Dean&Francis

ISSN 2959-6157

# Association Rule-Based Analysis of Contributing Factors to Sleep Disorders

**Yifei Xu**

School of Economics and Management, Shanghai Maritime University, Shanghai, 201306, China
Corresponding author: xyf.azzzzz@gamil.com

**Abstract:**

This study employs the Apriori algorithm in association rule analysis to investigate influencing factors of sleep disorders, with the ultimate goal of providing actionable sleep recommendations. The methodology consists of three key phases: first, dataset preprocessing involving discretization of continuous variables. Second, chi-square testing was used to select five significant variables as independent factors. Finally, association rule mining using the Apriori algorithm to extract meaningful patterns. The findings reveal three key patterns: first, younger populations show significantly elevated insomnia risk with 0.471 support and 0.83 confidence, and this risk persists even with sufficient sleep duration, as evidenced by 0.906 confidence. Second, middle-aged women demonstrate greater susceptibility to sleep apnea, supported by a combined rule confidence of 0.938. Third, younger males exhibit better sleep health outcomes with 0.607 support, while low stress levels show a strong positive association with normal sleep patterns, yielding a 1.205 lift. Additionally, the observed association between moderately high physical activity levels and insomnia, though showing 0.946 confidence, may be influenced by confounding factors. Consequently, insomnia management in younger populations should focus on stress reduction and lifestyle modifications, while middle-aged women with sleep apnea may benefit from therapies such as non-invasive positive pressure ventilation. For the general population, public education initiatives should be implemented to enhance overall sleep health literacy.

**Keywords:** association rule; Apriori; sleep disorder.

## 1. Introduction

Sleep, as a vital physiological phenomenon, plays a crucial role in maintaining life activities and physical health. Sleep disorders have been associated with a large number of harmful health consequences, including raising the risk of diabetes, obesity, stroke, and many other diseases, so it is essential to under-

stand various factors affecting sleep quality.

Research has identified multiple influencing factors across different populations. Liu found physical discomfort to be the primary factor impairing sleep quality in hospitalized patients, followed by psychological stress and environmental changes [1]. Further studies by Huang and Shen categorized these influences into three factors: iatrogenic factors, such as the discomfort caused by diseases; environmental factors, such as uncomfortable hospital beds; and psychological factors, such as patients' worry about diseases [2]. Among respiratory patients, Lu identified cough, chest pain, financial burden, and dyspnea as major contributors to sleep disorders through multivariate logistic regression analysis [3].

In non-clinical populations, Meng employed a series of assessment tools, including the Pittsburgh Sleep Quality Index (PSQI), to investigate sleep quality among military personnel stationed in high-altitude regions. The study revealed that psychological stress levels and fatigue severity significantly impacted sleep quality, while the soldiers' confidence in training exercises and self-assessed health status exerted positive effects on sleep outcomes [4]. By using an online survey of PSQI, IAT, and PHQ-9, the rate of depressive symptom level, internet addiction, and poor sleep quality showed strong correlations among them were concluded [5]. Thalia reported that moderate physical activity indirectly improved sleep through enhanced emotion regulation [6]. However, Srirangaramasamy found no significant association (p=0.659) between physical activity and sleep quality, despite a high prevalence (74.9%) of poor sleep among young adults, suggesting complex and sometimes contradictory relationships between lifestyle factors and sleep outcomes [7].

Therefore, current researches focus on specific populations, using PSQI and other tools to assess sleep quality and multifactorial logistic regression or hypothesis testing to analyze influencing factors. The public may use PSQI to evaluate sleep patterns, but lacks insight into behaviors contributing to sleep disorders. Moreover, logistic regression, the predominant method employed in existing studies, has limited capability in uncovering potential behavioral patterns. To address these gaps, this study employs association rule mining to systematically analyze 13 relevant factors through multiple association rule extractions. This approach aims to identify key factors affecting sleep quality and provide actionable sleep improvement recommendations for the general population.

# 2. Methods

## 2.1 Data Source and Introduction

The Sleep Health and Lifestyle Dataset was obtained from Kaggle, comprising over 350 rows and 13 columns [8]. The detailed variations of this dataset include gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress level, BMI category, blood pressure, heart rate, daily steps, and sleep disorder. While gender, occupation, BMI category, and sleep disorder are categorical types, the remaining variables are all of the numeric type.

