To Determine LDL Phenotypes Using Lipids, Lipoproteins, Apoproteins, and sdLDL Through Association Rule Mining

Mehtap Atak,¹ Mehmet Kıvrak,² Hatice Sevim Nalkıran,³ Hüseyin Avni Uydu,⁴ Ömer Şatıroğlu⁵

¹Department of Biochemistry, Recep Tayyip Erdoğan University Faculty of Medicine, Rize, Türkiye
²Department of Biostatistics and Medical Informatics, Recep Tayyip Erdoğan University Faculty of Medicine, Rize, Türkiye
³Department of Medical Biology, Recep Tayyip Erdoğan University Faculty of Medicine, Rize, Türkiye
⁴Department of Medical Biochemistry, Samsun University Faculty of Medicine, Samsun, Türkiye
⁵Department of Cardiology, Recep Tayyip Erdoğan University Faculty of Medicine, Rize, Türkiye

ABSTRACT

Objective: The atherogenic lipoprotein phenotype is closely associated with the risk assessment of Coronary Artery Disease (CAD) and the monitoring of treatment processes. Particularly, high levels of small dense low-density lipoprotein (sdLDL) and low levels of large buoyant low-density lipoprotein (lbLDL) are critical in determining Pattern B. This study aims to determine the lipid phenotype using the Association Rule Mining (ARM) method, based on concentrations of lipids, lipoproteins, apoproteins, and sdLDL.

Materials and Methods: This retrospective case-control study utilized analytical research methods. Numerical variables were expressed as mean, standard deviation, median, and min-max values. Statistically significant differences were observed between the low-density lipoprotein (LDL) size categories in terms of triglycerides (TG), LDL, high-density lipoprotein (HDL), apolipoprotein B (ApoB), apolipoprotein E (ApoE), sdLDL, and lbLDL distributions. ARM was employed to detect the lipoprotein phenotype.

Results: Statistically significant differences were found between the LDL size categories in distributions of TG, LDL, HDL, ApoB, ApoE, sdLDL, and lbLDL (p<0.001, p<0.03, p<0.001, p<0.016, p<0.004, p<0.001, and p<0.001). The ARM method revealed that the probability of phenotype B is 100% for sdLDL values in the range of 15.5–109 and lbLDL values in the range of 0–31.5.

Conclusion: This study introduces a contemporary approach for detecting lipoprotein phenotypes using ARM, further substantiating the strong correlation between atherogenic phenotypes and sdLDL.

Keywords: Lipoproteins, LDL phenotype, coronary artery disease, association rule mining.

INTRODUCTION

Cardiovascular diseases (CVDs) are a significant cause of death globally. The World Health Organization stated that an estimated 17.9 million people died from CVDs in 2019, representing
32% of all global deaths. The risk of CVD is closely associated with changes in the lipoprotein profile in blood plasma. High levels of low-density lipoprotein cholesterol (LDL-C) in plasma are one of the major risk factors for the development of CVD. However, not only plasma LDL-C levels but also particle properties are of great importance in predicting disease risk. Low-density lipoprotein (LDL) particles are divided into subclasses with different diameters and densities, physiochemical compositions, metabolic behaviors, and atherogenic potentials. These subclasses of LDL are determined by various analytical methods, such as density gradient ultracentrifugation, nuclear magnetic resonance, gradient gel electrophoresis, and polyacrylamide gel electrophoresis. LDL consists of two subclasses differing in size and density: large and medium-sized LDL (lbLDL, pattern A) and smaller-sized LDL (sdlDL, pattern B). SdlDL is strongly associated with CVD risk due to its easier penetration into arterial tissue, more efficient passage through the subendothelium, low affinity for the LDL receptor, and high oxidative sensitivity. Additionally, the diameter of LDL particles is an important parameter in determining their atherogenic tendency, and is critical for predicting cardiovascular events. The LDL phenotype in plasma is typically determined by gradient gel electrophoresis. If the LDL particle diameter is 258-263 Å or more, it is classified as ‘Type A’; below this threshold, it is classified as ‘Type B’. The ‘Type A’ phenotype, or pattern A, is usually characterized by large buoyant LDL (lbLDL) particles, while ‘Type B’ phenotype, or pattern B, is characterized by small density LDL (sdlDL) particles.

