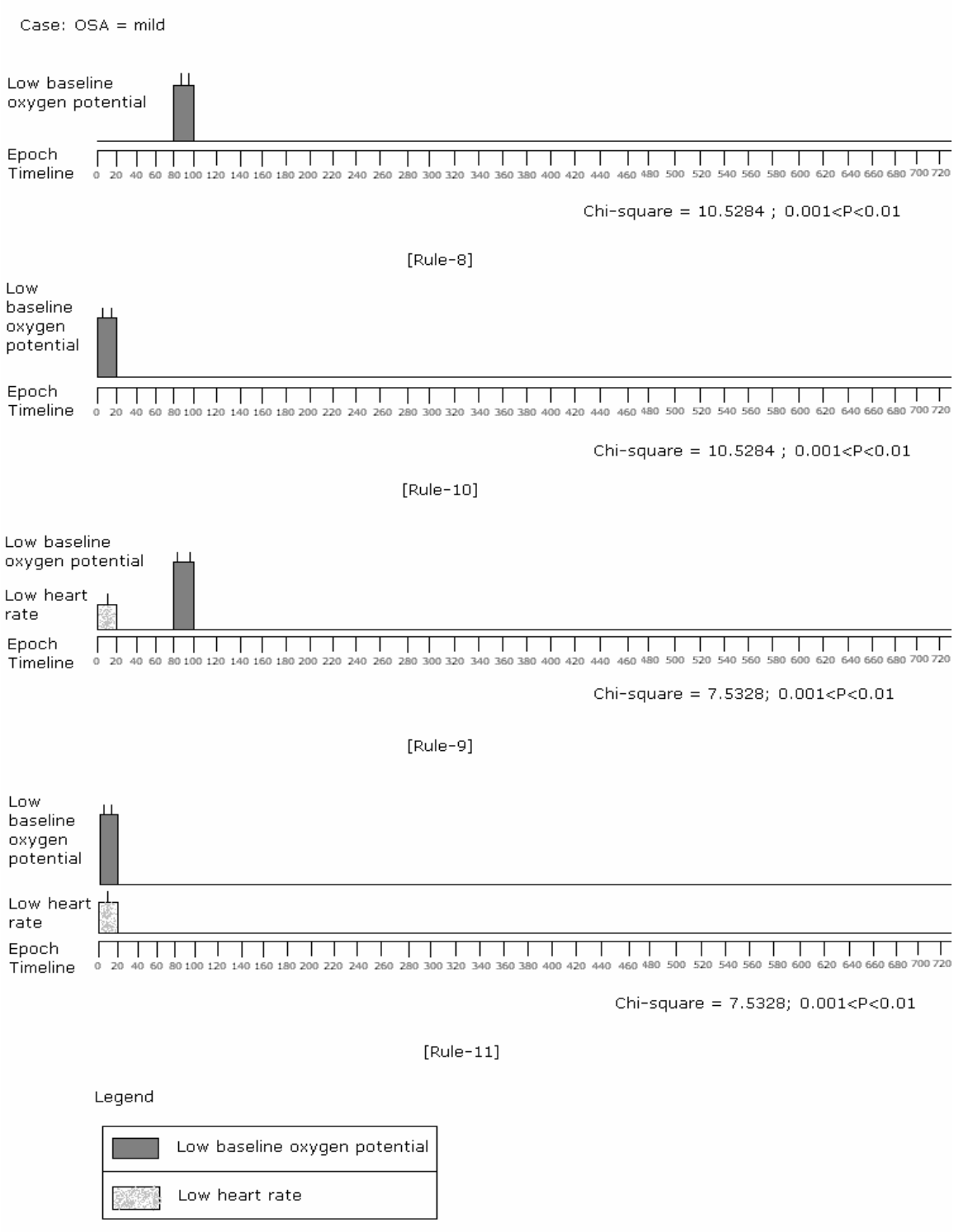


Fig. 7.3: The figure represents epoch windows with low baseline oxygen potential and low heart rate that correlate with mild OSA (Experiment-4b).

Analysis of Results:

The diagrammatic representation of the rules above helps in distinguishing how below baseline oxygen potential and low heart rate influences the degree of Obstructive Sleep Apnea (OSA) that a patient suffers from. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. The chi-square values suggest that the rules generated from the experiment are fairly robust. Figure 7.3 illustrates that patients who do not suffer from severe cases of OSA tend to experience low oxygen potentials only during the first hour of sleep, sometimes as soon as 10 minutes into sleep. However, their sleep is not disturbed after the first hour since oxygen potentials recover and do not fall below the normal threshold value. Consequently, they do not suffer from serious OSA-related problems. On the contrary, patients who have moderately severe to very severe cases of OSA seem to suffer from intermittent periods of low oxygen potential as illustrated in figure 7.2. This seems to suggest that intermittent low oxygen potentials have a high correlation with inducing increasingly severe levels of OSA. The intermittent low oxygen potentials are frequently observed between 2 ½ to 4 hours after the patient goes to sleep.

In experiment 4a, we obtained rules that correlated mild to moderate degrees of OSA with below baseline oxygen potential. However, it was difficult to identify which of the rules were more likely to correlate with mild OSA and which with moderate OSA. With the help of experiment 4b, we can clearly identify that below baseline oxygen potentials witnessed only during the first 100 epochs correlates highly with mild OSA while presence of low oxygen potentials at times beyond the first 100 epochs correlate with higher degrees of OSA. Thus, we see that a smaller window of 20 epochs helps us to better analyze results accruing from a larger window of 50 epochs.

Experiment-5:

Objective:

Determine variations in stage 2 and REM of sleep that associate with PLMS and OSA.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	4
Attribute Names:	stage 2 (Stage 2 of sleep),
	stage R (REM stage of sleep),
	OSA (Obstructive Sleep Apnea),
	PLMS (Periodic Leg Movement Syndrome)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	stage2, stageR
Required attributes in the consequent:	none
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage2, stageR
Epoch-window:	50

Sample of statistically significant rules obtained from association rule mining over the full training set:

1. stage-2 (700-750#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9943; P<0.001]
2. stage-2 (650-700#2) && stage-R (750-800#1) ➔ PLMS (infrequent)
   [Confidence = 100%; Support = 7.41%; Lift = 1.6875; Chi-square = 4.4547; 0.01<P<0.05]
3. stage-2 (300-350#2) && stage-R (750-800#1) ➔ PLMS (infrequent)
   [Confidence = 100%; Support = 7.41%; Lift = 1.6875; Chi-square = 4.4547; 0.01<P<0.05]
4. stage-2 (150-200#1) && stage-R (800-850#1) ➔ OSA (moderately-severe)
   [Confidence = 55.56%; Support = 6.17%; Lift = 2.5; Chi-square = 6.5056; 0.01<P<0.05]
5. stage-2 (800-850#1) && stage-R (700-750#1) ➔ OSA (mild)
   [Confidence = 50%; Support = 6.17%; Lift = 2.1316; Chi-square = 4.4745; 0.01<P<0.05]

Analyses of Results:

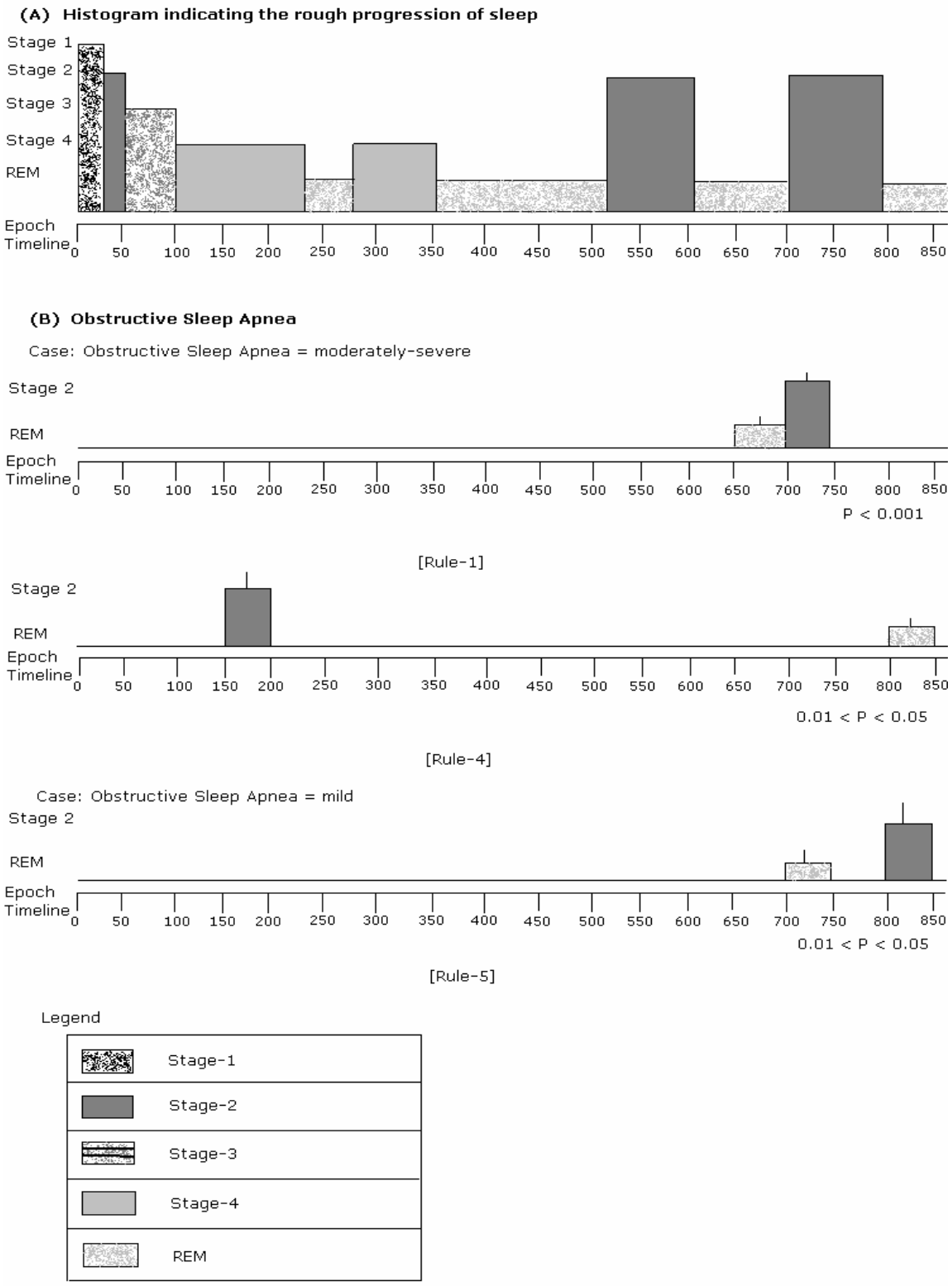


Fig. 7.4: Variation of stages 2 and REM of sleep in patients suffering from moderately severe and mild obstructive sleep apnea (OSA) (Experiment-5).

This experiment focuses on obtaining rules correlating OSA and PLMS with stage-2 and REM of sleep. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. Rule 1 is the most statistically significant rule obtained from this experiment. It indicates specific epoch windows of size 50 towards the end of sleep which show a high correlation with fairly severe OSA. Rule 2 is an interesting result indicating that within the 50-epoch window, patients experiencing two fragmentations in stage-2 of sleep between the 650 and 700th epochs and one during REM sleep between the 750th and 800th epochs do not suffer from significant PLMS. Though the confidence of this rule is 100%, the lift for the rule is very small thus making it statistically less significant. Rule 3 similarly indicates that two fragmentations in stage-2 of sleep between 300th and 350th epochs and one during REM stage of sleep between the 750th and 800th epochs do not result in severe or frequent PLMS. We also obtain rules correlating stage-2 and REM of sleep with mild OSA. This rule though has a relatively low confidence compared to other rules.

Experiment-6:

Objective:

Determine variations in stage-2 and REM stage of sleep that associate with gender, Body-Mass Index (BMI), epworth or depression score.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	6
Attribute Names:	stage-2, Stage-REM, gender, bmi-class, epworth, depression
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%

Lower bound for support: 1%

Required attributes in the antecedent: stage2, stage-REM

Required attributes in the consequent: none

Disallowed attributes in the antecedent: none

Disallowed attributes in the consequent: stage-2, stage-REM

Epoch-window: 50

Sample of statistically significant rules obtained from association rule mining over the full training set:

1. stage-2 (500-550#1) && stage-R (200-250#1) ➔ depression (moderate)
   [Confidence = 100%; Support = 7.41%; Lift = 2.7; Chi-square = 11.0153; P<0.001]
2. stage-2 (500-550#1) && stage-R (500-550#1) ➔ depression (moderate)
   [Confidence = 100%; Support = 9.88%; Lift = 2.7; Chi-square = 15.0912; P<0.001]
3. stage-2 (250-300#2) && stage-R (200-250#1) ➔ depression (moderate)
   [Confidence = 100%; Support = 7.41%; Lift = 2.7; Chi-square = 11.0153; P<0.001]
4. stage-2 (50-100#1) && stage-R (400-450#1) ➔ bmi-class (obese) && depression (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 2.8929; Chi-square = 11.1312; P<0.001]
5. stage-2 (50-100#1) && stage-R (350-400#1) ➔ bmi-class (obese) && depression (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 2.8929; Chi-square = 11.1312; P<0.001]
6. stage-2 (700-750#1) && stage-R (450-500#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3738; P<0.001]
7. stage-2 (500-550#1) && stage-R (500-550#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3799; P<0.001]
8. stage-2 (500-550#1) && stage-R (450-500#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 71.43%; Support = 6.17%; Lift = 3.8571; Chi-square = 14.2082; P<0.001]
9. stage-2 (600-650#1) && stage-R (500-550#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3738; P<0.001]
10. stage-2 (700-750#1) && stage-R (700-750#1) ➔ bmi-class (obese) && epworth (mild)
    [Confidence = 55.56%; Support = 6.17%; Lift = 3.4615; Chi-square = 11.7223; P<0.001]
11. stage-2 (600-650#1) && stage-R (200-250#1) ➔ bmi-class (obese) && epworth (moderate)
    [Confidence = 62.5%; Support = 6.17%; Lift = 3.6161; Chi-square = 12.6879; P<0.001]
12. stage-2 (300-350#1) && stage-R (350-400#1) ➔ gender (female) && bmi-class (overweight)
    [Confidence = 71.43%; Support = 6.17%; Lift = 6.4286; Chi-square = 28.2109; P<0.001]
13. stage-2 (150-200#1) && stage-R (500-550#1) ➔ gender (female) && bmi-class (obese)
    [Confidence = 85.71%; Support = 7.41%; Lift = 3.3061; Chi-square = 14.2609; P<0.001]
14. stage-2 (450-500#1) && stage-R (650-700#1) ➔ depression (moderate)
    [Confidence = 62.5%; Support = 6.17%; Lift = 3.8942; Chi-square = 14.2078; P<0.001]
15. stage-2 (150-200#1) && stage-R (400-450#1) ➔ epworth (moderate) && depression (mild)
    [Confidence = 83.33%; Support = 6.17%; Lift = 3.5526; Chi-square = 12.9316; P<0.001]

16. stage-2 (400-450#1) && stage-R (800-850#1) ➔ epworth (severe)
    [Confidence = 63.64%; Support = 8.64%; Lift = 2.7129; Chi-square = 11.445; P<0.001]
17. stage-2 (450-500#2) && stage-R (250-300#1) ➔ gender (male) && epworth (moderate)
    [Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9842; P<0.001]
18. stage-2 (550-600#2) && stage-R (250-300#1) ➔ gender (male) && epworth (moderate)
    [Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9779; P<0.001]

Analyses of Rules:

The sample sets of rules provided above are statistically robust with the probability of the correlation having occurred by chance being only 1 in 1000. What makes the rules even more robust is that they display high-levels of confidence. The rules focus on determining the stage 2 and REM epoch windows (window size = 50 epochs) that display a high correlation with depression, gender, Body-mass index, epworth index and the whole gamut of combinations that arise by combining these factors with each other.