## 2.2 Chi-Square Test

The chi-square ($\chi^2$) test measures the discrepancy between observed sample frequencies and theoretically expected values under a null hypothesis. The process involves formulating the null hypothesis H0, dividing data into k intervals with ≥5 expected observations each, and calculating a certain test statistic. It follows a $\chi^2$ distribution with (k-1) degrees of freedom.

Its advantages lie in broad applicability to categorical and discrete data, straightforward interpretation, and versatility for goodness-of-fit tests or independence analysis. The method is also computationally simple and requires minimal assumptions, making it a robust tool in statistical inference.

## 2.3 Association Rules Mining(ARM)

Association Rule Mining is an unsupervised data mining method designed to discover meaningful relationships among large-scale data items. An association rule is expressed in the form $A \to B$, indicating that A leads to B, where A is the antecedent, B is the consequent, and $A \cap B = \varnothing$.

Three critical metrics are used to evaluate association rules: support, confidence, and lift. The support of an itemset X, denoted as $support(X)$, is defined as the proportion of transactions in dataset D that contain all items in X.

The confidence of $A \to B$ measures the likelihood that B occurs given A, as the calculation formula are shown in Equation (1).

$$confidence(A \to B) = \frac{support(A \cup B)}{support(A)} \tag{1}$$

The lift of $A \to B$ quantifies the correlation between A and B, calculated as:

$$lift(A \to B) = \frac{confidence(A \to B)}{support(B)} \tag{2}$$

ISSN 2959-6157

Rules meeting minimum support and confidence thresholds are strong association rules. Those with $lift(A \rightarrow B) > 1$ are considered meaningful.

## 2.4 Apriori

In 1994, R. Agrawal proposed the Apriori algorithm based on association rules, which remains one of the most fundamental algorithms in association rule mining until these days. The Apriori algorithm primarily involves two key steps, which are identifying all frequent itemsets (itemsets with $support(A \rightarrow B) \geq min\_sup$), and generating strong association rules from these frequent itemsets.

Using $L_k$ to denote the set of all frequent k-itemsets. The Apriori algorithm employs an iterative approach to discover frequent itemsets. First, it identifies C1 (candidate 1-itemsets) as all unique items, and $L_1$ (frequent 1-itemsets). Then, it uses $L_{k-1}$ to derive $L_k$ until no further frequent itemsets can be found.

The step-by-step process is shown in Fig. 1.