Data mining can be summarized as a collection of methods used to extract information from data, focusing on the process of discovering previously unknown, hidden patterns. It involves transforming data into qualified information by utilizing statistical analysis methods and artificial intelligence algorithms. Data Mining Methods are instrumental in diagnosing various diseases such as cardiovascular diseases, diabetes, obesity, and cancer. These methods support the process through Machine Learning, developing solutions with self-learning methods.

Association Rule Mining (ARM) aims to extract statistically significant relational patterns from large databases. Increasingly used in medical literature, ARM has become a key method for identifying factors associated with diseases. The Apriori algorithm is the most commonly used in the ARM model, but the Frequent Pattern Growth (FP-Growth) and Eclat Algorithms are also employed. Support and confidence are two crucial measures in creating strong, meaningful rules. The size of the LDL diameter, an important parameter indicating the atherogenic tendencies of lipoproteins, is also significant for predicting cardiovascular events.

### Materials and Methods

This is an observational study within the framework of quantitative research. We employed a case-control analytical research method for this retrospective study. All ARM inferences were generated using the ‘arules’ package in the R programming language. RStudio version 1.1.456 was used for the analysis. The Statistical Package for the Social Sciences (SPSS) 22.0 Package Program was utilized for statistical analysis.

### Dataset

The acquisition of the dataset comprises three steps. The first stage involves collecting serum samples, the second stage pertains to the results obtained through biochemical and clinical analyses, and the third stage focuses on the outputs obtained from Low-Density Lipoprotein (LDL) subfraction analysis. This research adhered to the Declaration of Helsinki and received approval from the Non-Interventional Clinical Research Ethics Committee (Recep Tayyip Erdoğan University Faculty of Medicine, Rize, Türkiye; Decision Number: 2017/163; Date: 27.10.2017). Each participant provided written, informed consent prior to registration. The study included 516 patients who visited the Recep Tayyip Erdoğan University (RTEU) Medical Faculty Teaching & Research Hospital Cardiology Outpatient Clinic, were diagnosed with Coronary Artery Disease (CAD), and met the criteria for elective conventional coronary angiography. Subsequently, fasting blood samples were collected in pre-cooled Ethylenediaminetetraacetic acid (EDTA) tubes at baseline from each patient. After centrifugation at 3,500 rpm for 20 minutes at 4 °C, all fresh plasma aliquots were immediately analyzed in the clinical biochemistry laboratory. In this study, triglyceride (TG) and total cholesterol (TC) concentrations were measured using an autoanalyzer (Abbott Architect C16000).

High-Density Lipoprotein Cholesterol (HDL-C) levels were determined by the dextran sulfate-Mg²⁺ precipitation method. Total LDL-C levels were calculated using Friedewald’s formula (LDL-C=TC-(TG/5+HDL-C)). Apolipoprotein (ApoA, ApoB, ApoE) concentrations were determined using the nephelometric method (Siemens BN 2). Finally, LDL subfractions were measured using the Lipoprint system according to the manufacture
This method separates plasma lipoproteins in a non-denaturing gel gradient polyacrylamide. In the Lipoprint system, lipoproteins are separated based on their varying net surface charges and the size of the LDL particle. Seven types of LDL are identified by this method, classified as lb_LDL and sd_LDL based on LDL particle size. Additionally, LDL particles are divided into ‘large’ and ‘small’ types; the cut-off point for classification was a particle size of 264 Å as determined by Receiver Operating Characteristic (ROC) curve analysis. The area under the ROC curve was found to be 0.997. A particle diameter larger than 264 Å is defined as phenotype A, while a diameter smaller than 264 Å is defined as phenotype B. Individuals with phenotype B are considered to be in a higher risk group for CAD. The cut-off point for LDL particles is presented in Figure 1.

In our dataset, the LDL size variable is used as the output variable. The input variables include TG, TC, LDL, HDL, ApoA, ApoB, ApoE, sd_LDL, and lb_LDL. Association Mining is carried out in two steps: Frequent Itemset (FI) discovery and Association Rule (AR) generation. A randomly selected sample may contain statistical errors, particularly in support and confidence calculations. The error is calculated either at the FI discovery step or the AR generation step and is compared with the corresponding value of the universe, either absolutely or relatively. The classification of sampling size estimation techniques depends on the type of error and the step where the error emerges. The formula for absolute frequent itemset ($f_{abs}$) is given as:

$$f_{abs} = -\frac{2\ln{(1-\gamma)}}{\delta^2}$$

(1)

Where:

- $\gamma$: minimum confidence of AR,
- $\delta$: minimum support of FI,
- $\delta$: failure probability in FI discovery/AR generation step.

Using this formula, the minimum sample size required for Association Rule Mining was calculated to be 395. The classification performance is depicted in a deviation plot in Figure 2.

**Figure 1.** The cut-off point of LDL particles: The area under the ROC curve was 0.997 with a standard error of 0.002 ($p<0.001$). The cut-off value is the one that maximizes the sensitivity/(1-specificity) ratio.

**Figure 2.** Using the specified formula, the minimum sample size in association rule mining was calculated as 395.

$$Hz = \frac{1}{\sqrt{2}} \left( \frac{2d+1}{4} \right)^{1/(d+4)}$$

(2)

Missing values were imputed in the dataset using the Random Forest Method. Additionally, 31 outliers with a random distribution were identified and removed from the dataset, as they were determined to have no effect on the analysis results (Fig. 3).

**Data Preprocessing**

Initially, Henze-Zirkler’s Multiple Normality Test was conducted. The data were found not to be multi-normally distributed at the 95% confidence interval ($p<0.001$) (Table 1). This finding aligns with Henze and Zirkler’s work on multivariate normal distribution, as shown in the following setting:

**Association Rule Mining**

ARM aims to uncover meaningful relationship structures by generating rules from pattern structures within a specific dataset. Support and confidence are two important statistical criteria for interpreting these ratios. Originally utilized in...
marketing. ARM is now employed in medicine, particularly for identifying disease risk factors. The general formulation in ARM is \( X \Rightarrow Y \), where \( X \) represents the rule's premise and \( Y \) its conclusion. The most commonly used method in ARM is the Apriori Algorithm.

**Apriori Algorithm**

The Apriori Algorithm is a method used to identify frequently recurring significant items from a given database through multiple iterative scanning operations. By establishing minimum support and confidence values, items below the minimum support value are disregarded. Subsequently, triple association rules are formulated, and those falling below the minimum support value are excluded from these rules.

**Support**

In a dataset (XY), the support value is expressed as the percentage of rows that include both \( X \) and \( Y \) values. It is represented as:

\[
\text{Support (X} \Rightarrow \text{Y)} = P(XUY).
\]

**Confidence**

Another essential measure for association rules is confidence. Confidence indicates the degree of association discovered. \( P(Y|X) \) represents the probability of including \( Y \) in a transaction that also includes \( X \), and it is defined as follows:

\[
\text{Confidence (X} \Rightarrow \text{Y)} = P(Y|X).
\]

**RESULTS**

**Statistical Analysis**

Numerical variables were expressed as mean, standard deviation, median, and min-max values. As the dataset did not meet the assumptions of multivariate normality, the Mann-Whitney U test, a non-parametric test, was used for group comparisons in quantitative data. The Type 1 error (\( \alpha \)) was set at 0.05. ApoA: Apolipoprotein A; ApoB: Apolipoprotein B; ApoE: Apolipoprotein E; HDL: High density lipoprotein; lb_LDL: Large-buoyant low density lipoprotein; SdLDL: Small-density low density lipoprotein; TC: Total cholesterol; TG: Triglyceride.

---

**Table 2. Descriptive statistics and significance between groups**

<table>
<thead>
<tr>
<th>Variable [Mean±SD/Median (Min–Max)]</th>
<th>Phenotype A (n=283)</th>
<th>Phenotype B (n=163)</th>
<th>U statistics</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>TG</td>
<td>127.8±64.1/112 (37–423)</td>
<td>187.4±106.5/156 (43–674)</td>
<td>-6.5956</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TC</td>
<td>204.3±44.3/203 (109–376)</td>
<td>197.1±44.7/192 (98–322)</td>
<td>1.8195</td>
<td>0.069</td>
</tr>
<tr>
<td>LDL</td>
<td>133.6±38.1/132 (62–301)</td>
<td>125.7±38.01/122.7 (54.9–236)</td>
<td>2.1719</td>
<td>0.030</td>
</tr>
<tr>
<td>HDL</td>
<td>45.3±12.5/43.8 (18–89)</td>
<td>36.9±8.7/36 (20.3–65.8)</td>
<td>7.2168</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ApoA</td>
<td>129.1±30.9/126 (72.9–281)</td>
<td>126.1±27.4/123 (71.1–203)</td>
<td>0.6968</td>
<td>0.486</td>
</tr>
<tr>
<td>ApoB</td>
<td>103.8±28.1/104 (35–188)</td>
<td>111.1±28.8/108.5 (49.3–223)</td>
<td>-2.4128</td>
<td>0.016</td>
</tr>
<tr>
<td>ApoE</td>
<td>4.2±1.3/4.1 (1.6–14.1)</td>
<td>4.75±1.6/4.35 (2.1–11.8)</td>
<td>-2.8618</td>
<td>0.004</td>
</tr>
<tr>
<td>SdLDL</td>
<td>4.5±3.6/4 (0–23)</td>
<td>50.6±25.3/51 (6–109)</td>
<td>-17.115</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>lbLDL</td>
<td>66.2±23.3/65 (12–155)</td>
<td>25.1±20.5/19 (0–121)</td>
<td>14.115</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Numerical variables are expressed as mean, standard deviation, median, and minimum–maximum values. Due to the dataset's inability to satisfy multiple normality assumptions, the Mann-Whitney U test, a non-parametric test, was used for comparing groups in quantitative data. The Type 1 error (\( \alpha \)) was set at 0.05. ApoA: Apolipoprotein A; ApoB: Apolipoprotein B; ApoE: Apolipoprotein E; HDL: High density lipoprotein; lb_LDL: Large-buoyant low density lipoprotein; SdLDL: Small-density low density lipoprotein; TC: Total cholesterol; TG: Triglyceride.

**Figure 3.** Outlier distribution: The x-axis represents the local outlier factor, while the y-axis shows the observation points. A total of 31 rows containing outliers or excessive values were identified in the dataset and subsequently removed from the dataset based on random selection.
DISCUSSION

In this study, ARM was employed to determine the impact of serum lipid, lipoprotein, and LDL subfractions on the formation of atherogenic (Pattern B) and anti-atherogenic (Pattern A) LDL phenotypes in CAD. Additionally, individuals were categorized based on Patterns A and B, and lipid parameters were analyzed within these groups. It was observed that serum lipid levels (TG, TC, and LDL-C) were higher in individuals with Pattern B. Furthermore, sdLDL levels were found to be lower in the Pattern A group compared to the Pattern B group, while lbLDL levels were higher in the Pattern A group compared to the Pattern B group.

Previous literature has indicated that serum sdLDL is associated with CVD risk, independent of LDL-C concentration. In this study, no difference in LDL-C concentration was observed between the groups. However, the high concentration of TG and sdLDL, combined with low levels of HDL-C, are considered significant markers of dyslipidemia and are characteristic of the atherogenic lipoprotein phenotype (Pattern B). Additionally, the Pattern B exhibited high levels of TG and sdLDL along with low HDL-C levels. These findings align with the existing literature.

SdLDL formation is closely associated with high serum TG levels (>120 mg/dL), low HDL-C levels, and increased hepatic lipase enzyme activity. TGs and cholesterol esters are exchanged between lipoprotein particles via the cholesteryl ester transfer protein (CETP). When serum TG levels are high and HDL levels are low, TGs in the Very Low-Density Lipoprotein (VLDL) structure are transferred to LDL, instead of HDL, via CETP. The newly formed LDL particle becomes rich in TGs. Following the hydrolysis of TGs in this particle by hepatic lipase, the residue is called sdLDL. Our findings showed that TG levels were higher and HDL levels were lower in the Pattern B group, leading to a more dominant sdLDL particle in the plasma, which is consistent with previous research.

LDL particle diameter is a key factor in calculating sdLDL concentration. Those with an LDL particle diameter greater than 264 Å are classified as lbLDL (Pattern A), and those with smaller diameters as sdLDL (Pattern B). LDL particle diameter is critical in determining the pattern. However, Munusuru et al. stated that, in addition to particle diameter, factors such as VLDL, LDL, LDL subfractions (sdLDL, lbLDL), and HDL subfractions (HDL2, HDL3) play significant roles in determining the atherogenic lipoprotein phenotype. Similarly, in our study, concentrations of lipids, lipoproteins, apoproteins, and sdLDL were used to determine the LDL phenotype using ARM. It is essential to note that for identifying Pattern B, it is not only necessary to have high sdLDL levels but also to have low lbLDL levels. Serum apoprotein (ApoA, ApoB, ApoE) concentrations within certain limits are also effective in determining the pattern. Likewise, in defining Pattern A, it was determined that, in addition to lbLDL, ApoA, ApoE, HDL, and sdLDL should be within specific ranges.
Table 3. Generated association rules

<table>
<thead>
<tr>
<th>Rule no</th>
<th>Association rules (X⇒Y)</th>
<th>Support</th>
<th>Confidence</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sdLDL=(15.5–109), lbLDL=(0–31.5)</td>
<td>0.289</td>
<td>1</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>LDL_Size=Phenotype B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>HDL=(18–37), sdLDL=(15.5–109)</td>
<td>0.207</td>
<td>1</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>LDL_Size=Phenotype B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>TG=(40–232), TC=(200–376), sdLDL=(0–15.5)</td>
<td>0.298</td>
<td>0.982</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>LDL_Size=Phenotype A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HDL: High density lipoprotein; lb_LDL: Large-buoyant low density lipoprotein; SdLDL: Small-density low density lipoprotein; TC: Total cholesterol; TG: Triglyceride.

Figure 4. Generated association rules: LDL size was set as the response output variable, and Phenotype B was identified as the positive predictor class. These results were obtained using the Ameva discretization method, incorporating predictive variables with confidence values of 0.2 for support and 0.5 for classification. The classification-based associations rules method was utilized in the algorithm.
CONCLUSION
In conclusion, the atherogenic lipoprotein phenotype is closely associated with the determination of CAD risk and the monitoring of treatment. This study presents a novel approach to defining lipoprotein phenotypes using ARM and corroborates prior research on the impact of sdLDL on lipoprotein phenotypes.

Peer-review: Externally peer-reviewed.

Ethics Committee Approval: The Recep Tayyip Erdoğan University Non-Interventional Clinical Research Ethics Committee granted approval for this study (date: 27.10.2017, number: 2017/163).

Informed Consent: Written informed consent was obtained from patients who participated in this study.

Author Contributions: Concept – MA, MK; Design – MA, MK; Supervision – MA, HAU; Resource – HAU, ÖŞ, MA; Materials – ÖŞ; Data Collection and/or Processing – ME, ÖŞ; Analysis and/or Interpretation – MA, ÖŞ, HAU; Literature Search – MA, MK, HSN; Writing – HSN, MA, MK; Critical Reviews – HSN, HAU, ÖŞ.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES
4. Rizzo M, Berneis K. Should we measure routinely the LDL peak particle size? Int J Cardiol 2006; 107(2): 166–70. [CrossRef]
5. Diffenderfer MR, Schaefer EJ. The composition and metabolism of large and small LDL. Curr Opin Lipidol 2014; 25(3): 221–6. [CrossRef]
14. Vega GL, Ma PTS, Cater NB. Effects of adding fenofibrate (200 mg/day) to simvastatin (10 mg/day) in patients with combined hyperlipidemia and metabolic syndrome. Am J Cardiol 2003; 91(8): 956–60. [CrossRef]
16. Riondato M, Upfal E. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. ACM Transactions on Knowledge Discovery from Data (TKDD) 2014; 8(4): 1–32. [CrossRef]


27. Hanak V, Munoz J, Teague J, Stanley AJr, Bittner V. Accuracy of the triglyceride to high-density lipoprotein cholesterol ratio for prediction of the low-density lipoprotein phenotype B. Am J Cardiol 2004; 94(2): 219–22. [CrossRef]