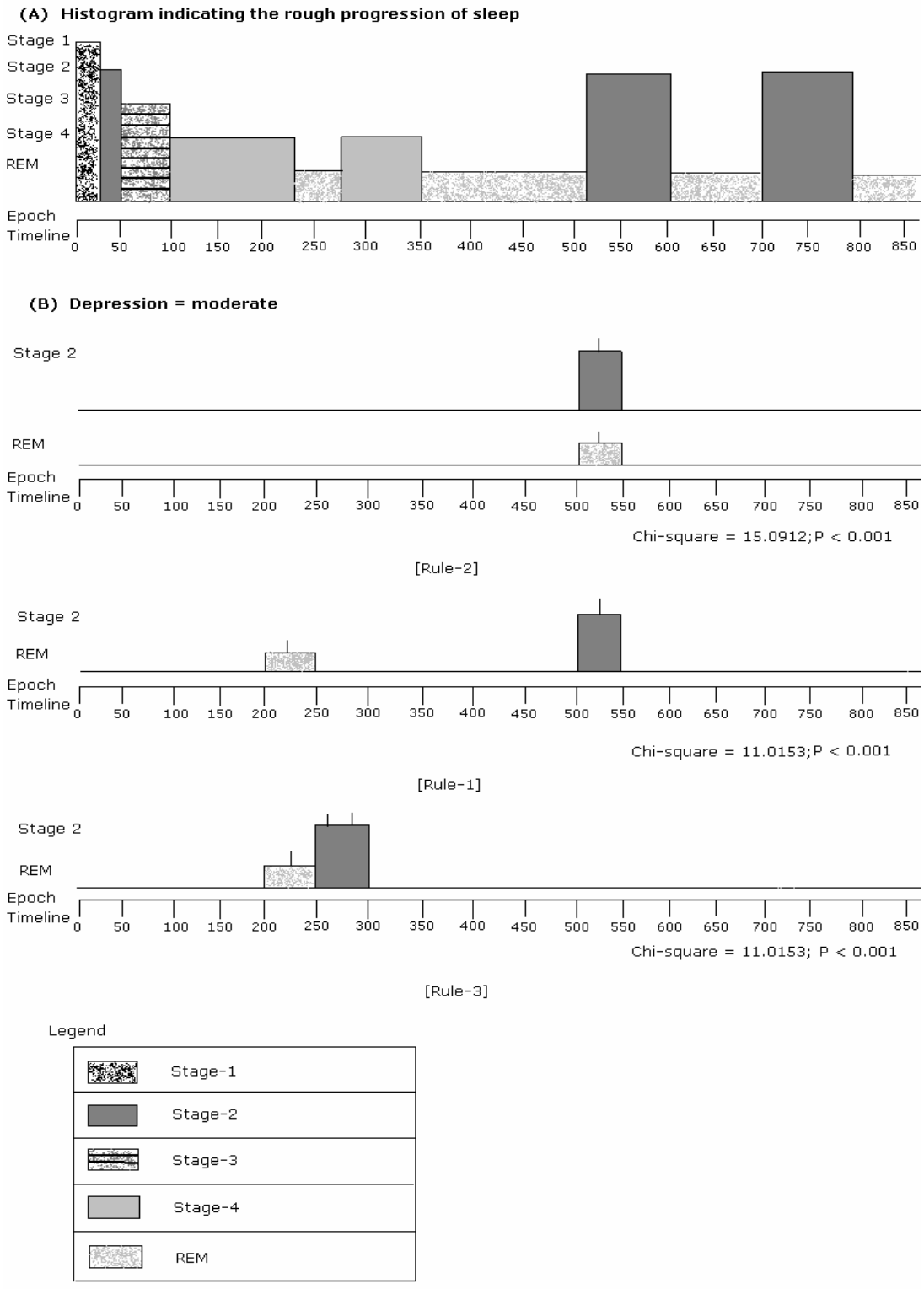


Fig. 7.5: Most frequent windows of stage 2 and REM sleep noticed in patients having moderate depression (Experiment-6).

The figure above depicts the frequently observed windows of stage-2 and REM of sleep that correlate with moderate degrees of depression in patients. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. Significant rules for mild or severe cases of depression in patients were not obtained from the experiment. We obtain robust rules with very high chi-square values. It is interesting that stage-2 and REM are both frequently observed between the 500^{th} and 550^{th} epochs. In addition, REM is also frequently observed between the 200^{th} and 250^{th} epochs. It is interesting that one of the rules identifies a correlation between moderate depression and presence of two fragments of stage-2 sleep between the 250^{th} and 300^{th} epochs.

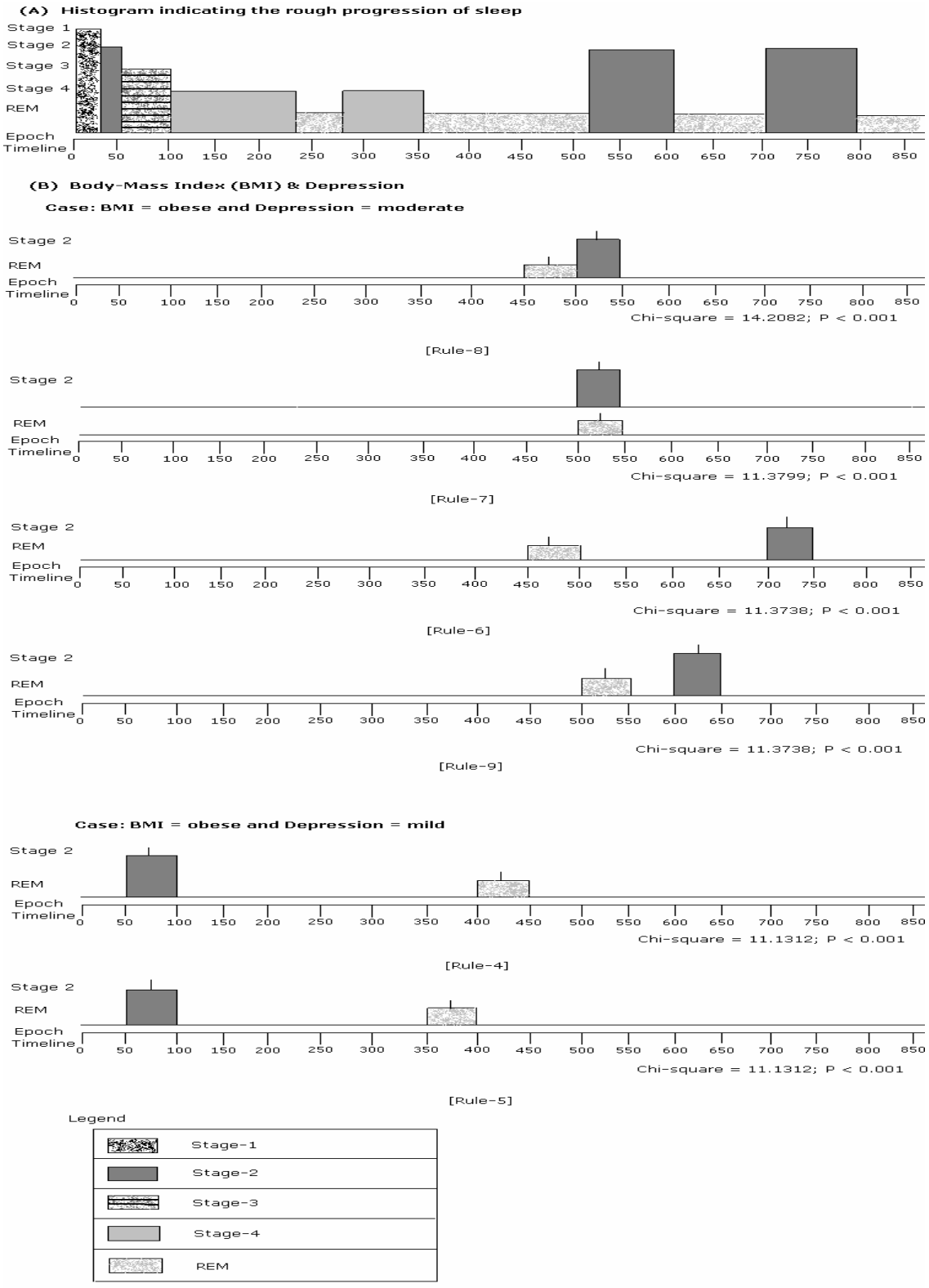


Fig. 7.6: Most frequent windows of stage 2 and REM sleep noticed in obese patients having mild to moderate depression (Experiment-6).

The diagrammatic representation of the rules above clearly indicates the manner in which frequently occurring windows of stage-2 and REM vary in obese patients who suffer from insignificant (mild) to moderate bouts of depression. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. Obese patients with moderate levels of depression seem to frequently experience epoch windows with stages-2 and REM towards the middle and terminal stages of sleep. In contrast, for obese patients with insignificant and mild depression complaints, stage-2 is witnessed within the first hour of sleep. REM in both cases occurs nearly within the same windows. Hence, stage-2 can be used as an identifier for segregating obese patients suffering from mild or moderate levels of depression.

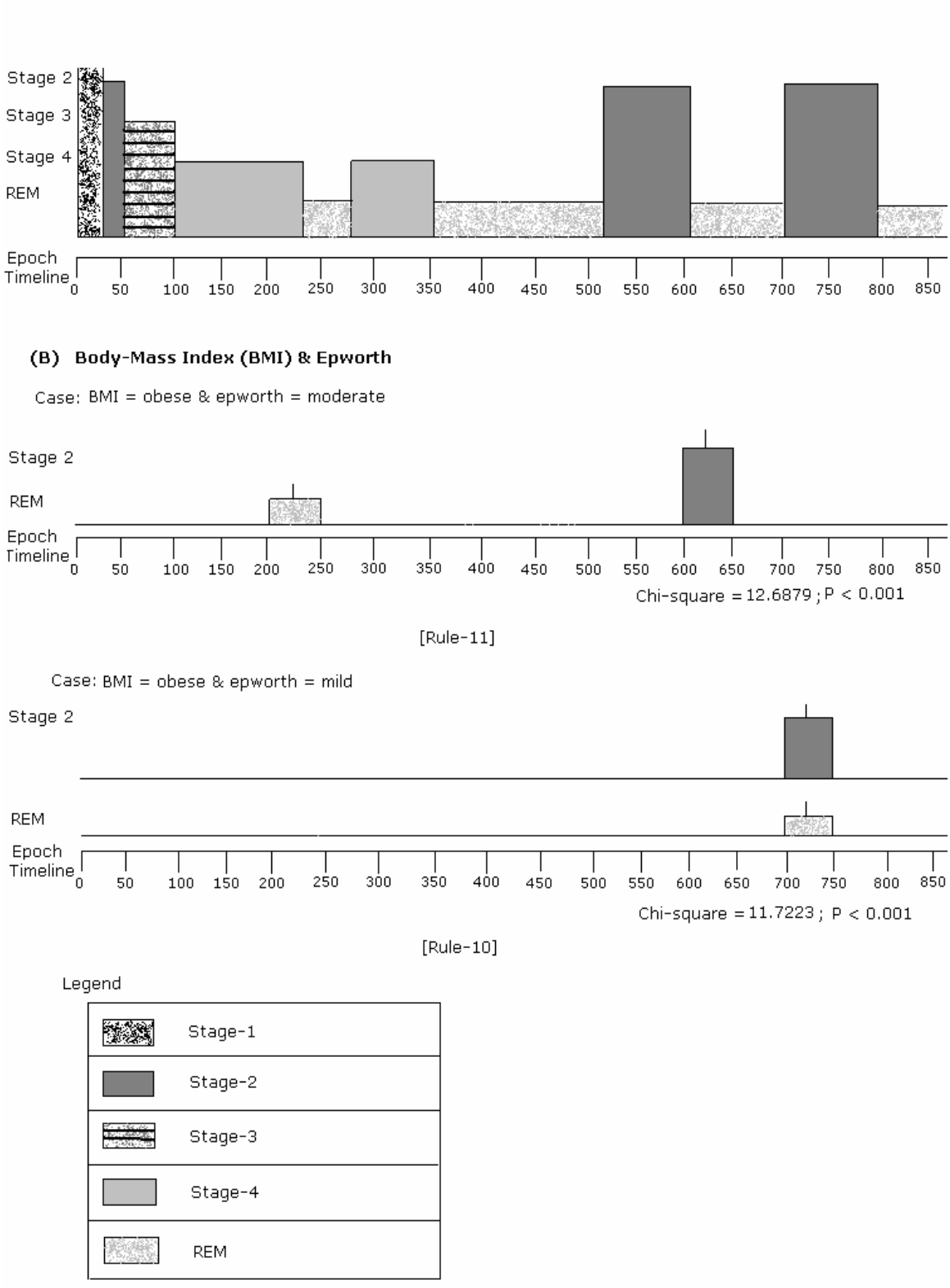


Fig.7.7: Most frequent windows of stage 2 and REM sleep noticed in obese patients having mild to moderate epworth indices (Experiment-6).

The above diagram illustrates the differences seen in obese patients who have mild to moderate epworth ratings. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. Unlike the previous experiment, REM can be used as an aid in distinguishing obese patients with mild from those with moderate epworth ratings. The most frequent REM window in obese patients with moderate epworth ratings is 200^{th} to 250^{th}. In contrast for obese patients with mild epworth ratings, the most frequent REM window is between the 700^{th} and 750^{th} epochs.

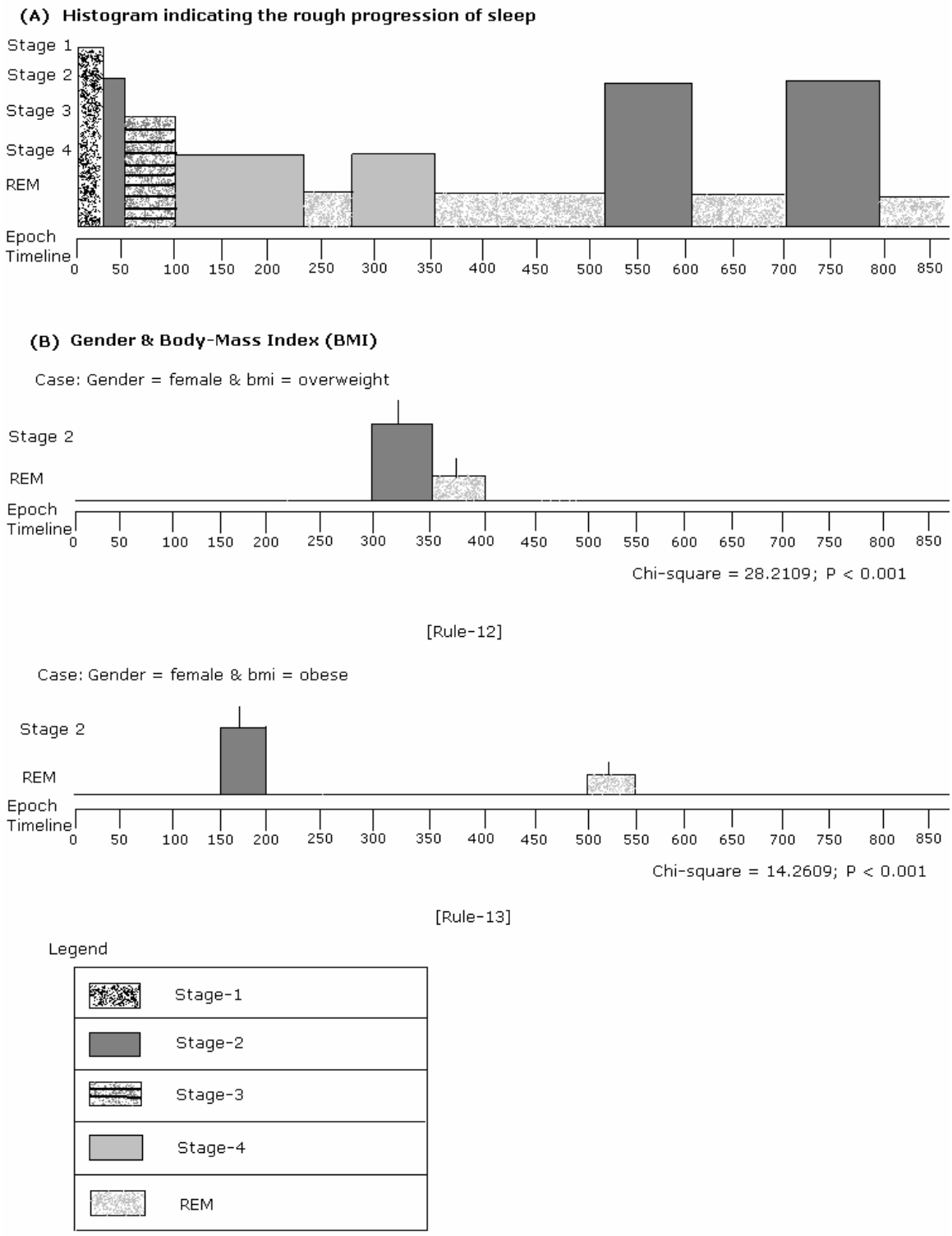


Fig. 7.8: Most frequent windows of stage 2 and REM sleep noticed in female patients who reported being overweight and obese (Experiment-6).

The above diagram illustrates two rules correlating stages 2 and REM of sleep with female patients who are either overweight or obese. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. The rule correlating stage 2 and REM with overweight female patients has a very high chi-square value. The high statistical significance of both these rules make it very interesting.

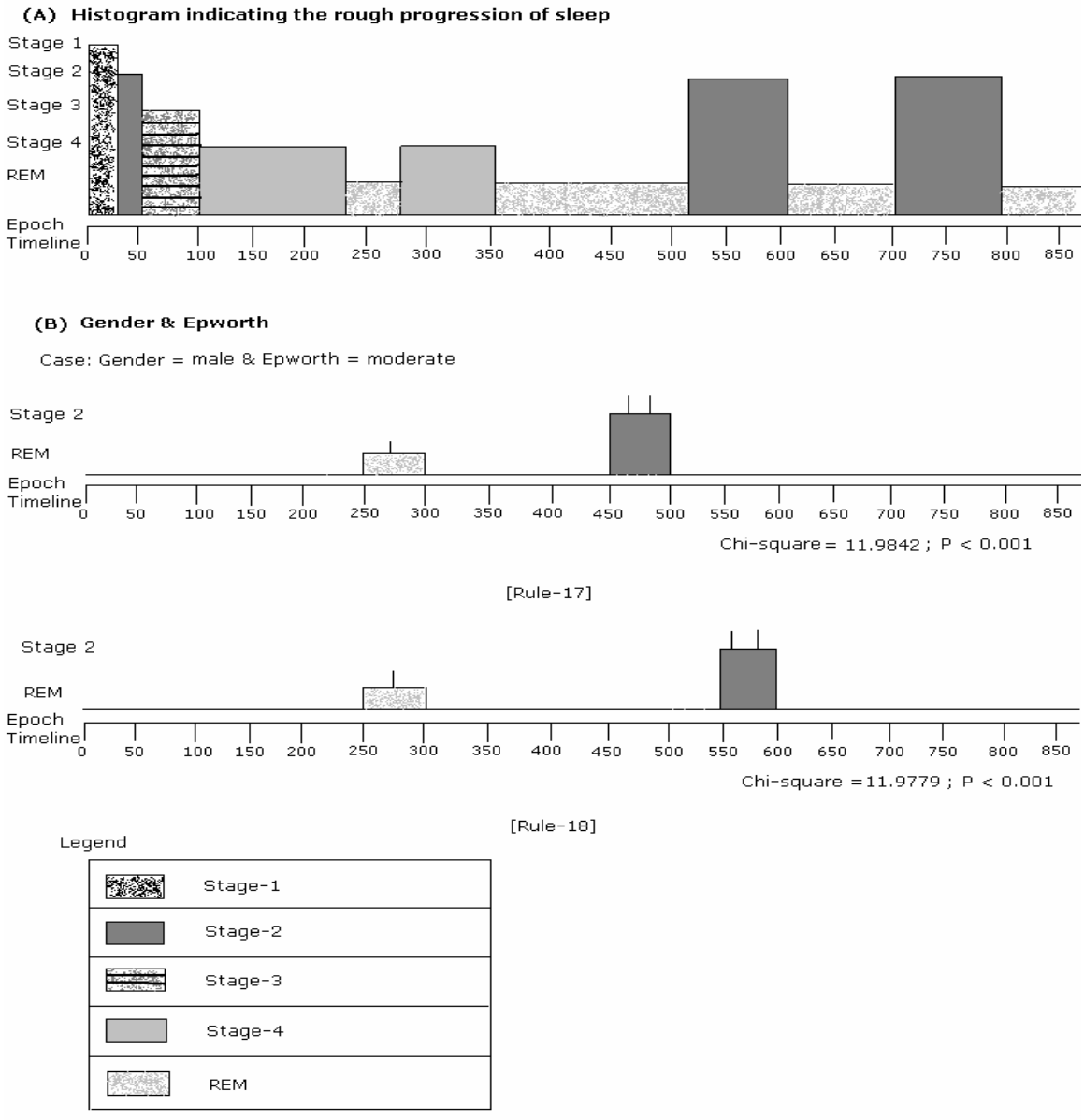


Fig. 7.9. Most frequent windows of stage 2 and REM sleep noticed in male patients with moderate epworth ratings (Experiment-6).

The above diagrammatic representation indicates the frequent windows of stage-2 and REM noticed in male patients with moderate epworth ratings. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. It is interesting to note that the window between 200th and 250th epochs shows presence of REM stage in both the rules. Besides, stage-2 with two fragments is also observed in the middle stages of sleep.

Experiment-7:

Objective:

Determine stage 2 and REM sleep variations that associate with gender, body-mass index, collar-size, caffeine-intake and exercise related information.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	8
Attribute Names:	stage-1, stage-2, stage-REM, gender, bmi-class, collar-size, exercise, caffeine
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	stage-1, stage-2, stage-REM
Required attributes in the consequent:	none
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage-1, stage-2, stage-REM
Epoch-window:	50

Sample of statistically significant rules obtained from association rule mining over the full training set:

1. stage-1 (750-800#1) && stage-2 (300-350#1) && stage-R (750-800#1) ➔ exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 3.8571; Chi-square = 15.2179; P<0.001]

2. stage-1 (750-800#1) && stage-2 (750-800#1) && stage-R (250-300#1) ➔ exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 3.8571; Chi-square = 15.2179; P<0.001]

3. stage-1 (750-800#1) && stage-2 (200-250#1) && stage-R (200-250#1) ➔ exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 3.8571; Chi-square = 15.2179; P<0.001]

4. stage-1 (300-350#1) && stage-2 (800-850#1) && stage-R (650-700#1) ➔ exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 3.8571; Chi-square = 15.2179; P<0.001]

5. stage-1 (400-450#1) && stage-2 (800-850#1) && stage-R (650-700#1) ➔ exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 3.8571; Chi-square = 15.2179; P<0.001]

6.stage-1 (750-800#1) && stage-2 (750-800#1) && stage-R (750-800#1) ➔ exercise (3-5 times/week)

[Confidence = 85.71%; Support = 7.41%; Lift = 3.3061; Chi-square = 14.2609; P<0.001]

7.stage-1 (750-800#1) && stage-2 (450-500#2) && stage-R (750-800#1) ➔ exercise (3-5 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]

8.stage-1 (550-600#1) && stage-2 (800-850#1) && stage-R (650-700#1) ➔ exercise (3-5 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1135; P<0.001]

9.stage-1 (250-300#1) && stage-2 (300-350#1) && stage-R (350-400#1) ➔ bmi-class (overweight)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9769; P<0.001]

10.stage-1 (250-300#1) && stage-2 (300-350#1) && stage-R (350-400#1) ➔ gender (female) && bmi-class (overweight)

[Confidence = 83.33%; Support = 6.17%; Lift = 7.5; Chi-square = 34.2049; P<0.001]

11. stage-1 (300-350#1) && stage-2 (800-850#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 7.3636; Chi-square = 33.8947; P<0.001]

12. stage-1 (300-350#1) && stage-2 (600-650#2) && stage-R (650-700#1) ➔ caffeine-intake(0cups/day)

[Confidence = 100%; Support = 6.17%; Lift = 6.2308; Chi-square = 27.861; P<0.001]

13. stage-1 (600-650#1) && stage-2 (600-650#2) && stage-R (650-700#1) ➔ caffeine-intake (0cups/day)

[Confidence = 83.33%; Support = 6.17%; Lift = 5.1923; Chi-square = 21.761; P<0.001]

14. stage-1 (400450#1) && stage-2 (600-650#2) && stage-R (650-700#1) ➔ caffeine-intake (0 cups/day)

[Confidence = 83.33%; Support = 6.17%; Lift = 5.1923; Chi-square = 21.761; P<0.001]

15. stage-1 (150-200#1) && stage-2 (100-150#1) && stage-R (750-800#1) ➔ caffeine-intake (1-3 cups/day)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.8571; Chi-square = 14.5696; P<0.001]

16. stage-1 (50-100#1) && stage-2 (500-550#1) && stage-R (750-800#1) ➔ bmi-class (obese) && caffeine-intake (1-3 cups/day)

[Confidence = 100%; Support = 6.17%; Lift = 3.24; Chi-square = 11.931; P<0.001]

17. stage-1 (500-550#1) && stage-2 (550-600#2) && stage-R (750-800#1) ➔ collar-size (16-18 inches)

[Confidence = 71.43%; Support = 6.17%; Lift = 3.8571; Chi-square = 14.2082; P<0.001]

18. stage-1 (200-250#2) && stage-2 (200-250#2) && stage-R (500-550#1) ➔ exercise (0 times/week)
[Confidence = 100%; Support = 6.17%; Lift = 2.3143; Chi-square = 7.0; 0.001<P<0.01]

Analyses of Rules:

Due to the large number of rules generated by the mining system, we present only the most important associations with high confidence and statistical robustness. We find interesting patterns of stage 1,2 and REM sleep variation that correlates highly with overweight problems, the per day caffeine intake or exercise frequency over the week of patients.

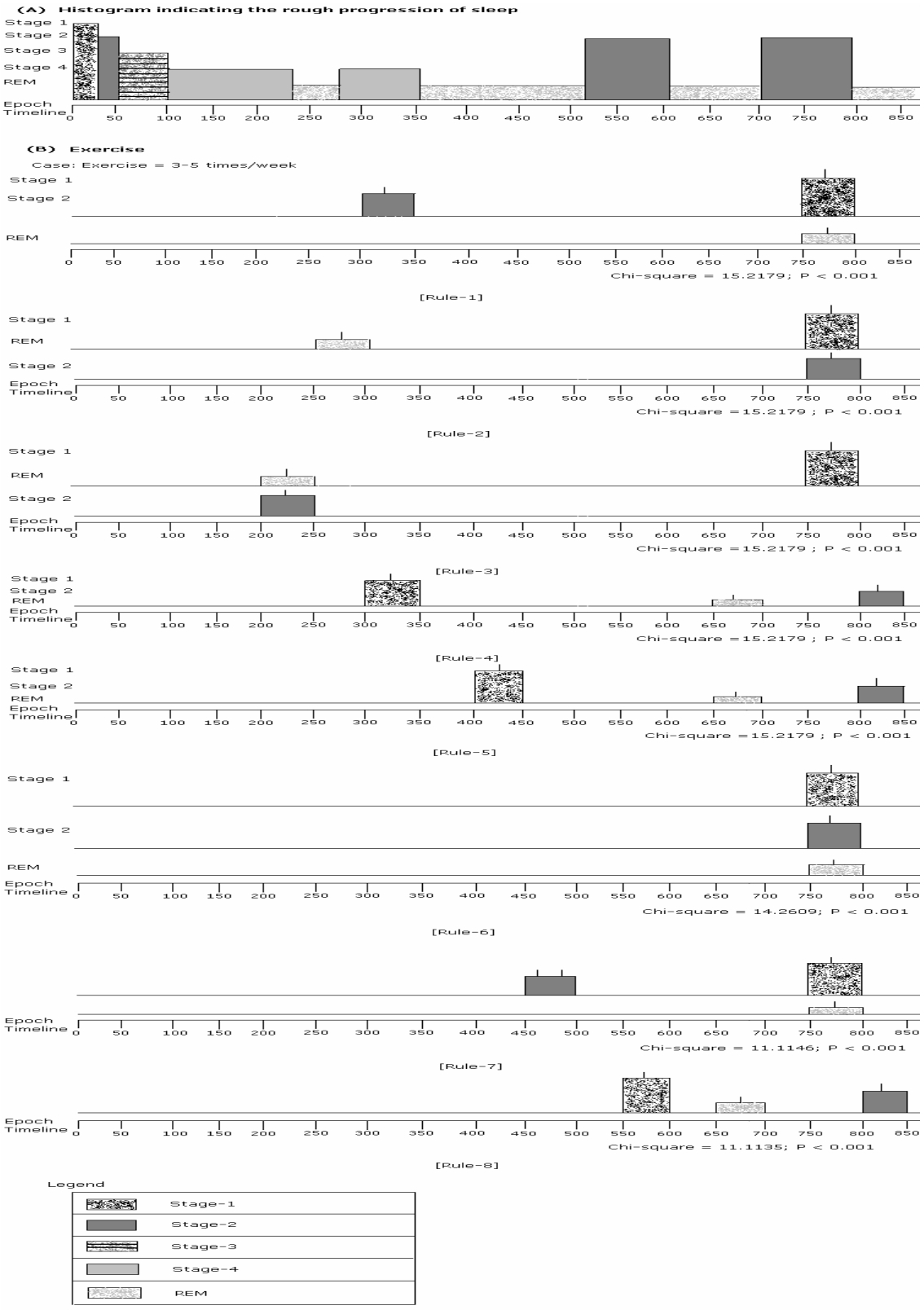


Fig. 7.10: Most frequent epoch windows with stages 1,2 and REM noticed in patients who exercise 3-5 times/week (Experiment-7).

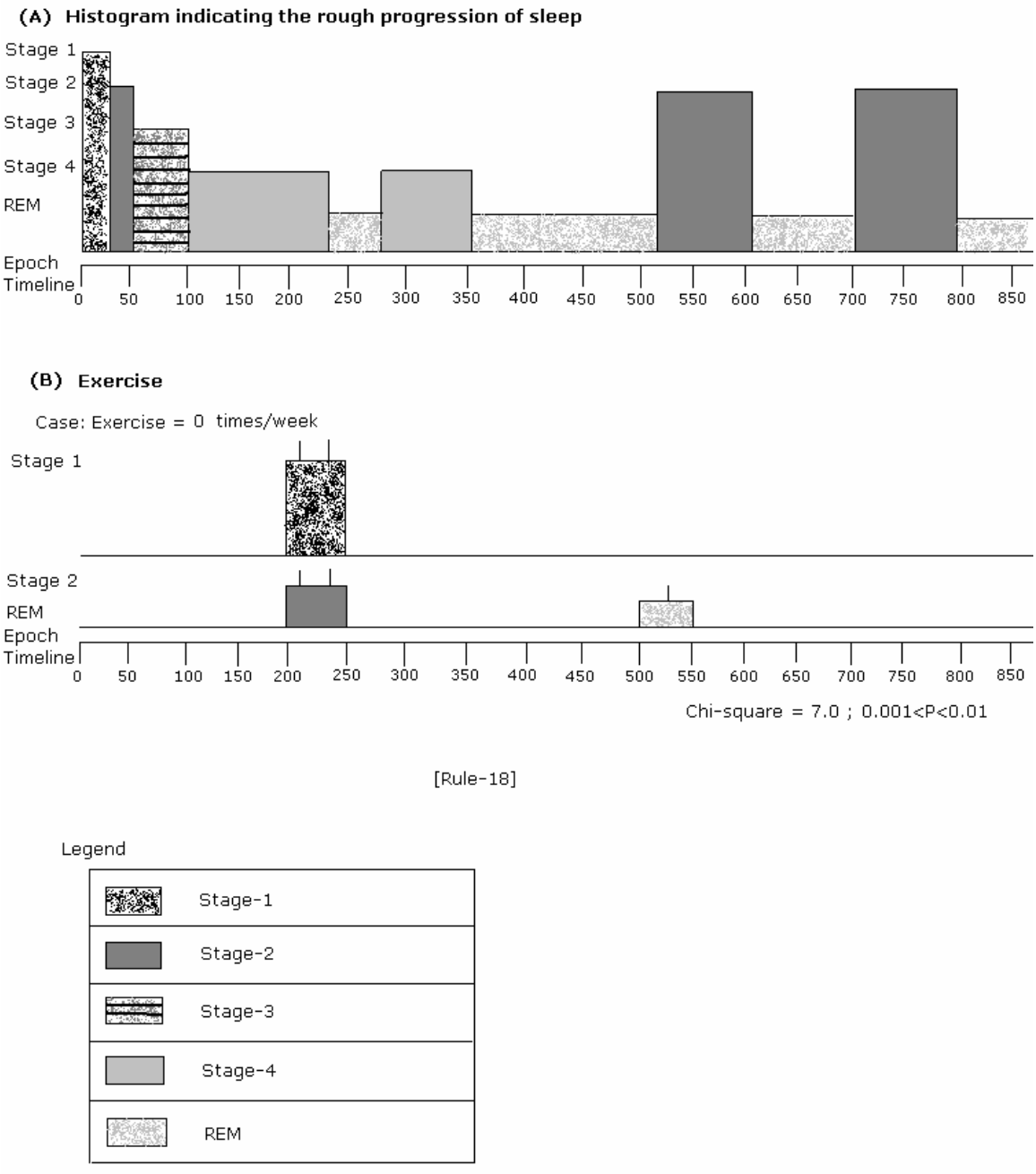


Fig. 7.11. Most frequent epoch windows with stages 1,2 and REM noticed in patients who do not exercise at all (Experiment-7).

Figures 7.10 and 7.11 diagrammatically illustrate frequent epoch windows of stages 1,2 and REM seen in patients who do not exercise at all and those who exercise fairly frequently. 'Case' in the above figure represents the consequent of the rules, while the

blocks represent the antecedent attributes in the rules. Consider the epoch window 200-250. Patients who do not exercise at all show presence of both stages 1 and 2 within this window. However, it is interesting to note that the stages have two fragmentations each. For patients who exercise 3-5 times/week while stages 1 and REM are noticed within the same epoch window, stage-1 is not frequently observed. Also for patients, who exercise frequently, stages 1and 2 are frequently noticed towards the terminal stages of sleep (750-800 epoch window). In contrast, stages 1and 2 are not frequently observed during the terminal stages of sleep in case of patients who do not exercise at all.