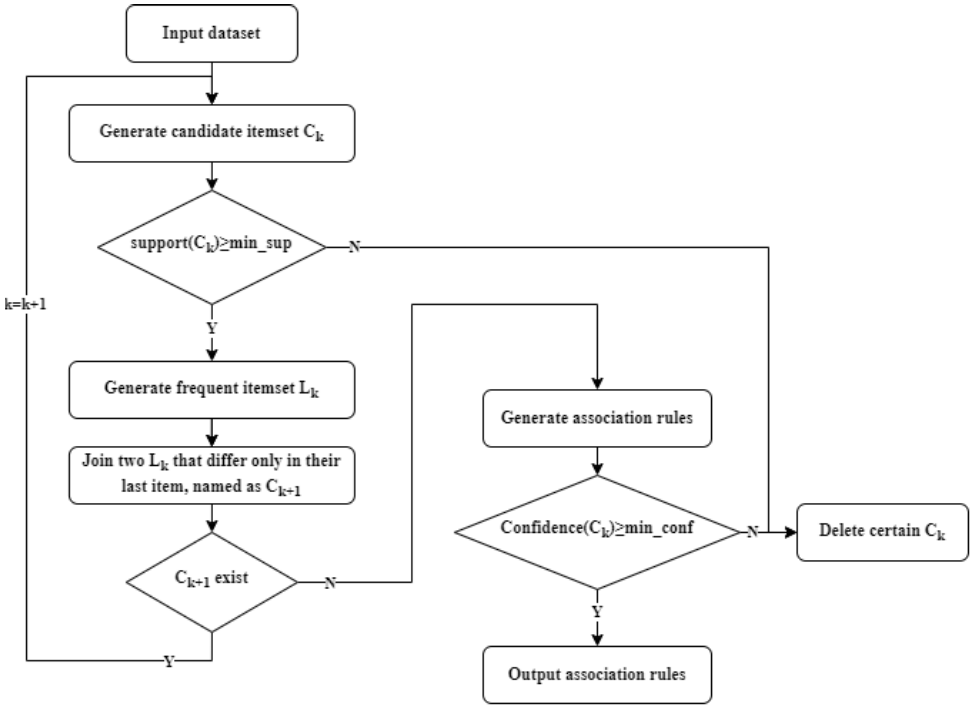


**Fig. 1 The flow chart of the Apriori algorithm (original)**

## 2.5 Data Preprocessing

As discrete variables are often needed in Association Rules Mining, known as the ARM, this paper discretizes some variables by the following methods.

According to the World Health Organization, they defined the young age as 25 to 44 and the middle age as 45 to 59. Experts from the National Heart, Lung, and Blood Institute recommend that adults sleep between 7 and 9 hours a night, so the sleep duration is divided into adequate and inadequate. As the numbers included in the Quality of Sleep are unevenly distributed, they are reclassified into 4 categories using clustering algorithms. The figures in Physical Activity Level are processed in the same way. What's more, the numbers in the Stress Level are evenly divided into 6 groups because they are approximately equidistributed. With merely 0.5% of cases showing abnormal blood pressure and resting heart rate values, Blood Pressure and Heart Rate are excluded due to insufficient sample size for reliable results. Daily step counts for adults were dichotomized as adequate, which refers to 7,000 to 10,000 steps, or inadequate, which means fewer than 7,000 steps, according to Paluch AE et al [9].

After data preprocessing, the changes of some variables are shown in Table 1.

**Table 1. New Variable**

| Original Variable Name | New Variable Name | Range |
|---|---|---|
| Age | $Age_{cl}$ | Young/Middle age |
| Sleep Duration | $SD_{cl}$ | Adequate/Inadequate sleep |
| Quality of Sleep | $QS_{cl}$ | QS4~5, QS6, QS7, QS8, QS9 |
| Physical Activity Level | $PAL_{cl}$ | PAL1, PAL2 ... |
| Stress Level | $SL_{cl}$ | High/Medium/Low Stress |
| Daily Steps | $DS_{cl}$ | Adequate/Inadequate Daily Steps |

# 3. Results

## 3.1 Chi-Square Test

After testing the chi-square test for 45 pairs of variable groups, this paper only shows a total of five variable groups where the two variables are uncorrelated, for the null hypothesis is rejected (Table 2).

**Table 2. Chi-square test**

| Variable 1 | Variable 2 | p-value |
|---|---|---|
| $Age_{cl}$ | $SD_{cl}$ | 0.980 |
| Gender | $SD_{cl}$ | 0.862 |
| Gender | $PAL_{cl}$ | 0.825 |
| Gender | $SL_{cl}$ | 0.240 |
| $SD_{cl}$ | $SL_{cl}$ | 0.051 |

The p-values of all five variable pairs exceed 0.05; therefore, they fail to reject the null hypothesis that these variables are not independent. Although the fifth pair‘s p-value is slightly above 0.05, suggesting a potential weak association that doesn‘t reach statistical significance, it is still included in subsequent ARM since both variables appear in the first four independent pairs. Thus, these five variable relationships are selected for ARM, which effectively reduces multicollinearity in rule generation, focuses computational resources on the most informative variable relationships, and ensures the derived association rules reflect independent effects rather than spurious correlations.

## 3.2 Chi-Square Test

Association rule mining requires discrete data, as raw data with scattered and continuous distributions is difficult to analyze directly for patterns. After discretization, the data is aggregated into clear categories, meeting algorithm requirements while simplifying the structure and reducing noise. This enables the mined association rules to be more concise and interpretable, improving analysis efficiency and the usability of results.

In order to go on the subsequent association rule mining analysis, the original continuous variables need to be discretized. Fig. 2 presents a comparative distribution of raw data in blue and discretized data in orange across four numerical dimensions: age, step count, physical activity level, and stress level.

