Acta Medica Okayama

Volume 59, *Issue* 1 2005 *Article* 3

FEBRUARY 2005

Determination of depression risk factors in children and adolescents by regression tree methodology.

Handan Camdeviren[*] Mehmet Mendes[†] M. Muhip Ozkan[‡]
Fevziye Toros[**] Tayyar Sasmaz[††] Seva Oner[‡‡]

[*]Mersin University,
[†]&#xFEFF;Çanakkale Onsekiz Mart University,
[‡]&#xFEFF;Ankara University,
[**]Mersin University,
[††]&#xFEFF;Mersin University,
[‡‡]&#xFEFF;Mersin University,

Determination of depression risk factors in children and adolescents by regression tree methodology.[*]

Handan Camdeviren, Mehmet Mendes, M. Muhip Ozkan, Fevziye Toros, Tayyar Sasmaz, and Seva Oner

Abstract

We used a regression tree method (RTM) to determine risks of depression in children/adolescents. The survey records of 4,143 children/adolescents in a study based in Mersin, Turkey served as data in this study, and multi-step, stratified, and cluster sampling were used. Effects of 24 variables (sex, smoking, parental problems, etc.) were evaluated on depression scores. The Child Beck Depression Inventory (CBDI) was used to determine the level of depression. Subjects were into 12 different groups based on magnitudes of mean depression scores. The interactions among 7 variables determined to be risk factors are shown on a schema. The STATISTICA (ver.6.0) package program was used for all computations. Although traditional statistical methods have often been used for analysis in this field, such approaches are associated with certain disadvantages such as missing values, ignorance of interaction effects, or restriction of the shape of the distribution. To avoid such disadvantages, we therefore suggest the use of the RTM in studies involving numerical-based outcome variables and for the investigation of a large number of variables and it may be more effective than traditional statistical methods in epidemiological studies which determine risk factors.

KEYWORDS: children and adolescents, Beck depression inventory, classification and regression trees, cross-yalidation, diagnostic models

[*]PMID: 15902995 [PubMed - indexed for MEDLINE]

Acta Med. Okayama, 2005
Vol. 59, No. 1, pp. 19-26

Acta Medica
Okayama

http://www.lib.okayama-u.ac.jp/www/acta/

Original Article

Determination of Depression Risk Factors in Children and Adolescents by Regression Tree Methodology

Handan Çamdeviren[a]*, Mehmet Mendeş[b], M. Muhip Özkan[c], Fevziye Toros[d], Tayyar Şaşmaz[e], and Seva Öner[e]

[a]*Department of Biostatistics, Mersin University School of Medicine,* [b]*Department of Biometry & