An evidence based Decision Support Model to Discover the Characteristics of the Elderly with Depression

Myonghwa Park¹, Chang Sik Son², Sun Kyung Kim¹

¹ College of Nursing, Chungnam National University, Daejeon, 301-747, Republic of Korea
² Department of Medical Informatics, School of Medicine, Keimyung University, Daegu
mhpark@cnu.ac.kr

Abstract The aim of this study was to develop a reliable evidence based decision support model with combination of statistical analysis and decision tree algorithms. We used the large data set of 2008 Korean Elderly Survey (KES) which was consisted of 14,970 elderly data. Target variable was depression and input variables were demographic, health related and socioeconomic characteristics. Statistical analysis including the Chi-square, Fisher’s exact test, the Mann-Whitney U-test and Wald logistic regression was used as a feature selection process. The final decision support models were built and C5.0 tree showed the accuracy of 81.6%. The decision model can be applied as an aid in the decision making for clinicians to increase vigilance with suspected depression in elderly population

Keywords: Data mining, Logistic regression, Depression, Aged

1 Introduction

1.1 Background

Huge amounts of data generated by healthcare settings are complex and voluminous to be processed. Data mining can improve decision-making by discovering patterns and trends in complex healthcare data. Understandings gained from data mining can influence cost and managing efficiency while maintaining a high level of care [1].

Depression is a major contributor to healthcare costs in the elderly populations, causing overwhelming burden on the healthcare system [2]. Having experienced all those losses, elderly became more vulnerable to depressions. A number of factors are involved in the elderly depression not only socio-demographic factors but also functional level and social support are associated with depression [3].

Using large database, application of data mining can result in the diagnosis and prognosis and even the discovery of hidden patterns in depression [4]. It is therefore

* This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0024922).
required to build decision tree model of depression for the better understanding of characteristics of elderly depression and prediction of its risk factors.

1.2 Purpose of the Study

The aim of this study is to develop a reliable hybrid decision support model with combination of statistical analysis and decision tree algorithms to ensure high accuracy in screening of elderly with depression.

2 Method

2.1 Data Source

This study used the data from the 2008 national study of Korean elderly to determine the characteristics according to the depressive symptom of the elderly in Korea with the Institutional Review Board (IRB) approval. The 2008 national study of Korean elderly surveyed 15,146 older adults (Male: 6,452, Female: 8,694) over 60 years who lived in community and investigated their health and economic condition, and welfare status [5].

2.2 Decision Tree Models and Statistical Data Analysis

Feature selection node was used based on the p-value for categorical prediction and analysis node was used to compare the percentage of correct and wrong classification of the models. ROC (Receiver Operating Characteristics) curves were driven to see the predictive performance of the models. The C5.0 Tree was selected to design the final decision support model which provides a simple representation of accumulated knowledge, facilitating a decision making process. This model selects the best decision node that separates the different classes from the empirical data. The whole process repeats until the subsamples are not able to split any further. The splits in the lowest level are reexamined, filtering those with no significant contribution to the value of the model.

The decision tree model was built using the default experimental parameters of Clementine version 12.0 (SPSS Inc., Chicago, IL, USA). The rules were evaluated with the testing data set for their predictability. Statistical analysis was performed using SPSS 18.0 for Windows (SPSS Inc., Chicago, IL, USA). Univariate correlations were evaluated using the Chi-square test or Fisher’s exact test [6] and Kolmogorov-Smirnov test were conducted first to test normality. The variables were then analyzed further by either of Student t-test and Mann-Whitney U-test. A two tailed \( p<0.05 \) was selected as the level of statistical significance. In the multivariate analysis, Wald forward selection was used with entry and removal criteria of 0.05 and 0.10. The results were expressed as the odd ratios (OR) with 95% confidence intervals.
3 Testing and Results

3.1 Demographic Characteristics of the Subjects

Of the 14,970 subjects in the data set and 4,423 (29.54%) participants had depression. Significant differences were observed in terms of gender, age, education level, marital status, family style, work status, and religion.

3.2 Performance of Decision Support Models Based on Multivariate Analysis

The performance of the model was evaluated using six standard measures including accuracy (ACC), sensitivity (SENS), positive predictive value (PPV), negative predictive value (NPV), and the area under the ROC curve (AUC) (Table 1).

<table>
<thead>
<tr>
<th>Methods</th>
<th>ACC</th>
<th>SENS</th>
<th>SPEC</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
<th>no. rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>81.6</td>
<td>60.0</td>
<td>90.6</td>
<td>73.3</td>
<td>84.4</td>
<td>75.3</td>
<td>11</td>
</tr>
</tbody>
</table>

ACC, accuracy; SENS, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value; AUC, area under ROC curve

3.3 Decision Support Model Based on Multivariate Analysis

Ten out of the total of 23 variables, including overall satisfaction of living, average life satisfaction score, limitation in ADL and IADL, nutritional status score, exercise capacity score, health condition and perceived financial status were selected using the C5.0 tree algorithm. Having cut off points determined by the decision tree algorithm, the criteria for dichotomizing the continuous variables were all statistically significant (<0.001).

<table>
<thead>
<tr>
<th>Levels</th>
<th>Contents</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Overall life satisfaction</td>
<td>10.728 (9.794–11.752)</td>
</tr>
<tr>
<td>Level 1</td>
<td>Life satisfaction ≤ 3.462</td>
<td>0.161 (0.145–0.178)</td>
</tr>
<tr>
<td></td>
<td>Life satisfaction &gt; 3.462</td>
<td>6.225 (5.615–6.901)</td>
</tr>
<tr>
<td></td>
<td>Health status</td>
<td>5.377 (4.393–6.580)</td>
</tr>
<tr>
<td>Level 2</td>
<td>Limitation in ADL (no or yes)</td>
<td>4.194 (3.406–5.164)</td>
</tr>
<tr>
<td></td>
<td>Nutritional status ≤ 4</td>
<td>0.292 (0.241–0.355)</td>
</tr>
<tr>
<td></td>
<td>Nutritional status &gt; 4</td>
<td>3.424 (2.820–4.158)</td>
</tr>
<tr>
<td></td>
<td>Life satisfaction ≤ 2.944 or &gt; 2.944</td>
<td>4.241 (2.776–6.479)</td>
</tr>
<tr>
<td>Level 3</td>
<td>Nutritional status ≤ 5 or &gt; 5</td>
<td>3.211 (2.794–3.691)</td>
</tr>
<tr>
<td></td>
<td>Overall life satisfaction (low)</td>
<td>0.177 (0.106–0.295)</td>
</tr>
</tbody>
</table>
We identified ten associated rules with the depression and these rules were as follows: 1) moderate to high overall life satisfaction, average life satisfaction score was ≤3.462, having limitation in ADL; 2) moderate to high overall life satisfaction, average life satisfaction score was ≤3.462, no limitation in ADL, nutritional status score was between 5 and 10, exercise capacity score < 73.333; 3) moderate to high overall life satisfaction, average life satisfaction score was ≤ 3.462, no limitation in ADL, exercise capacity score ≤ 73.333, nutritional status score was > 10, no limitation in IADL; 4) moderate to high overall life satisfaction, average life satisfaction score was ≤ 3.462, no limitation in ADL, exercise capacity score < 73.333, nutritional status score was > 10, having limitation in IADL, perceived financial status was poor; 5) moderate to high overall life satisfaction, average life satisfaction score was ≤ 3.462, no limitation in ADL, average life satisfaction score ≥ 3.214, having limitation in IADL, perceived health status was bad, perceived functional status was either rich or poor; 6) low overall life satisfaction, perceived health status was bad, nutritional status score was > 4; 7) low overall life satisfaction, perceived health status was moderate to good, average life satisfaction score was ≤ 2.944; 8) low overall life satisfaction, perceived health status was bad, nutritional status score was ≤ 4, very low overall life satisfaction score; 9) low overall life satisfaction, perceived health status was bad, nutritional status score was ≤ 4, low overall life satisfaction score, having limitation in IADL; 10) low overall life satisfaction, perceived health status was bad, nutritional status score was ≤ 4, low overall life satisfaction score, no limitation in IADL, days of having sound sleep were 1-2 days per week or none.

4 Conclusion

Generally, the C5.0 Tree model showed good performance in predicting characteristics of depression in this study. Although, there are a variety of methods to analyze the clinical decisions, tree classification techniques have a few benefits over the alternative techniques [7]. The interpretation of results in a tree model can be useful not only for rapid classifying new clinical observations, but for explaining why observations are classified or predicted in a particular manner. In addition, performing data mining tasks in health care data, tree decision support models are particularly
suitable as there is little priori knowledge that show which variables are related and how [8].

This study developed a reliable decision support model performing statistical analyses and using a decision tree algorithm to provide high accuracy in decision making process using the derived rules to facilitate discriminatory knowledge discovery in a large healthcare data set. The experimental results show that the C 5.0 Tree with logistic regression provided excellent discrimination and we demonstrated its feasibility. Therefore, the decision model developed in this study can be applied to identify the high risk group of depression and to increase vigilance when detecting depression in the elderly in the community.

References