The comparative visualization demonstrates distinct distribution patterns before and after discretization across four variables: age transitions from a dispersed 30~60-year continuum to a dominant young vs. middle-aged dichotomy. Sleep duration condenses its 6~8-hour continuum into a binary classification favoring sufficient sleep. Physical activity level collapses its scattered 30–90 range into six tiers with one sparse outlier category. Stress level shifts from a bimodal 3–8 distribution into a tripartite structure where low stress predominates, collectively transforming continuous variability into interpretable discrete categories for association rule mining.

ISSN 2959-6157

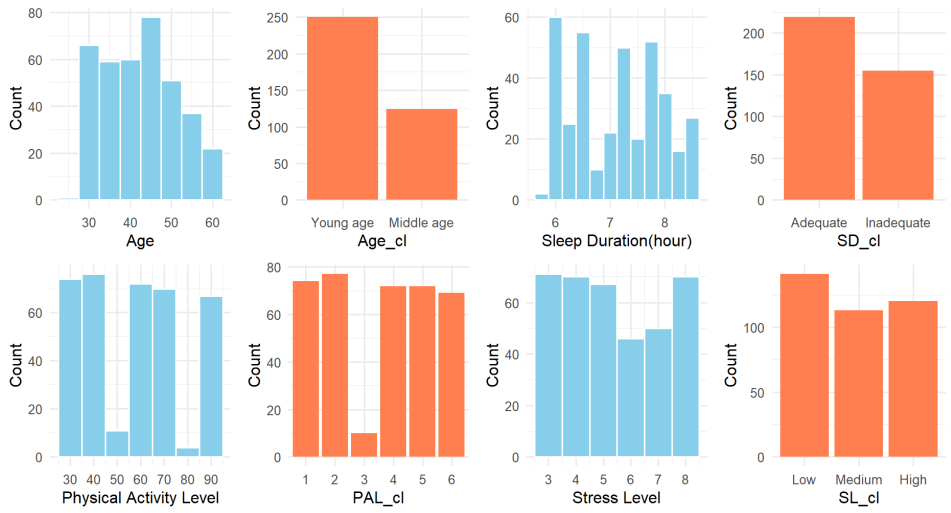


**Fig. 2 The histogram of variables (original)**

Fig. 3 presents the gender-specific distribution of sleep disorders, revealing distinct patterns between males and females. The bar chart uses orange to represent Insomnia, blue for Normal sleep, and green for Sleep Apnea, showing their distribution counts between 0-200 across genders. It shows that a higher proportion of males maintain normal sleep status, while females exhibit a greater prevalence of sleep disorders. Both genders have similar numbers of insomnia cases, but the frequency of insomnia is significantly higher in women than in men.

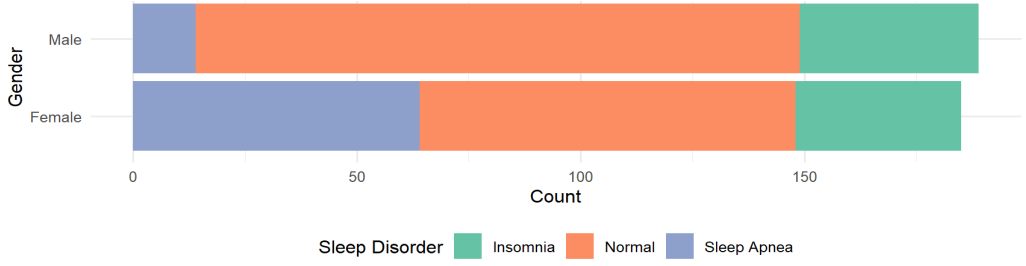


**Fig. 3 The stacked bar chart of Sleep disorder and Gender (original)**

### 3.3 Association Rules Mining Results

According to the ARM method, based on the preprocessed data, with the minimum support set to 0.3 and the minimum confidence set to 0.4, a total of 70 association rules are found. After further filtering, the association rules revealing the influencing factors of sleep disorder are screened out and arranged in descending order of support, which are shown in Tables 3 and 4, respectively. Table 3 lists the top 10 association rules involving sleep disorders with the highest support metric, while Table 4 presents the top 5 highest-support rules with normal sleep.

Table 3. Rules including sleep disorders

| Rules | Support Metric | Confidence Metric | Lift Metric |
|---|---|---|---|
| Young age→Insomnia | 0.471 | 0.830 | 1.670 |
| Female→Sleep Apnea | 0.413 | 0.634 | 1.259 |
| Middle age→Sleep Apnea | 0.406 | 0.941 | 1.869 |
| Adequate sleep→Insomnia | 0.400 | 0.614 | 1.236 |
| Female, Middle age→Sleep Apnea | 0.394 | 0.938 | 1.865 |
| Adequate sleep, Young age→Insomnia | 0.374 | 0.906 | 1.824 |
| PAL4→Insomnia | 0.342 | 0.946 | 1.905 |
| PAL4, Young age→Insomnia | 0.342 | 0.946 | 1.905 |
| PAL4, Adequate sleep→Insomnia | 0.335 | 0.963 | 1.938 |
| PAL4, Adequate sleep, Young age→Insomnia | 0.335 | 0.963 | 1.938 |

A relatively high support value suggests that this association occurs frequently in the data. According to Table 4, the support metric of Young age → Insomnia is 0.471, indicating that 47.1% of the records in the dataset contain both young age and insomnia. Similarly, the support metric values for Female → Sleep Apnea and Middle age → Sleep Apnea both exceed 0.4, implying that young individuals are more likely to suffer from insomnia, while middle-aged individuals or females are more susceptible to sleep apnea.

A high confidence metric suggests that a certain rule has stronger predictive results than a low one. The confidence of PAL4 → Insomnia is 0.946, indicating that 94.6% of individuals with PAL4 (Physical Activity Level = 4, representing a moderately high activity level in this dataset) suffer from insomnia. Similarly, the confidence metric of Adequate sleep, Young age → Insomnia is 0.906, implying that even among young individuals with sufficient sleep, the probability of insomnia remains high. This could be attributed to stronger pre-sleep excitement in younger populations than others, which may lead to difficulty falling asleep and, consequently, insomnia.

A lift metric greater than 1 indicates a positive correlation between the antecedent and the consequent. The lift metric of PAL4 → Insomnia is 1.905, suggesting that a moderately high activity level, which is particularly PAL4, has a significant influence on insomnia.

**Table 4. Rules excluding sleep disorders**

| Rules | Support Metric | Confidence Metric | Lift Metric |
|---|---|---|---|
| Young age, Male→Normal | 0.607 | 0.821 | 1.332 |
| Inadequate sleep→Normal, Young age | 0.361 | 0.782 | 1.057 |
| Male, Adequate sleep→Normal | 0.352 | 0.570 | 1.059 |
| Young age, Adequate sleep→Normal, Male | 0.342 | 0.904 | 1.466 |
| Low Stress Level→Normal, Young age | 0.338 | 0.892 | 1.205 |

Among the four rules, the first rule shows a significantly higher support figure than the others, indicating that young age, male, and no sleep disorder frequently co-occur. Additionally, the confidence and lift values of this rule are relatively high, supporting the reasonable inference that young males generally do not suffer from sleep disorders.

The second rule shows moderately high support and confidence, but only an average lift value, suggesting it may serve as a potentially valid rule, though not strongly predictive. Specifically, it implies that insufficient sleep tends to correlate with the absence of sleep disorders in young individuals. A possible explanation is that sleep deprivation leads to physical exhaustion, promoting deeper sleep as the body attempts to recover, thereby reducing the likelihood of insomnia or other sleep disturbances.

However, when high support, low confidence, and low lift occur simultaneously, the association rule is likely to be spurious, merely reflecting the frequent co-occurrence of events in the dataset. The third rule raises the possibility of a spurious association rather than a meaningful causal relationship.

The fourth and fifth rules demonstrate moderate support figures but notably high confidence and lift figures, strongly suggesting that when these conditions are met, the association is highly reliable. Specifically, young individuals with adequate sleep and low stress levels are very likely to be free from sleep disorders.

Additionally, the analysis reveals that while PAL4 appears frequently in insomnia-related association rules, no significant rules are found linking other activity levels to sleep disorders. This pattern suggests that the apparent association of PAL4 may stem from its frequent co-occurrence with youth and adequate sleep duration rather than direct causation. The absence of consistent activity-sleep disorder relationships across all PAL categories implies that physical activity level, when isolated from confounding variables, demonstrates no meaningful impact on sleep disorder prevalence.

## 4. Discussion

For younger populations, insomnia should not be simply attributed to insufficient sleep duration, but may require stress reduction interventions or more comprehensive analysis of lifestyle factors to identify targeted mitigation strategies [7]. The preceding analysis indicates that middle-aged women are more susceptible to sleep apnea. Notably, Researcher Tang‘s findings highlight a significant increase in the incidence of obstructive sleep apnea among women during pregnancy and postmenopause. For clinical management, non-invasive positive pressure ventilation (NPPV) and hormone replacement therapy (HRT) represent viable therapeutic options for this condition [10]. Meanwhile, for other demographic groups, regular

ISSN 2959-6157

public education campaigns should be implemented to enhance sleep health literacy, complemented by community medical services and health seminars to improve population-wide understanding of sleep health management. Furthermore, the research outcomes can be directly applied to public health guidance, such as designing behavioral intervention programs for high-stress young adults or raising public awareness of sleep disorder prevention through health education.

This study not only uncovers population-specific risk patterns of sleep disorders but also innovatively translates data mining findings into tiered intervention strategies: designing stress-reduction programs for high-risk young adults, promoting sleep apnea prevention and treatment measures for middle-aged women, and enhancing public sleep health literacy through community health networks. Although the findings hold significant public health value, the conclusions still require further validation through multicenter studies to strengthen their generalizability and clinical applicability.

It should be noted that while this study presents relevant findings, the conclusions should be interpreted as preliminary observations requiring further validation through additional datasets.

## 5. Conclusion

This study employed the Apriori algorithm for comprehensive association rule mining to systematically examine sleep disorder patterns across different demographic groups. The analysis yielded several clinically significant findings, most notably the strong association between younger age groups and insomnia prevalence, demonstrated by high support and confidence metrics. This relationship persisted even when accounting for adequate sleep duration, suggesting other underlying contributing factors. The research also identified middle-aged women as a particularly vulnerable population for sleep apnea, with rule-based analysis showing exceptionally high confidence levels. Younger male participants conversely displayed the most favorable sleep health outcomes among all studied groups, with robust statistical support. These core findings were complemented by additional insights into stress-sleep relationships and physical activity impacts.

From a translational perspective, this study not only uncovered population-specific sleep disorder risk patterns but also innovatively transformed data mining results into targeted interventions. Specifically, it informed stress reduction programs for high-risk youth and optimized treatment measures for sleep apnea in middle-aged women, while simultaneously improving public sleep health literacy through community health networks.

Nevertheless, certain limitations persist. Despite their significant public health value, these findings require further validation through multicenter studies to strengthen generalizability and clinical utility.

## References

[1] Liu F. Investigation and influencing factors analysis of sleep quality in internal medicine inpatients. China Health Industry, 2013, 10(6), 111.

[2] Huang H., Shen A. Investigation and influencing factors analysis of nocturnal sleep quality in gynecological inpatients. Chinese General Practice Nursing, 2016, 14(35), 3756-3757.

[3] Lu L. Investigation and influencing factors analysis of sleep quality in respiratory inpatients. Electronic Journal of Practical Clinical Nursing Science, 2020, 5(12), 81-82.

[4] Meng X., Zhang Q., Wang J., et al. Investigation and influencing factors analysis of sleep quality in plateau stationed soldiers. Chongqing Medicine, 2017, 46(25), 3571-3573.

[5] Chaleshi D, Badrabadi F, Anari F G, et al. Depressive Symptom Level, Sleep Quality, and Internet Addiction among Medical Students in Home Quarantine during the COVID-19 Pandemic. Mental Illness, 2023, Article 1787947.

[6] Thalia S. T. Long-term links between physical activity and sleep quality. Medicine and science in sports and exercise, 2018.

[7] Srirangaramasamy J, Karthikeyan V, Ramanathan R, et al. Assessment of Physical Activity and Sleep Quality Among Doctors and Medical Students: A Cross-Sectional Study From South India. Cureus, 2024.

[8] Kaggle. Sleep Health and Lifestyle Dataset, 2023/1/1, 2025/7/1, https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

[9] Daily steps and all-cause mortality: a meta-analysis of 15 international cohorts. The Lancet Public Health, 7(3), e219-e228.

[10] Tang X., Guo H., Cui X. Research status of obstructive sleep apnea in women. Journal of Nanjing Medical University (Natural Sciences), 2025, 45(05): 727-736.