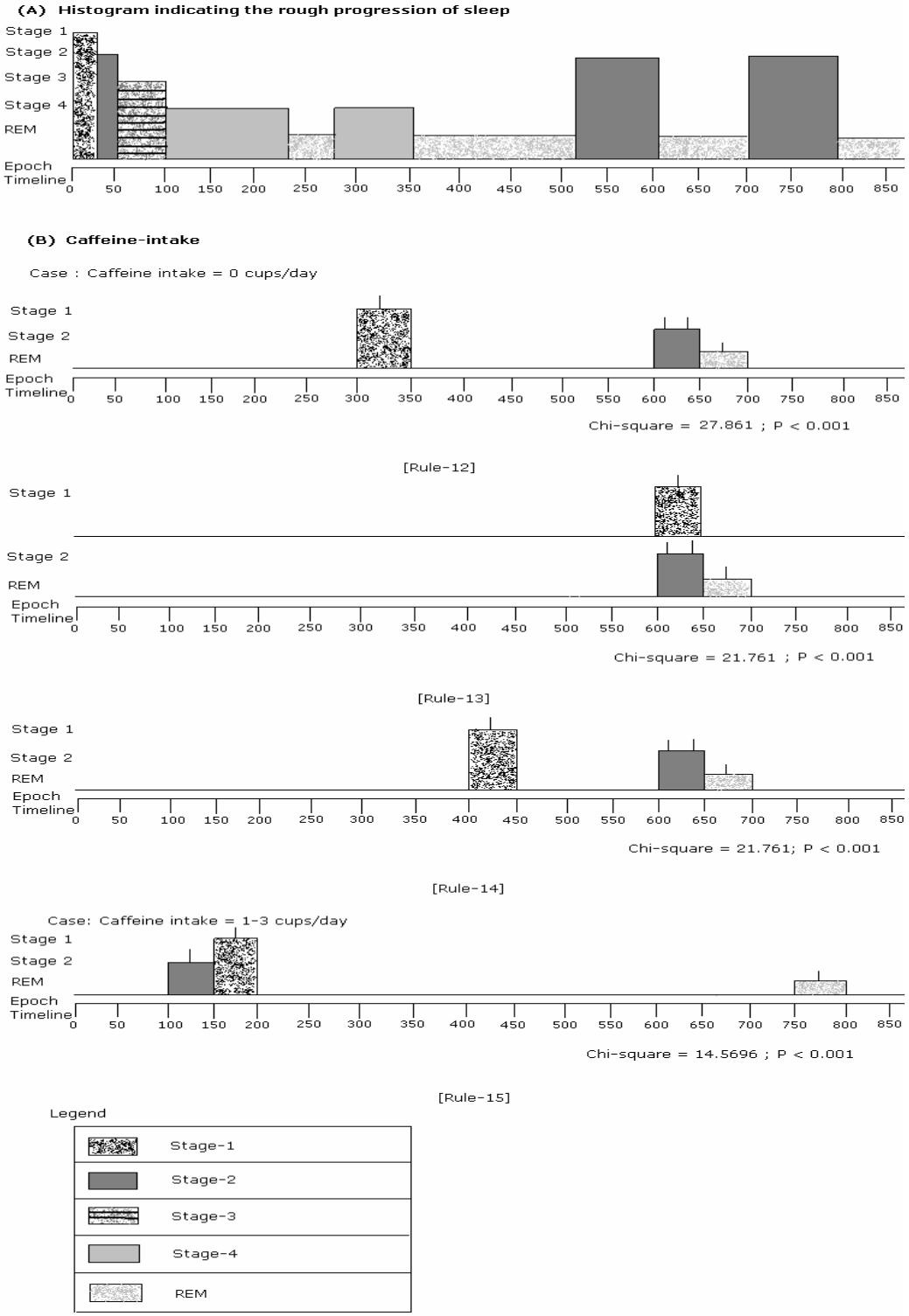


Fig. 7.12: Most frequently witnessed windows of stages 1,2 and REM in patients who do not drink caffeinated beverages and those who have little amounts of it each day (Experiment-7).

This experiment draws rules with strong correlations of caffeine intake on the basis of stages 1,2 and REM of sleep. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. Patients who do not have caffeinated beverages have a very high likelihood of witnessing stage-2 between the 600^{th} and 650^{th} epochs followed by REM between the 650^{th} and 700^{th} epochs. Stage-2 is witnessed with two fragments. As against this, for patients who have some (small quantity) intake of caffeinated beverages each day, stage-2 is more frequently observed during the 100^{th} to 150^{th} epoch window while REM is frequently seen between the 750^{th} and 800^{th} epoch window. Stages-2 and REM can be used to differentiate between patients taking caffeinated beverages from those that do not.

Experiment-8:

Objective:

Determine association of stage 1,2 and REM sleep with sleep efficiency, arousal-index, sleep latency till stage 1 and REM and mean oxygen potential.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	8
Attribute Names:	stage-1, stage-2, stage-REM, sleep-efficiency, arousal-index, sleep-latency-till-1 (sleep latency till stage-1), sleep-latency-till-REM (sleep latency till stage-REM), oxy-mean (mean oxygen potential)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	40%
Lower bound for support:	1%
Required attributes in the antecedent:	stage-1, stage-2, stage-REM
Required attributes in the consequent:	none
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage-1, stage-2, stage-REM

Epoch-window: 50

Sample of statistically significant rules obtained from association rule mining over the full training set:

1. stage-1 (450-500#2) && stage-2 (800-850#1) && stage-REM (600-650#1) ➔ sleep-efficiency (normal)
   [Confidence = 100%; Support = 6.17%; Lift = 3.5217; Chi-square = 13.43; P<0.001]
2. stage-1 (600-650#1) && stage-2 (650-700#1) && stage-R (450-500#1) ➔ sleep-efficiency (normal)
   [Confidence = 100%; Support = 6.17%; Lift = 3.5217; Chi-square = 13.4314; P<0.001]
3. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-R (800-850#1) ➔ sleep-efficiency (low)
   [Confidence = 71.43%; Support = 6.17%; Lift = 3.4034; Chi-square = 11.7503; P<0.001]
4. stage-1 (200-250#1) && stage-2 (200-250#1) && stage-R (750-800#1) ➔ sleep-latency-till-1 (short)
   [Confidence = 100%; Support = 6.17%; Lift = 5.0625; Chi-square = 21.6382; P<0.001]
5. stage-1 (0-50#1) && stage-2 (0-50#1) && stage-2 (400-450#1) && stage-R (200-250#1) ➔ sleep-latency-till-REM (prolonged) && oxy-mean (normal)
   [Confidence = 100%; Support = 6.17%; Lift = 3.375; Chi-square = 12.65; P<0.001]
6. stage-1 (750-800#1) && stage-2 (200-250#1) && stage-R (600-650#1) ➔ sleep-efficiency (mild) && sleep-latency-till-REM (prolonged)
   [Confidence = 83.33%; Support = 6.17%; Lift = 4.5; Chi-square = 18.0309; P<0.001]
7. stage-1 (0-50#1) && stage-2 (0-50#1) && stage-2 (400-450#1) && stage-R (650-700#1) ➔ sleep-efficiency (mild) && sleep-latency-till-1 (normal)
   [Confidence = 100%; Support = 6.17%; Lift = 4.5; Chi-square = 18.6422; P<0.001]
8. stage-1 (50-100#1) && stage-2 (50-100#1) && stage-R (450-500#1) ➔ arousal-index (moderate)
   [Confidence = 100%; Support = 6.17%; Lift = 3.24; Chi-square = 11.931; P<0.001]
9. stage-1 (0-50#1) && stage-2 (400-450#1) && stage-R (650-700#1) ➔ sleep-latency-till-1 (normal) && sleep-latency-till-REM (prolonged)
   [Confidence =83.33%; Support = 6.17%; Lift = 4.5; Chi-square = 18.0309; P<0.001]
10. stage-1 (500-550#1) && stage-2 (400-450#1) && stage-R (200-250#1) ➔ arousal-index (mild) && oxy-mean (normal)
    [Confidence = 100%; Support = 6.17%; Lift = 3.375; Chi-square = 12.65; P<0.001]

Analyses of Rules:

Like the previous experiment, this experiment also produces a number of interesting rules since it covers conditions, which have an increased likelihood of inducing sleep disorders. All these rules have a very high confidence besides being statistically robust.

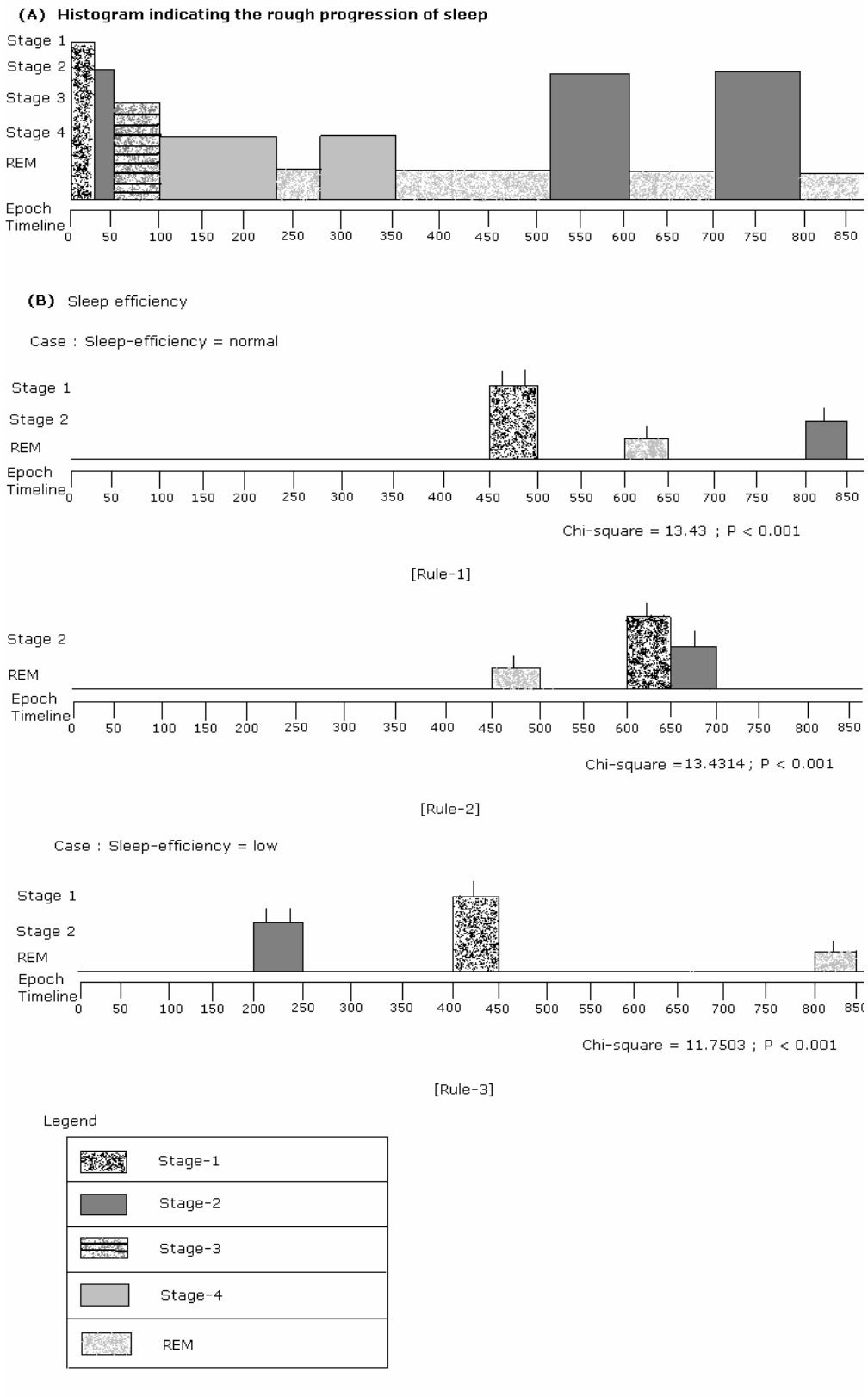


Fig. 7.13. Sleep-efficiency variations with stages 1,2 and REM (Experiment-8)

From the figure above we notice that sleep-efficiency is reduced in patients who witness stage-2 of sleep with two fragmentations between the 200^{th} and 250^{th} epochs. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. The fragments possibly cause the patients to go to a wake state resulting in lowering of sleep-efficiency. In case of normal sleep-efficiency, stage-2 is witnessed during the terminal stages of sleep.

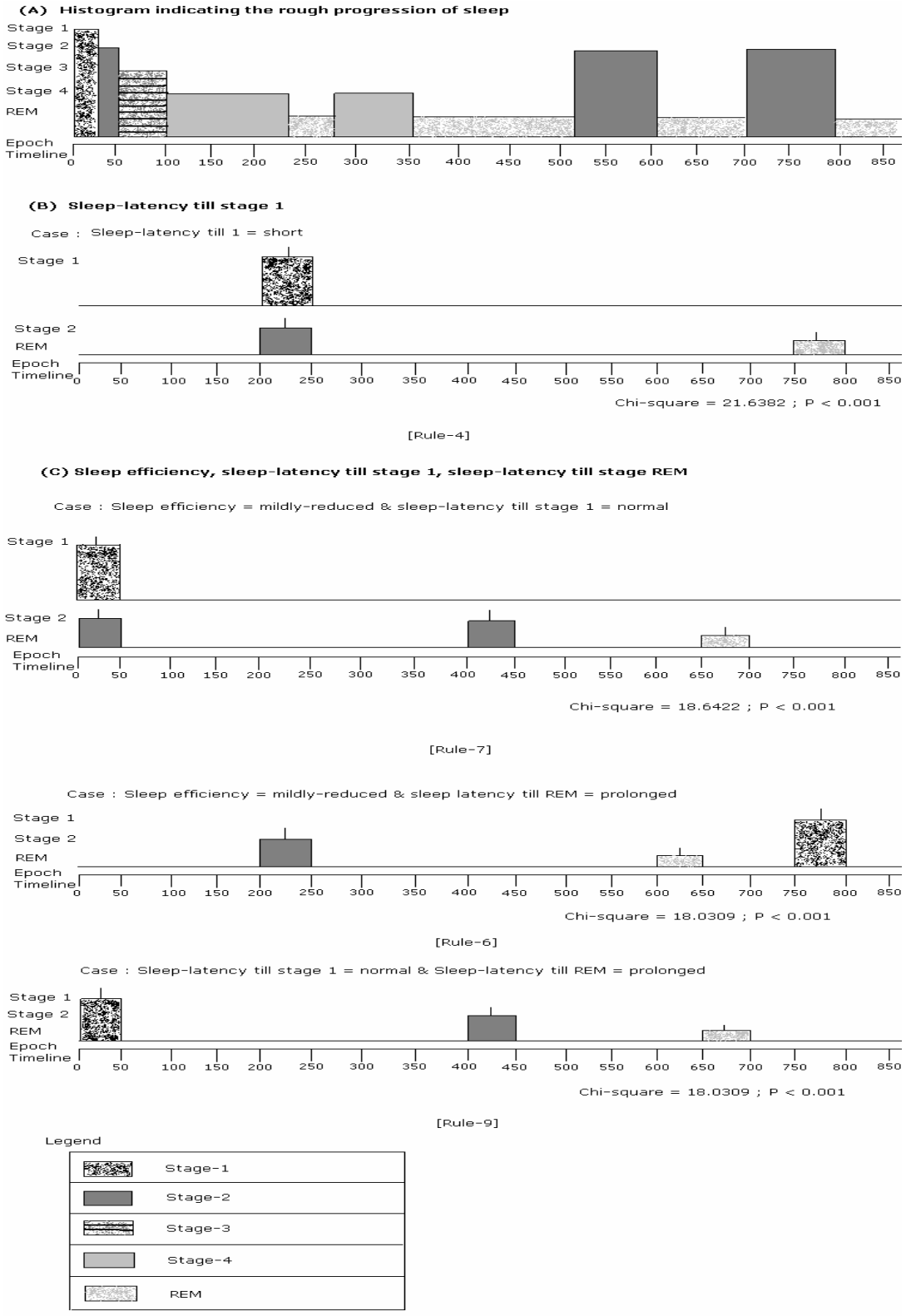


Fig.7.14. Most frequent epoch windows of stages 1,2 and REM of sleep that correlate with sleep latency till stage 1, sleep latency till REM and sleep efficiency (Experiment-8)

The above figure depicts frequent epoch windows with stages 1,2 and REM that have high correlation with different permutations that can be conjured with sleep latencies to stage 1 and REM and sleep-efficiency. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. The rules have very high chi-square values and are hence statistically robust.

Experiment-9:

Objective:

Determine association of stage 1,2,3,4 and REM sleep with male or female patients with varying exercise habits.

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	7
Attribute Names:	stage-1, stage-2, stage-3, stage-4, stage-REM, gender, exercise
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	gender, exercise
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage-1, stage-2, stage-3, stage-4, stage-REM
Epoch-window:	50

Sample of statistically significant rules obtained from association rule mining over the full training set:

1. stage-2 (600-650#1) && stage-2 (800-850#1) ➔ gender (male) && exercise (0 times/week)
   [Confidence = 56.25%; Support = 11.11%; Lift = 2.398; Chi-square = 11.9404; P<0.001]
2. stage-1 (550-600#1) && stage-2 (600-650#1) && stage-2 (800-850#1) ➔ gender (male) && exercise (0 times/week)

[Confidence = 75%; Support = 7.41%; Lift = 3.1974; Chi-square = 13.1399; P<0.001]

3. stage-1 (200-250#2) && stage-2 (600-650#1) && stage-2 (800-850#1) ➔ gender (male) && exercise (0 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 4.2632; Chi-square = 17.3809; P<0.001]

4. stage-1 (300-350#1) && stage-2 (600-650#1) && stage-2 (800-850#1) ➔ gender (male) && exercise (0 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.5526; Chi-square = 12.9327; P<0.001]

5. stage-1 (350-400#2) && stage-2 (600-650#1) && stage-2 (800-850#1) ➔ gender (male) && exercise (0 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 4.2632; Chi-square = 17.3809; P<0.001]

6. stage-2 (550-600#1) && stage-2 (600-650#1) && stage-2 (800-850#1) ➔ gender (male) && exercise (0 times/week)

[Confidence = 60%; Support = 7.41%; Lift = 3.1974; Chi-square = 12.7301; P<0.001]

7. stage-2 (550-600#1) && stage-2 (600-650#1) && stage-2 (800-850#1) && stage-R (600-650#1) ➔ gender (male) && exercise (0 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.1974; Chi-square = 11.0238; P<0.001]

8. stage-2 (200-250#2) && stage-3 (150-200#1) && stage-3 (400-450#1) ➔ gender (female) && exercise (0 times/week)

[Confidence = 75%; Support = 7.41%; Lift = 3.0375; Chi-square = 12.0869; P<0.001]

9. stage-1 (700-750#1) && stage-2 (200-250#2) && stage-3 (400-450#1) ➔ gender (female) && exercise (0 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 4.05; Chi-square = 16.2453; P<0.001]

10. stage-2 (200-250#2) && stage-2 (500-550#2) && stage-2 (700-750#2) ➔ gender (female) && exercise (0 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9779; P<0.001]

11. stage-1 (700-750#1) && stage-2 (300-350#1) && stage-3 (400-450#1) ➔ gender (female) && exercise (0 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9779; P<0.001]

12. stage-1 (650-700#1) && stage-2 (600-650#2) && stage-2 (700-750#2) ➔ gender (female) && exercise (0 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9779; P<0.001]

13. stage-2 (750-800#1) && stage-R (250-300#1) ➔ gender (male) && exercise (3-5 times/week)

[Confidence = 55.56%; Support = 6.17%; Lift = 4.5; Chi-square = 17.4602; P<0.001]

14. stage-2 (550-600#2) && stage-R (450-500#1) ➔ gender (male) && exercise (3-5 times/week)

[Confidence = 71.43%; Support = 6.17%; Lift = 5.7857; Chi-square = 24.7034; P<0.001]

15. stage-1 (750-800#1) && stage-2 (450-500#2) && stage-2 (750-800#1) ➔ gender (male) && exercise (3-5 times/week)

[Confidence = 55.56%; Support = 6.17%; Lift = 4.05; Chi-square = 14.9665; P<0.001]

16. stage-2 (200-250#1) && stage-2 (300-350#1) && stage-3 (150-200#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 62.5%; Support = 6.17%; Lift = 4.6023; Chi-square = 18.0918; P<0.001]

17. stage-1 (550-600#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 71.43%; Support = 6.17%; Lift = 5.2597; Chi-square = 21.8368; P<0.001]

18. stage-2 (800-850#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 58.33%; Support = 8.64%; Lift = 4.2955; Chi-square = 24.0344; P<0.001]

19. stage-1 (550-600#1) && stage-2 (800-850#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 83.33%; Support = 6.17%; Lift = 6.1364; Chi-square = 26.8514; P<0.001]

20. stage-3 (750-800#1) && stage-R (600-650#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 50%; Support = 6.17%; Lift = 3.6818; Chi-square = 12.8869; P<0.001]

21. stage-1 (300-350#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 71.43%; Support = 6.17%; Lift = 5.2597; Chi-square = 21.8368; P<0.001]

22. stage-1 (300-350#1) && stage-2 (800-850#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 100%; Support = 6.17%; Lift = 7.3636; Chi-square = 33.8947; P<0.001]

23. stage-3 (150-200#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 50%; Support = 6.17%; Lift = 3.6818; Chi-square = 12.8869; P<0.001]

24. stage-2 (800-850#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 62.5%; Support = 6.17%; Lift = 4.6023; Chi-square = 18.0918; P<0.001]

25. stage-1 (550-600#1) && stage-2 (450-500#1) ➔ gender (female) && exercise (3-5 times/week)

[Confidence = 50%; Support = 6.17%; Lift = 3.6818; Chi-square = 12.8869; P<0.001]

Analysis of Results:

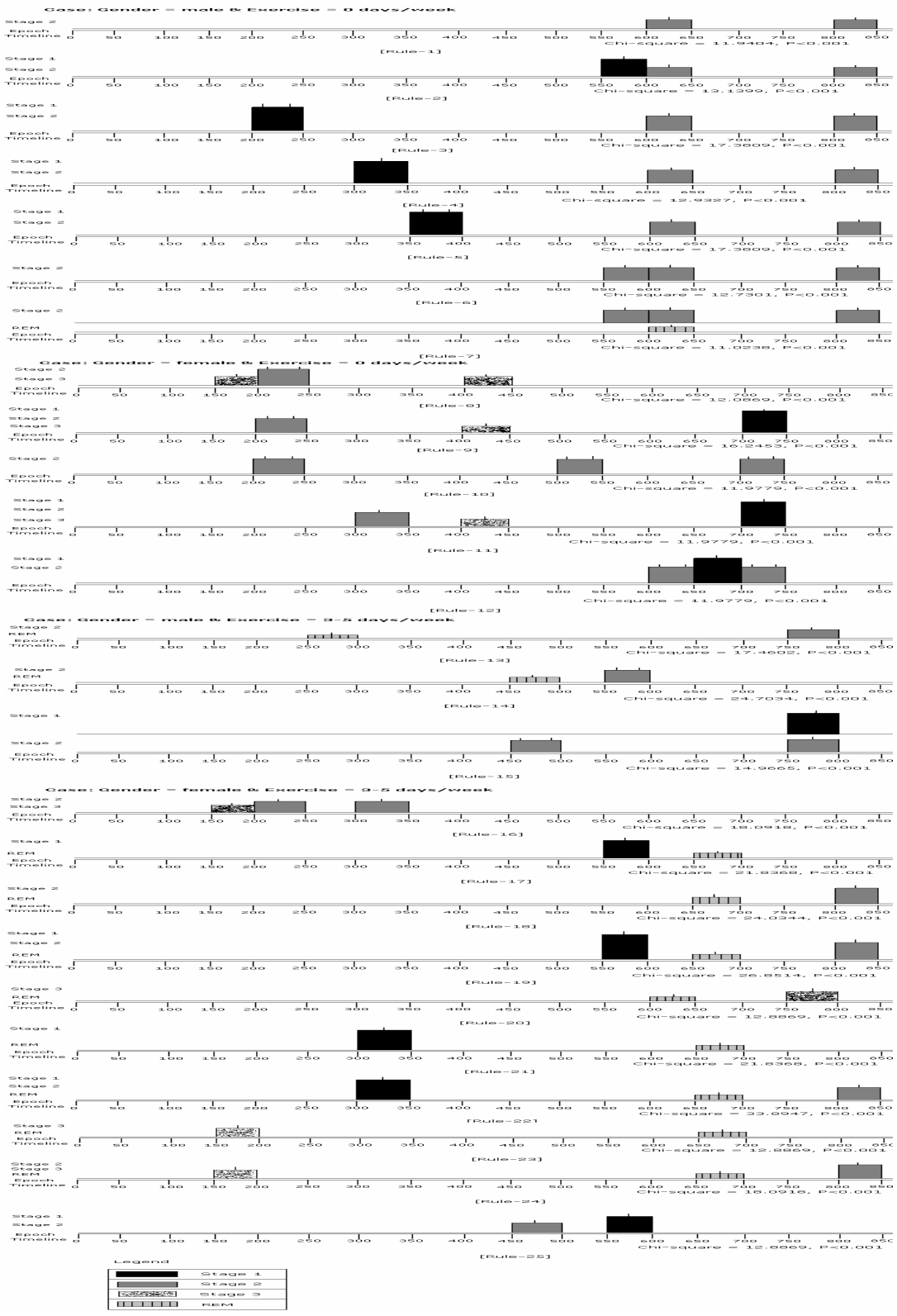


Fig.7.15. Variations in stages 1,2,3,4 and REM correlating with male and female patients with varying exercise habits (Experiment 9).

'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. For male patients who do not exercise at all, rules with highest chi-square values indicate two fragments of stage-1 occurring frequently towards the early part of sleep. Similarly in female patients who do not exercise at all, we notice two fragments of stage-2 during the early part of sleep. These fragments indicate disturbances in normal sleep pattern. The result hints at lack of exercise being a reason for disturbed sleep patterns. Also we notice that REM during the middle and towards the latter periods of sleep is frequently noticed in patients who exercise more frequently. This hints at a more normal sleep pattern being witnessed in patients who exercise fairly frequently (say 3 to 5 days in a week).

Experiment-10:

Objective:

Determine association of stage 1,2,3,4 and REM sleep with Obstructive Sleep Apnea (OSA).

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	6
Attribute Names:	stage-1, stage-2, stage-3, stage-4, stage-R (REM stage), OSA (Obstructive Sleep Apnea)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	OSA
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage-1, stage-2, stage-3, stage-4, stage-R
Epoch-window:	50

Sample of statistically significant rules obtained from association rule mining over the full dataset:

1. stage-3 (400-450#1) && stage-3 (500-550#1) ➔ OSA (mild)
   [Confidence = 62.5%; Support = 6.17%; Lift = 4.2634; Chi-square = 16.2308; P<0.001]
2. stage-1 (500-550#1) && stage-3 (400-450#1) && stage-3 (500-550#1) ➔ OSA (mild)
   [Confidence = 83.33%; Support = 6.17%; Lift = 4.2634; Chi-square = 16.7577; P<0.001]
3. stage-1 (200-250#2) && stage-1 (750-800#2) && stage-2 (200-250#2) ➔ OSA (mild)
   [Confidence = 100%; Support = 6.17%; Lift = 4.2632; Chi-square = 17.3809; P<0.001]
4. stage-2 (150-200#2) && stage-2 (800-850#1) && stage-3 (300-350#1) ➔ OSA (mild)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.5526; Chi-square = 12.9327; P<0.001]
5. stage-2 (200-250#1) && stage-3 (400-450#1) ➔ OSA (mild)
   [Confidence = 85.71%; Support = 7.41%; Lift = 3.6541; Chi-square = 16.5468; P<0.001]
6. stage-1 (150-200#1) && stage-2 (50-100#2) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
7. stage-1 (150-200#1) && stage-2 (150-200#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
8. stage-1 (150-200#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
9. stage-2 (150-200#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)
   [Confidence = 63.64%; Support = 8.64%; Lift = 2.8636; Chi-square = 12.6274; P<0.001]
10. stage-1 (150-200#1) && stage-2 (150-200#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
11. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-3 (50-100#1) ➔ OSA (moderately-severe)
   [Confidence = 63.64%; Support = 8.64%; Lift = 2.8636; Chi-square = 12.6307; P<0.001]
12. stage-1 (400-450#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]
13. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]
14. stage-1 (400-450#1) && stage-1 (700-750#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
15. stage-1 (400-450#1) && stage-1 (700-750#1) && stage-2 (200-250#2) && stage-3 (50-100#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
16. stage-1 (500-550#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]
17. stage-2 (700-750#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

18. stage-2 (700-750#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)
[Confidence = 75%; Support = 7.41%; Lift = 3.0; Chi-square = 11.8402; P<0.001]

19. stage-2 (700-750#1) && stage-3 (50-100#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)
[Confidence = 100%; Support = 6.17%; Lift = 4.5; Chi-square = 18.6422; P<0.001]

20. stage-2 (700-750#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)
[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

21. stage-3 (50-100#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)
[Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]

22. stage-2 (700-750#1) && stage-3 (50-100#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)
[Confidence = 100%; Support = 6.17%; Lift = 3.75; Chi-square = 14.6474; P<0.001]

23. stage-1 (350-400#1) && stage-1 (700-750#1) ➔ OSA (moderately-severe)
[Confidence = 66.67%; Support = 7.41%; Lift = 3.0;Chi-square = 11.5761; P<0.001]

24. stage-1 (300-350#1) && stage-2 (700-750#1) && stage-3 (500-550#1) ➔ OSA (moderately-severe)
[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

25. stage-2 (150-200#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)
[Confidence = 75%; Support = 7.41%; Lift = 3.375; Chi-square = 14.3113; P<0.001]

26. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-2 (300-350#2) ➔ OSA (moderately-severe)
[Confidence = 100%; Support = 6.17%; Lift = 4.5; Chi-square = 18.6422; P<0.001]

Analysis of Results:

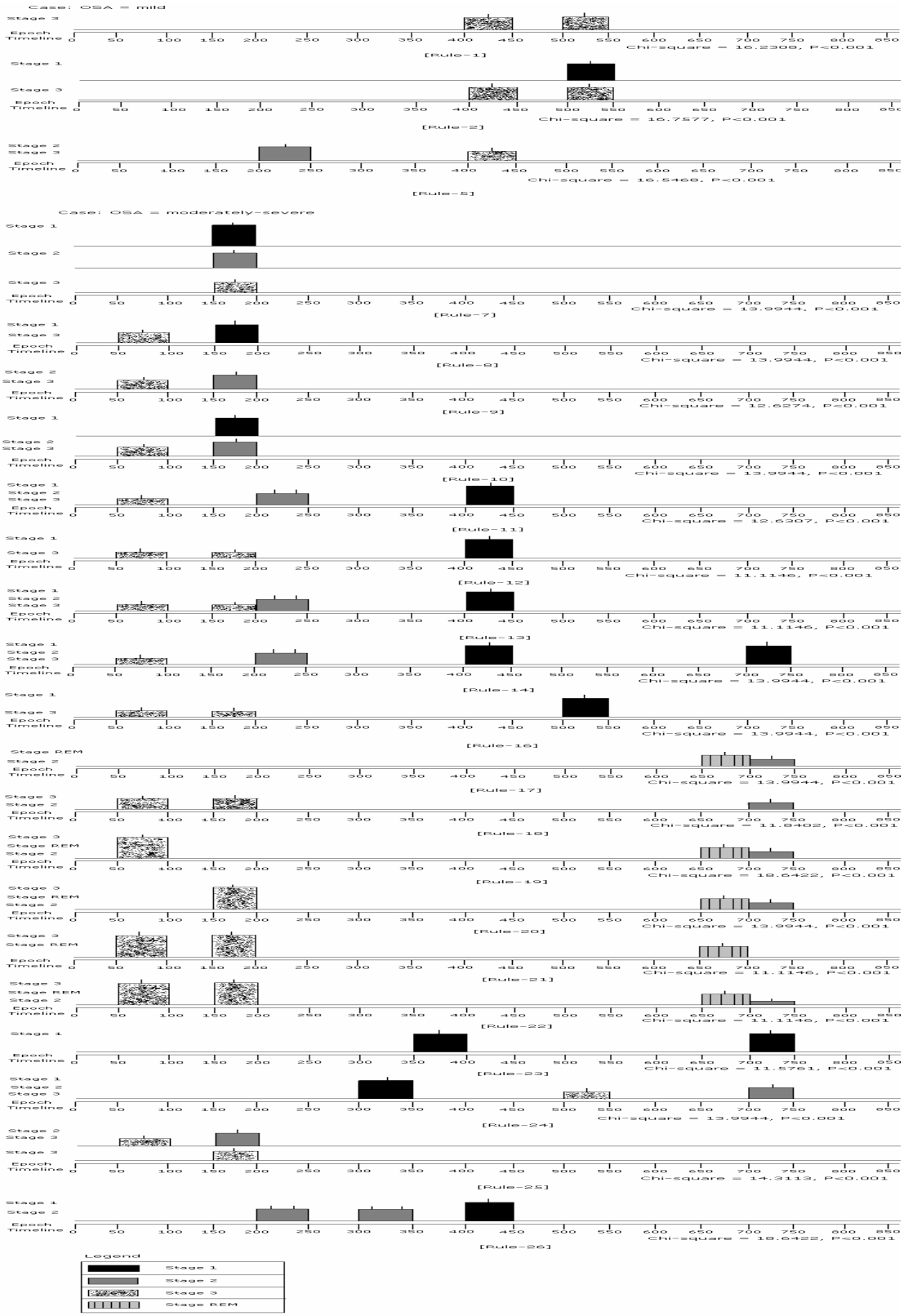


Fig.7.16. Correlation between stages-1,2,3,4, REM of sleep with varying degrees of OSA Experiment 10).

'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. From the above diagram it is clearly seen that in patients with severe cases of Obstructive Sleep Apnea (OSA) stage-3 of sleep is frequently observed during the early epochs of sleep. This may be attributed to administration of CPAP or conducting of split-night study thus causing the patients to enter into deep sleep earlier than usual (which is what is seen for cases where patients are diagnosed with mild OSA). Another interesting observation is the presence of two fragments of stage-2 between the 200^{th} and 250^{th} epochs of sleep in patients with moderately severe OSA. The fragmented stage-2 window is indicative of fractured sleep typical in patients suffering from severe cases of OSA. To determine if more interesting patterns can be obtained within a smaller window we conduct the same experiment with a window size of 25.

Experiment-11:

Objective:

Determine association of stage 1,2,3,4 and REM sleep with Obstructive Sleep Apnea (OSA) (Epoch Window = 25)

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	6
Attribute Names:	stage-1, stage-2, stage-3, stage-4, stage-R (REM stage), OSA (Obstructive Sleep Apnea)
Minimum number of rules sought:	10
Minimum confidence:	50%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	OSA
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage-1, stage-2, stage-3, stage-4, stage-R

Epoch-window: 25

Sample of statistically significant rules obtained from association rule mining over the full dataset:

1. stage-1 (525-550#1) && stage-3 (525-550#1) ➔ OSA (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 3.1974; Chi-square = 13.14; P<0.001]
2. stage-1 (675-700#1) && stage-3 (475-500#1) ➔ OSA (mild)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.5526; Chi-square = 12.9327; P<0.001]
3. stage-1 (550-575#1) && stage-3 (475-500#1) ➔ OSA (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 3.1974; Chi-square = 13.14; P<0.001]
4. stage-1 (550-575#1) && stage-1 (675-700#1) && stage-3 (475-500#1) ➔ OSA (mild)
   [Confidence = 100%; Support = 6.17%; Lift = 4.2632; Chi-square = 17.3809; P<0.001]
5. stage-1 (100-125#1) && stage-1 (425-450#1) && stage-2 (425-450#1) ➔ OSA (mild)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.5526; Chi-square = 12.9327; P<0.001]
6. stage-1 (550-575#1) && stage-3 (500-525#1) ➔ OSA (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 3.1974; Chi-square = 13.14; P<0.001]
7. stage-2 (425-450#1) && stage-3 (50-75#1) && stage-3 (150-175#1) ➔ OSA (moderately-severe)
   [Confidence = 66.67%; Support = 7.41%; Lift = 3.0; Chi-square = 11.5761; P<0.001]
8. stage-2 (175-200#1) && stage-2 (550-575#1) && stage-3 (125-150#1) ➔ OSA (moderately-severe)
   [Confidence = 66.67%; Support = 7.41%; Lift = 3.0; Chi-square = 11.5761; P<0.001]
9. stage-2 (175-200#1) && stage-2 (725-750#1) && stage-3 (125-150#1) ➔ OSA (moderately-severe)
   [Confidence = 66.67%; Support = 7.41%; Lift = 3.0; Chi-square = 11.5761; P<0.001]
10. stage-1 (750-775#1) && stage-2 (300-325#1) ➔ OSA (moderately-severe)
    [Confidence = 70%; Support = 8.64%; Lift = 3.15; Chi-square = 15.0634; P<0.001]

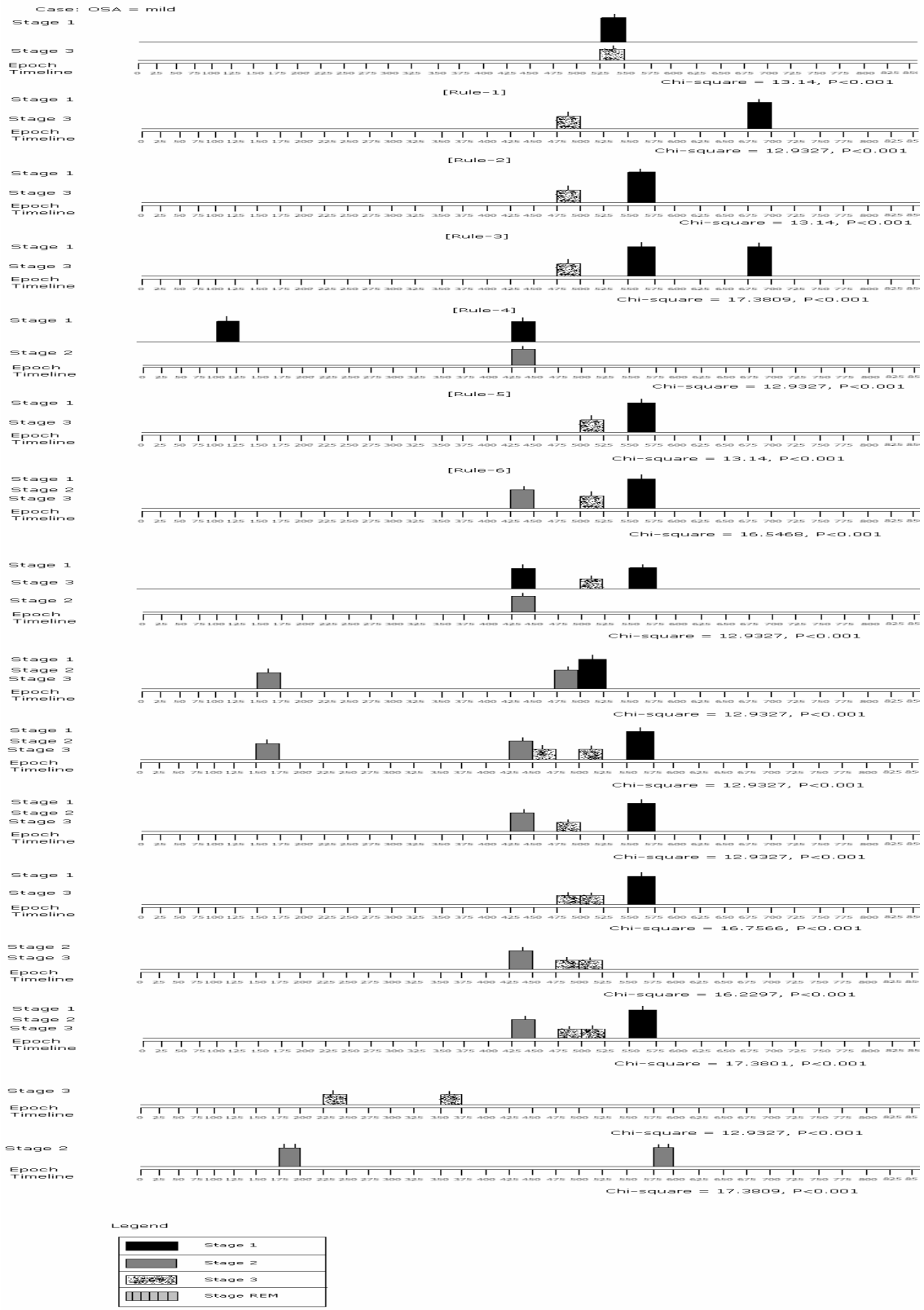


Fig.7.17 Correlation between mild-OSA and sleep stages observed with window-size=25 epochs (Experiment 11).

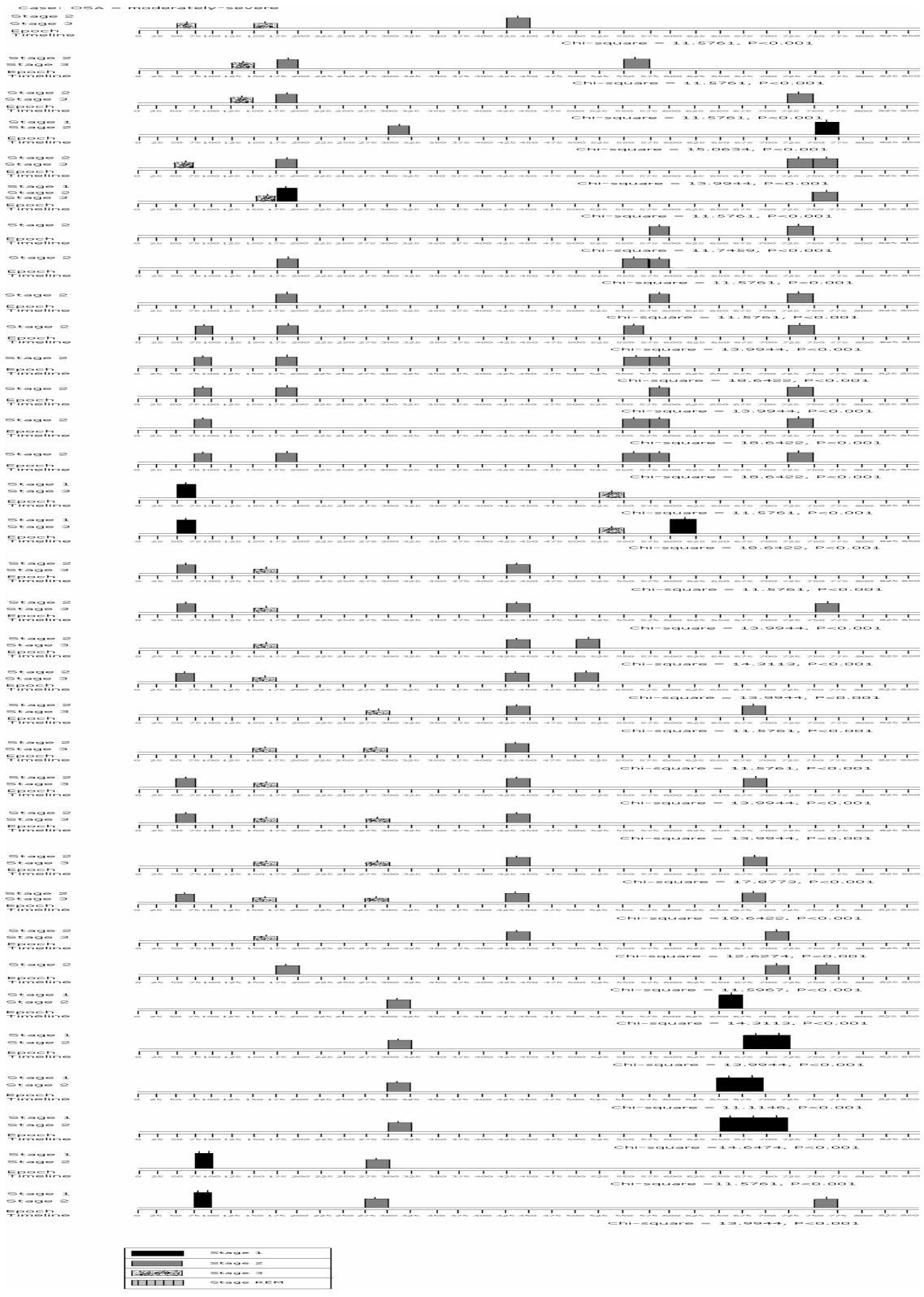


Fig. 7.18. Correlation between mild-OSA and sleep stages observed with window-size = 25 epochs (Experiment 11).

From the figures 7.17 and 7.18 we see a pattern similar to that noticed with a larger window of 50 epochs as in experiment 9. 'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. In other words, we notice the same phenomenon of stage-3 being prevalent during the early hours of sleep in patients with increased severity of OSA. The reason for this is the same as explained in the analysis of experiment-9. In addition we find another interesting observation. For patients suffering from severe cases of OSA we obtain a rule in which stage-3 of sleep in a particular window is immediately followed by stage-1. This is suggestive of a disturbance large enough to cause awakening resulting in fractured sleep. In case of patients with mild OSA symptoms, stage-3 and stage-1 may overlap within the same window. This suggests that the fragmentation may probably be very small reducing the chances of patients from suffering a fractured sleep pattern.

Experiment-12:

Objective:

Determine association of stage 1,2,3,4 and REM sleep with Obstructive Sleep Apnea (OSA) (Epoch Window = 25)

Input Parameters:

Instances:	81 (Each instance corresponds to one patient)
Total number of attributes:	4
Attribute Names:	stage-1, stage-2, sleep-eff (sleep efficiency), OSA (Obstructive Sleep Apnea)
Minimum number of rules sought:	10
Minimum confidence:	90%
Upper bound for support:	30%
Lower bound for support:	1%
Required attributes in the antecedent:	none
Required attributes in the consequent:	sleep-eff
Disallowed attributes in the antecedent:	none
Disallowed attributes in the consequent:	stage-1, stage-2, OSA
Epoch-window:	50

Sample of statistically significant rules obtained from association rule mining over the full dataset:

1. stage-1 (500-550#1) && stage-1 (700-750#1) && stage-2 (200-250#2) ➔ sleep-eff (low)
   [Confidence = 100%; Support = 6.17%; Lift = 4.7647; Chi-square = 20.0521; P<0.001]
2. stage-2 (300-350#2) && stage-2 (550-600#1) ➔ sleep-eff (mildly-reduced)
   [Confidence = 90%; Support = 11.11%; Lift = 2.2781; Chi-square = 12.1694; P<0.001]
3. stage-1 (0-50#1) && stage-2 (300-350#2) && stage-2 (400-450#1) ➔ sleep-eff (mildly-reduced)
   [Confidence = 100%; Support = 8.64%; Lift = 2.5313; Chi-square = 11.7301; P<0.001]
4. stage-1 (850-900#1) && stage-2 600-650#1) && stage-2 (850-900#1) ➔ sleep-eff (normal)
   [Confidence = 100%; Support = 6.17%; Lift = 3.5217; Chi-square = 13.4314; P<0.001]
5. stage-1 (200-250#1) && stage-1 (550-600#1) && stage-2 (200-250#1) ➔ sleep-eff (normal)
   [Confidence = 100%; Support = 6.17%; Lift = 3.5217; Chi-square = 13.4314; P<0.001]

Analysis of Rules:

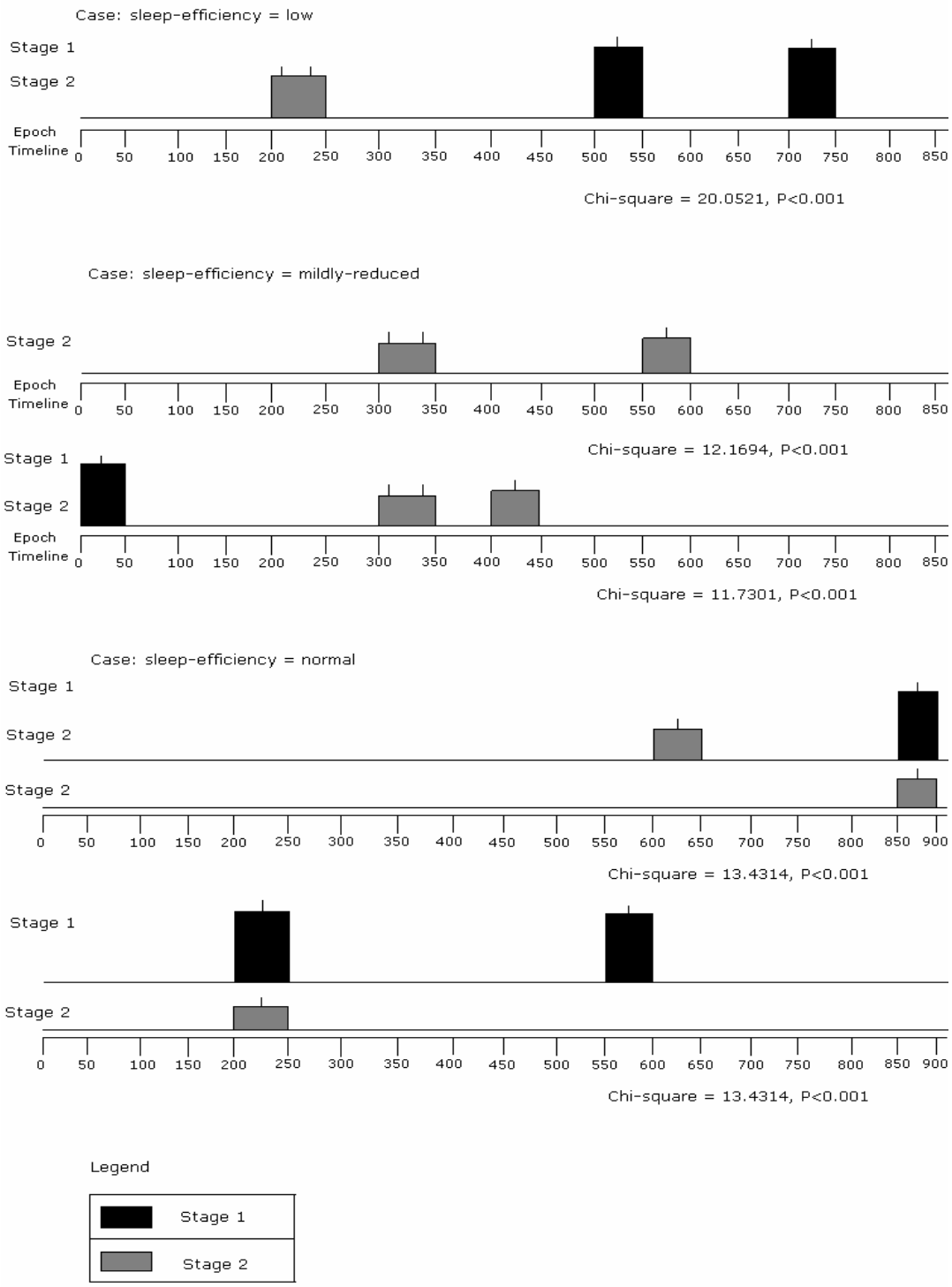


Fig. 7.19. Variations in stages 1, 2 of sleep and OSA influencing sleep efficiency (Experiment 12)

'Case' in the above figure represents the consequent of the rules, while the blocks represent the antecedent attributes in the rules. From the results and the figure above it can be easily deduced that sleep-efficiency is reduced with increase in the fragmentation in a stage within a particular window. It is interesting to note that stage-2 of sleep with two fragments within the 200^{th} to 250^{th} epoch window correlates highly with low sleep efficiency while an identical fragmentation within stage-2 between the 300^{th} and 350^{th} epochs affects the sleep efficiency only mildly. Patients with normal sleep efficiency do not suffer multiple fragmentations of sleep stages within the same window. It is interesting to note that stage-2 of sleep with two fragments within the 200^{th} to 250^{th} epoch window is also noticed in patients with severe cases of OSA (experiment 10). This result indirectly suggests that patients with severe cases of OSA will have low sleep efficiency.

7.3 Difference between Association rule mining and Logistic regression

In this section, we compare the two analysis approaches used in this thesis, viz., association rule mining and logistic regression. While association rule mining focuses on providing descriptive rules of associations between attribute-value pairs, logistic regression is geared towards building a predictive classification model from the attribute-value pool that forms the dataset. Moreover, logistic regression does not consider the correlations that exist among the independent variables or attributes that are used in predicting the dependent variable. Association rules on the other hand, do not have any dependent attribute or variable and they elicit all associations existing between the attribute values that satisfy the thresholds of support and confidence. There is another limitation when it comes to applying logistic regression for exploratory data analysis. It can be modeled to predict only one class. Thus, for instance, we may be able to develop a logistic regression model to predict a certain disorder. However, typically patients suffer from multiple disorders. In such cases, association rule mining can be used to identify multiple disorders that patient may suffer from.

Logistic regression can aid in providing more information on the rules obtained from association rule mining. More often than not, when we have multiple attribute-values participating in a rule, it is difficult to estimate the degree of influence a particular attribute-value has on the consequent that results in a rule. With logistic regression, it is possible to identify the contributing effect of the various attribute-values in a rule. We will point out this feature when we compare results from association rule mining with results obtained from logistic regression experiments.

7.4 Experimental analyses

Data Organization:

Our objective in performing logistic regression is to indicate or prove that association rule mining can help in narrowing down the independent attributes to be considered for predicting a given dependent variable. We use the chi-square test for statistical independence or correlation to identify the attributes that will best fit the classification model. This is one of the most important benefits of association rule mining. Note that the number of factors included in this study is 63 and therefore it is unlikely that one could identify a reasonable number of the key predicting factors by hand. If it were not for association rule mining, we would have no idea of which factors are most important or correlate the most with the dependent attribute under consideration.

Quite often we do not obtain a match between the results obtained using SAS and those that result from the mining process. There are several reasons for this. Firstly, SAS as mentioned earlier builds several models internally and selects the most parsimonious model by balancing the fit with other factors. In this process, some of the attribute-values obtained during association rule mining process that are probably present in one of those models fails to appear in the final result. Also, since the sample set of rules we provided as the result of the mining process are statistically very robust, some of the independent attribute-value variables identified by SAS to correlate with the dependent variable may

not show up in the rules from the association rule mining process if their statistical significance happens to be lower.

Experiment –1:

Objective:

Multinomial logistic regression to predict weight-related problems (both under and over-weight) in patients by selecting epoch windows showing presence of stages 1,2 and REM of sleep as independent variables based on the most statistically significant rules obtained from the association rule mining process.

The following is the list of association rules.

1. stage-2 (50-100#1) && stage-R (350-400#1) ➔ bmi-class (obese) && depression (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 2.8929; Chi-square = 11.1312; P<0.001]
2. stage-2 (700-750#1) && stage-R (450-500#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3738; P<0.001]
3. stage-2 (700-750#1) && stage-R (700-750#1) ➔ bmi-class (obese) && epworth (mild)
   [Confidence = 55.56%; Support = 6.17%; Lift = 3.4615; Chi-square = 11.7223; P<0.001]
4. stage-1 (250-300#1) && stage-2 (300-350#1) && stage-R (350-400#1) ➔ bmi-class (overweight)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.375; Chi-square = 11.9769; P<0.001]
5. stage-1 (250-300#1) && stage-2 (300-350#1) && stage-R (350-400#1) ➔ gender (female) && bmi-class (overweight)
   [Confidence = 83.33%; Support = 6.17%; Lift = 7.5; Chi-square = 34.2049; P<0.001]
6. stage-2 (200-250#2) && stage-2 (350-400#1) && stage-2 (450-500#2) && stage-R (800-850#1) ➔ bmi-class (obese)
   [Confidence = 100%; Support = 6.17%; Lift = 2.1316; Chi-square = 9.0037; 0.01<P<0.001]
7. stage-2 (800-850#1) && stage-R (750-800#1) ➔ bmi-class (normal)
   [Confidence = 35.71%; Support = 6.17%; Lift = 2.4107; Chi-square = 8.7458; 0.01<P<0.001]
8. stage-2 (700-750#1) && stage-R (400450#1) && stage-R (350-400#1) ➔ bmi-class (obese)
   [Confidence = 100%; Support = 6.17%; Lift = 2.1316; Chi-square = 9.0037; 0.01<P<0.001]
9. stage-2 (300-350#1) && stage-R (350-400#1) ➔ bmi-class (overweight)
   [Confidence = 71.43%; Support = 6.17%; Lift = 2.8929; Chi-square = 13.4395; P<0.001]

Output generated by SAS:

Model Information

Response Variable	Body-Mass index (BMI)
Number of Response Levels	5
Number of Observations	81
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	BMI	Total Frequency
1	underweight	8
2	normal	12
3	marginally overweight	3
4	overweight	20
5	obese	38

Explanatory variables:
Selected windows of stage-1, 2 and stage-R (REM) of sleep (Window size = 50 epochs)

Class level information

Class	Value	Design variables
sleep-stage epoch window	no	1
	yes	-1

Forward Selection Procedure adopted

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	st2-700-750#1	1	1	8.56	0.0034
2	st2-200-250#2	1	2	7.6452	0.0057
3	st2-800-850#1	1	3	4.4716	0.0345
4	stR-350-400#1	1	4	4.2518	0.0392

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-3.3382	0.5070	43.3486	<0.0001
Intercept2	1	-2.0469	0.3989	26.3288	<0.0001
Intercept3	1	-1.7966	0.3840	21.8841	<0.0001
Intercept4	1	-0.4328	0.3273	1.7486	0.1861
st2-200-250#2 no vs yes	1	0.6104	0.2518	5.8793	0.0153
st2-700-750#1 no vs yes	1	0.7582	0.2474	9.3908	0.0022
st2-800-850#1 no vs yes	1	-0.4790	0.2220	4.6541	0.0310
stR-350-400#1 no vs yes	1	0.5245	0.2786	3.5455	0.0597

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits		Odds Ratio
st2-200-250#2 no vs yes	3.390	1.264	9.095	1.8411
st2-700-750#1 no vs yes	4.556	1.727	12.015	2.1344
st2-800-850#1 no vs yes	0.384	0.161	0.916	0.6194
stR-350-400#1 no vs yes	2.855	0.958	8.508	1.6896

Association of Predicted Probabilities and Observed Responses

Percent Concordant	69.2	Somers' D	0.483
Percent Discordant	20.9	Gamma	0.536
Percent Tied	9.9	Tau-a	0.335
Pairs	2250	c	0.741

Analyses of Result:

The model above identifies and isolates the 50 epoch-sized windows of stage2 and REM sleep that are most likely to help in predicting obesity based on Body-Mass index values. Here the explanatory variables internally are represented by two values depending on whether the stage of sleep (stage-2 or REM) is witnessed within the concerned epoch window or not. If the stage is witnessed, it is internally represented by –1. While 1 represents the condition that the concerned stage is not witnessed within the window. To illustrate the internal representation of sleep stage windows consider the following example. Consider the epoch window 0-50 and stage-2 as the stage of sleep we are concerned with.

Now,

Presence of stage-2 of sleep within window 0-50 = st-2-0-50#x yes = -1,

Absence of stage-2 of sleep within window 0-50 = st-2-0-50#x no = 1.

The 'x' following the # sign in the above example indicates number of fragmentations that sleep stage-2 undergoes within the epoch window. It will always have a numeric value greater than or equal to 1.

The output under the section 'Analysis of Maximum Likelihood Estimates' provides the values of the regression co-efficients for the explanatory variables selected by forward selection. The negative values of the estimated regression co-efficients for some of the variables (st2-800-850#1) is indicative that decreasing their values (corresponds to presence of sleep stages) will increase the likelihood of patients suffering from excessive weight-associated problems (overweight and obesity problems).

A more intuitive way of interpreting regression co-efficients is exponentiating them to obtain the 'Odds ratio'. For instance, the value 0.6194 indicates that the odds in favor of patients suffering from acute weight problems when the presence of one fragment of stage-2 sleep between the 800^{th} and 850^{th} epoch window is almost 40%.

Experiment-2:

Objective:

Multinomial logistic regression to predict obstructive sleep apnea in patients by selecting epoch windows showing presence of stages 1, 2 and REM of sleep as independent variables based on the most statistically significant rules obtained from the association rule mining process.

The following is the list of association rules.

1. stage-3 (400-450#1) && stage-3 (500-550#1) ➔ OSA (mild)
   [Confidence = 62.5%; Support = 6.17%; Lift = 4.2634; Chi-square = 16.2308; P<0.001]
2. stage-1 (500-550#1) && stage-3 (400-450#1) && stage-3 (500-550#1) ➔ OSA (mild)
   [Confidence = 83.33%; Support = 6.17%; Lift = 4.2634; Chi-square = 16.7577; P<0.001]
3. stage-1 (200-250#2) && stage-1 (750-800#2) && stage-2 (200-250#2) ➔ OSA (mild)

[Confidence = 100%; Support = 6.17%; Lift = 4.2632; Chi-square = 17.3809; P<0.001]

4. stage-2 (150-200#2) && stage-2 (800-850#1) && stage-3 (300-350#1) ➔ OSA (mild)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.5526; Chi-square = 12.9327; P<0.001]

5. stage-2 (200-250#1) && stage-3 (400-450#1) ➔ OSA (mild)

[Confidence = 85.71%; Support = 7.41%; Lift = 3.6541; Chi-square = 16.5468; P<0.001]

6. stage-1 (150-200#1) && stage-2 (50-100#2) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

7. stage-1 (150-200#1) && stage-2 (150-200#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

8. stage-1 (150-200#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

9. stage-2 (150-200#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)

[Confidence = 63.64%; Support = 8.64%; Lift = 2.8636; Chi-square = 12.6274; P<0.001]

10. stage-1 (150-200#1) && stage-2 (150-200#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

11. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-3 (50-100#1) ➔ OSA (moderately-severe)

[Confidence = 63.64%; Support = 8.64%; Lift = 2.8636; Chi-square = 12.6307; P<0.001]

12. stage-1 (400-450#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]

13. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]

14. stage-1 (400-450#1) && stage-1 (700-750#1) && stage-3 (50-100#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

15. stage-1 (400-450#1) && stage-1 (700-750#1) && stage-2 (200-250#2) && stage-3 (50-100#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

16. stage-1 (500-550#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

17. stage-2 (700-750#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

18. stage-2 (700-750#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)

[Confidence = 75%; Support = 7.41%; Lift = 3.0; Chi-square = 11.8402; P<0.001]

19. stage-2 (700-750#1) && stage-3 (50-100#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)

[Confidence = 100%; Support = 6.17%; Lift = 4.5; Chi-square = 18.6422; P<0.001]

20. stage-2 (700-750#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

21. stage-3 (50-100#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.2143; Chi-square = 11.1146; P<0.001]

22. stage-2 (700-750#1) && stage-3 (50-100#1) && stage-3 (150-200#1) && stage-R (650-700#1) ➔ OSA (moderately-severe)

[Confidence = 100%; Support = 6.17%; Lift = 3.75; Chi-square = 14.6474; P<0.001]

23. stage-1 (350-400#1) && stage-1 (700-750#1) ➔ OSA (moderately-severe)

[Confidence = 66.67%; Support = 7.41%; Lift = 3.0;Chi-square = 11.5761; P<0.001]

24. stage-1 (300-350#1) && stage-2 (700-750#1) && stage-3 (500-550#1) ➔ OSA (moderately-severe)

[Confidence = 83.33%; Support = 6.17%; Lift = 3.75; Chi-square = 13.9944; P<0.001]

25. stage-2 (150-200#1) && stage-3 (50-100#1) && stage-3 (150-200#1) ➔ OSA (moderately-severe)

[Confidence = 75%; Support = 7.41%; Lift = 3.375; Chi-square = 14.3113; P<0.001]

26. stage-1 (400-450#1) && stage-2 (200-250#2) && stage-2 (300-350#2) ➔ OSA (moderately-severe)

[Confidence = 100%; Support = 6.17%; Lift = 4.5; Chi-square = 18.6422; P<0.001]

27. stage-1 (750-800#1) && stage-2 (300-350#1) ➔ OSA (moderately-severe)

[Confidence = 70%; Support = 8.64%; Lift = 3.15; Chi-square = 15.0634; P<0.001]

Output generated by SAS:

Model Information

Response Variable	Obstructive Sleep Apnea (OSA)
Number of Response Levels	6
Number of Observations	81
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	OSA	Total Frequency
1	absent	30
2	insignificant	1
3	mild	19
4	mildtomoderate	8
5	moderatelysevere	18
6	severe	5

Explanatory variables:

Selected windows of stage-1, 2 and stage-R (REM) of sleep (Window size = 50 epochs)

Class level information

Class	Value	Design variables
sleep-stage epoch window	no	1
	yes	-1

Forward Selection Procedure adopted

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	st2-300-350#1	1	1	4.4963	0.034

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	-0.4378	0.2376	3.3951	0.0654
Intercept2		1	-0.3825	0.2364	2.6182	0.1056
Intercept3		1	0.6076	0.2422	6.2954	0.0121
Intercept4		1	1.0637	0.2613	16.5727	<0.0001
Intercept5		1	2.8940	0.4717	37.6419	<0.0001
St2-300-350#1	no v/s yes	1	-0.4498	0.2132	4.4511	0.0349

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits		Odds Ratio
st2-300-350#1	no vs yes	0.407	0.176	0.938	0.6378

Analysis of Results:

The result from SAS indicates that there is a 40% probability of patients suffering from moderate to sever OSA on account of presence of stage-2 sleep during the 300th to 350th epoch.

Experiment-3:

Objective:

Multinomial logistic regression to predict depression in patients by selecting epoch windows showing presence of stages 1,2 and REM of sleep as independent variables based on the most statistically significant rules obtained from the association rule mining process.

The following is the list of association rules.

1. stage-2 (500-550#1) && stage-R (200-250#1) ➔ depression (moderate)
   [Confidence = 100%; Support = 7.41%; Lift = 2.7; Chi-square = 11.0153; P<0.001]
2. stage-2 (500-550#1) && stage-R (500-550#1) ➔ depression (moderate)
   [Confidence = 100%; Support = 9.88%; Lift = 2.7; Chi-square = 15.0912; P<0.001]
3. stage-2 (250-300#2) && stage-R (200-250#1) ➔ depression (moderate)
   [Confidence = 100%; Support = 7.41%; Lift = 2.7; Chi-square = 11.0153; P<0.001]
4. stage-2 (50-100#1) && stage-R (400-450#1) ➔ bmi-class (obese) && depression (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 2.8929; Chi-square = 11.1312; P<0.001]
5. stage-2 (50-100#1) && stage-R (350-400#1) ➔ bmi-class (obese) && depression (mild)
   [Confidence = 75%; Support = 7.41%; Lift = 2.8929; Chi-square = 11.1312; P<0.001]
6. stage-2 (700-750#1) && stage-R (450-500#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3738; P<0.001]
7. stage-2 (500-550#1) && stage-R (500-550#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3799; P<0.001]
8. stage-2 (500-550#1) && stage-R (450-500#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 71.43%; Support = 6.17%; Lift = 3.8571; Chi-square = 14.2082; P<0.001]
9. stage-2 (600-650#1) && stage-R (500-550#1) ➔ bmi-class (obese) && depression (moderate)
   [Confidence = 62.5%; Support = 6.17%; Lift = 3.375; Chi-square = 11.3738; P<0.001]
10. stage-1 (200-250#1) ➔ depression (mild)
   [Confidence = 80%; Support = 24.69%; Lift = 1.3224; Chi-square = 8.5976; 0.01<P<0.001]

Output generated by SAS:

Model Information

Response Variable	Depression
Number of Response Levels	3
Number of Observations	81
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Depression	Total Frequency
1	mild	49
2	moderate	30
3	severe	2

Explanatory variables:
Selected windows of stage-1, 2 and stage-R (REM) of sleep (Window size = 50 epochs)

Class level information

Class	Value	Design variables
sleep-stage epoch window	no	1
	yes	-1

Forward Selection Procedure adopted

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	st1-200-250#1	1	1	5.0572	0.0245
2	stR-500-550#1	1	2	5.8515	0.0156

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	0.4038	0.3109	1.6867	0.194
Intercept2		1	3.8966	0.7536	26.7384	<0.0001
st1-200-250#1	no v/s yes	1	-0.7099	0.2971	5.7094	0.0169
stR-500-550#1	no v/s yes	1	0.6725	0.2907	5.3530	0.0207

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits		Odds Ratio
st1-200-250#1	no vs yes	0.242	0.075	0.775	0.4917
stR-500-550#1	no vs yes	3.838	1.228	11.993	1.9591

Analysis of Results:

The SAS results indicate that there is almost a 50% likelihood of moderate to severe depression being observed for patients who experience stage1 of sleep between the 200th and 250th epoch. The REM stage with epoch periods between 500 and 550 is indicative of decreasing levels of depression. The REM stage with the same epoch periods is also obtained for moderate levels of depression as part of experiment 6. The association rule is not able to tell whether the REM stage was responsible for increasing

levels of depression from mild to moderate or decreasing it. This information can be obtained from the logistic regression result above.

Experiment-4:

Multinomial logistic regression to predict collar-size in patients by selecting epoch windows showing presence of stages 1,2 and REM of sleep as independent variables based on the most statistically significant rules obtained from the association rule mining process.

The following is the set of association rules.

1. stage-1 (500-550#1) && stage-2 (550-600#2) && stage-R (750-800#1) ➔ collar-size (16-18 inches)
   [Confidence = 71.43%; Support = 6.17%; Lift = 3.8571; Chi-square = 14.2082; P<0.001]
2. stage-1 (350-400#1) && stage-2 (300-350#1) && stage-R (750-800#1) ➔ collar-size (16-18 inches)
   [Confidence = 100%; Support = 6.17%; Lift = 2.7; Chi-square = 13.5263; P<0.001]
3. stage-1 (550-600#1) && stage-2 (300-350#1) && stage-R (750-800#1) ➔ collar-size (16-18 inches)
   [Confidence = 83.33%; Support = 6.17%; Lift = 2.25; Chi-square = 8.8924; 0.01<P<0.001]
4. stage-1 (350-400#1) && stage-2 (450-500#2) && stage-R (750-800#1) ➔ collar-size (16-18 inches)
   [Confidence = 83.33%; Support = 6.17%; Lift = 2.25; Chi-square = 8.8924; 0.01<P<0.001]

Output generated by SAS:

Model Information

Response Variable	collar-size
Number of Response Levels	5
Number of Observations	81
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	collar-size	Total Frequency
1	zero-to-fourteen inches	39
2	fourteen-sixteen inches	5
3	sixteen-eighteen inches	30
4	eighteen-twenty inches	6
5	twenty-plus	1

Explanatory variables:

Selected windows of stage-1, 2 and stage-R (REM) of sleep (Window size = 50 epochs)

Class level information

Class	Value	Design variables
sleep-stage epoch window	no	1
	yes	-1

Forward Selection Procedure adopted

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	stR-250-300#1	1	1	6.3072	0.012
2	st1-500-550#1	1	2	5.8346	0.0157
3	st1-500-550#1 && stR-250-300#1	1	3	5.8878	0.0152

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Odds Ratio
Intercept	1	-0.6710	0.3137	4.5751	0.0324	0.511
Intercept2	1	-0.3590	0.3073	1.3647	0.2427	0.6984
Intercept3	1	2.2381	0.4362	26.3228	<0.0001	9.3755
Intercept4	1	4.3327	1.0334	17.5782	<0.0001	76.1496
st1-500-550#1no v/s yes	1	0.1225	0.2793	0.1925	0.6609	1.1303
stR-250-300#1no v/s yes	1	0.5682	0.2841	4.0000	0.0455	1.7651
st1-500-550#1 && stR-250-300#1 no v/s yes	1	0.7228	0.2885	6.2757	0.0122	2.060

Analysis of Results:

This result from SAS indicates that presence of stage-1 during the 500^{th} to 550^{th} epoch and REM during the 250^{th} to 300^{th} epoch is suggestive of the person suffering from less weight problems.

Experiment-5:

Multinomial logistic regression to predict caffeine-intake in patients by selecting epoch windows showing presence of stages 1,2 and REM of sleep as independent variables based on the most statistically significant rules obtained from the association rule mining process.

The following is the set of association rules.

1. stage-1 (300-350#1) && stage-2 (600-650#2) && stage-R (650-700#1) ➔ caffeine-intake(0cups/day)
   [Confidence = 100%; Support = 6.17%; Lift = 6.2308; Chi-square = 27.861; P<0.001]
2. stage-1 (600-650#1) && stage-2 (600-650#2) && stage-R (650-700#1) ➔ caffeine-intake (0cups/day)
   [Confidence = 83.33%; Support = 6.17%; Lift = 5.1923; Chi-square = 21.761; P<0.001]
3. stage-1 (400450#1) && stage-2 (600-650#2) && stage-R (650-700#1) ➔ caffeine-intake (0 cups/day)
   [Confidence = 83.33%; Support = 6.17%; Lift = 5.1923; Chi-square = 21.761; P<0.001]
4. stage-1 (150-200#1) && stage-2 (100-150#1) && stage-R (750-800#1) ➔ caffeine-intake (1-3 cups/day)
   [Confidence = 83.33%; Support = 6.17%; Lift = 3.8571; Chi-square = 14.5696; P<0.001]
5. stage-1 (50-100#1) && stage-2 (500-550#1) && stage-R (750-800#1) ➔ bmi-class (obese) && caffeine-intake (1-3 cups/day)
   [Confidence = 100%; Support = 6.17%; Lift = 3.24; Chi-square = 11.931; P<0.001]

Output generated by SAS:

Model Information

Response Variable	caffeine
Number of Response Levels	5
Number of Observations	81
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	caffeine-intake (cups/day)	Total Frequency
1	absent	13
2	occasional	1
3	one-three	43
4	three-six	20
5	six-plus	4

Explanatory variables:

Selected windows of stage-1, 2 and stage-R (REM) of sleep (Window size = 50 epochs)

Class level information

Class	Value	Design variables
sleep-stage epoch window	no	1
	yes	-1

Forward Selection Procedure adopted

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	stR-650-700#1	1	1	7.4705	0.0063
2	st2-600-650#2	1	2	4.9881	0.0255
3	st1-0-50#1	1	3	8.4837	0.0036
4	st1-400-450#1	1	4	8.1733	0.0043
5	st2-650-700#1	1	5	5.1463	0.0233

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	-1.7239	0.4157	17.2	<0.0001
Intercept2		1	-1.5914	0.4061	15.3572	<0.0001
Intercept3		1	1.7868	0.3998	19.9758	<0.0001
Intercept4		1	4.2887	0.6458	44.0958	<0.0001
st1-0-50#1	no	1	1.0775	0.3116	11.9564	0.0005
st1-400-450#1	no	1	0.8497	0.2623	10.4961	0.0012
st2-600-650#2	no	1	-1.3510	0.3401	15.7773	<0.0001
st2-650-700#1	no	1	0.5758	0.2503	5.2910	0.0214
stR-650-700#1	no	1	-1.1525	0.3237	12.6801	0.0004

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits		Odds Ratio
st1-0-50#1	no vs yes	8.628	2.543	29.267	2.9373
st1-400-450#1	no vs yes	5.471	1.957	15.296	2.3389
st2-600-650#2	no vs yes	0.067	0.018	0.254	0.2589
st2-650-700#1	no vs yes	3.163	1.186	8.439	1.7786
stR-650-700#1	no vs yes	0.1	0.028	0.355	0.3158

Association of Predicted Probabilities and Observed Responses

Percent Concordant	73.3	Somers' D	0.578
Percent Discordant	15.5	Gamma	0.652
Percent Tied	11.2	Tau-a	0.368
Pairs	2063	c	0.789

Analysis of Results:

This is a very interesting result. There is almost a 75% likelihood that presence of stage-2 sleep during the 600^{th} to 650^{th} epoch with 2 fragmentations correlates with increased intake of caffeine. Similarly presence of stage-REM during the 650^{th} to 700^{th} epoch as also stage-1 between 400^{th} and 450^{th} epoch is an indicator of reduced intake of caffeine. It is useful to note that all these three epoch windows with the respective stages were obtained from experiment-7. This experiment reveals that it is presence of stage-1 between 400^{th} and 450^{th} epochs and REM between 650^{th} and 700^{th} epochs that actually correlate with low caffeine intake. Although, stage-2 with two fragmentations is present between the 600^{th} and 650^{th} epochs in rules obtained in association rules shown below for people with no caffeine intake, the logistic regression experiment performed above indicates that presence of stage-2 between the above mentioned epochs is in fact and indicator of increased consumption of caffeine. Thus, this experiment demonstrates how logistic regression results could enhance association rules.

7.5 Summary

The most interesting results are obtained with the mixed-level dataset. This is due to the fact that it combines time-sequence and ordinary attribute-values and tries to elicit associations among them. For instance, we obtain several rules illustrating correlations between certain patterns of stage-2 and REM sleep with demographic factors such as gender, epworth score, weight-category etc. We also obtain rules where sleep patterns associate with the severity of a sleep disorder for e.g., periodic leg movements. There are also a several rules which indicate how sleep pattern varies sharply in patients suffering mild obstructive sleep apnea when compared with more severe cases. There are also rules that identify sleep patterns in patients with variable degrees of sleep efficiency.

We apply logistic regression to predict some of the demographic factors as well as disorders based on the most significant association rules obtained by experimenting over the mixed level dataset. This methodology helps us to obtain better fitting models with

less number of independent variables. Also, selection of independent variables on the basis of association rules minimizes the size of the dataset. Thus, we gain an advantage in factor identification due to the descriptive nature of association rules. Logistic regression in turn helps to determine the significance of the different attributes that form part of association rules.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This thesis work explored and extracted intricate relationships among several factors relating to human sleep by using association rule mining technique. The complexity of the domain and the dataset allowed us to manipulate and organize data in different formats and groupings to do several interesting experiments. We demonstrated by way of experimental results how extended versions of association rule mining system could be used to obtain rules with more information. Our technique of mining rules using the window-based approach of association rule mining helps in obtaining rules with real-time information for easy analysis. Another advantage of this technique is that correlations produced do not provide details of merely the events but also the behavior of the events noticed within the different windows. This may have considerable importance in diagnosing pathology.

The descriptive rules adhering to the imposed constraints compare favorably with the results obtained from logistic regression experiments thereby validating our system. Along the way, while comparing results from association rule mining with those obtained from logistic regression experiments, we were able to detect that results obtained from either method could have a positive effect on the other. The positive effect could be compared with a symbiotic relationship that benefits the two objects in focus. While association rules can easily obtain associations between the participating attribute-values and provide a rule, logistic regression cannot automatically conjure up associations hidden within the dataset. On the other hand, logistic regression can provide insight into the extent of influence the attribute-values have on a predicted variable. In association rule mining where the predicted variable of logistic regression is the constraint forced on the consequent of a rule, the significance or influence of the attribute-values appearing in

the antecedent of a rule can be found using the results of the logistic regression experiment. Thus, the results can be of mutual benefit since they add to the information provided by any one technique when used in isolation.

Another advantage of window-based association rule mining system over logistic regression techniques is appearance of artifacts in the resulting rule is highly unlikely.

8.2 Future Work

This thesis concentrated principally on finding the behavioral, clinical or demographic factors that correlate with sleep progression with pathological and physiological conditions. However, we were not able to obtain data from patients suffering from many pathological disorders. In the future, we need to collect data from patients who have a history of having suffered or suffering from diabetes, hypertension, dementia, stroke etc. It would then be interesting to see if there are sleep pattern variations noticed following one of these diseases. It would also be interesting to compare the sleep pattern variations for patients suffering from say diabetes and hypertension.

We could work towards further extending the window-based approach for mining association rules by including tolerance bands for deciding whether events of interest should be considered as trivial or significant based on the degree of fragmentation and the window size selected.

Work can also be done in the area of intelligent information retrieval from the fuzzy responses to questionnaires. We could apply feature identification and explore methods to reduce the subjectivity of the interpretation.

Mining over raw unstaged sleep signals and developing a classification model for automatic sleep-stage classification is also an area to look at.

If results from these experiments prove to be successful, then their impact on health-care industry could be tremendous. We could build a tool that will find immediate application in the field of patient monitoring systems. Based on the experimental results, we would develop models that bear templates most frequently noticed in patients with particular pathological and physiological conditions. Patients undergoing a sleep study would then have their pattern matched against the reference template. This technique would considerably reduce the time of diagnosis.

We could extend the idea to pre-emptive healthcare. This is a powerful concept because till date medical treatment is not pre-emptive. Here, we could have normal people (suffering from no conscious ailment or disorder) undergo a sleep test. Their pattern can be matched with the several reference templates/patterns associated with different pathological disorders to determine their likelihood of suffering from it in the near future. Remedial action could then be taken by the doctors to thwart the appearance of the pathology.

Bibliography

[Alv03] Sergio A. Alvarez. Chi-Squared Computation for Association Rules: Preliminary Results. Technical Report BC-CS-2003-01, Computer Science Department, Boston College, July 2003.

[Aka74] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723, 1974.

[AK53] Aserinsky E, Kleitman N. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science*, 118:273-274, 1953.

[AS94] Rakesh Agrawal, Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. *In Proc. 20th International Conference on Very Large Data Bases, VLDB*, 487-499, 1994.

[But96] Nic Butkov. *Atlas of Clinical Polysomnography*. Medford, OR: Synapse Media, Inc., 1996, 2 volumes.

[BP01] Blake C, Pratt, W. Better Rules, Fewer Features: A Semantic Approach to Selecting Features From Text. *In Proceedings of the IEEE International Conference on Data Mining*, ICDM, San Jose, CA, 2001.

[BS90] Birbaumer N., Schmidt R.F. *Bilogische Psychologie*, Springer, 1990.

[CR94] Carskadon, Mary A., Rechtschaffen Allan. Monitoring and Staging Human Sleep. In: Kryger, Meir, Roth, Thomas and Dement. *Principles and Practice of Sleep Medicine*. 2nd edition, USA: W.B. Saunders Company, 1994.

[ED96] Everitt, B.S. and Der, G. *A Handbook of Statistical Analyses using SAS*. 1st edition, Chapman & Hall, 1996.

[FW00] E. Frank and I.H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers. 2000.

[HK00] Jiawei Han, Micheline Kamber. *Data Mining: Concepts and Techniques*, The Morgan Kaufmann series in Data Management Systems. Morgan Kaufmann Publishers. August 2000.

[Jas58] Jasper H.H. The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, 20, 371-375, 1958.

[LHM99] Bing Liu, Wynne Hsu and Yiming Ma. Pruning and Summarizing the Discovered

Associations; *In Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD)*, 125-134, 1999.

[NH99] David Nettleton, Lourdes Hernandez: Questionnaire screening of sleep apnea cases using fuzzy knowledge representation and intelligent aggregation techniques; *In Proceedings of the Workshop: Intelligent Data Analysis in Medicine and Pharmacology, IDAMAP*, 91-102, 1999.

[OOBS+01] Carlos Ordonez, Edward Omiecinski, Levien de Braal, Cesar A. Santana, Norberto Ezquerra, Jose A. Taboada, David Cooke, Elizabeth Krawczynska, Ernest V. Garcia. Mining Constrained Association Rules to Predict Heart Disease, *IEEE International Conf. on Data Mining*, *ICDM*, pages 433--440,2001.

[Pra02] Keith A. Pray. Mining Association rules from Time-sequence attributes. Master's thesis (in progress), Department of Computer Science, Worcester Polytechnic Institute

[Quin93] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.1993.

[Riss83] J. Rissamen. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*. 11(2). 416-431.1983.

[RK68] Rechtschaffen A. & Kales A. *A Manual of Standardized Terminology, Techniques and Scoring System for sleep stages of human subjects*. U.S. Dept. of Health, Education and Welfare, National Institute of Health, Publ. No. 204, Washington, 1968.

[Sho01] C.A. Shoemaker. Mining Association rules from set-valued data. Master's thesis, Department of Computer Science, Worcester Polytechnic Institute, May 2001.

[Sto02] Zachary Stoecker-Sylvia: A robust association rule mining tool for complex data. Undergraduate graduation project (MQP), Worcester Polytechnic Institute, April 2002.

Appendix-A

Sample Questionnaire Form

(A) Demographic information

1 Gender ☐ Male ☐ Female

2 Height _____ inches

3 Weight _____ pounds

4 collar-size _____ inches

…..

(B) Epworth score

(3- strongly likely, 2- often, 1- rarely, 0 –never)

1 Do you fall asleep often while driving?

☐	☐	☐	☐
3	2	1	0

1. Do you fall asleep while talking with people?

☐	☐	☐	☐
3	2	1	0

3 Do you fall asleep during meetings or while watching television?

☐	☐	☐	☐
3	2	1	0

…

(C) Habits

1 Do you worry in bed? ☐ Yes ☐ No

2 Do you eat in bed? ☐ Yes ☐ No

…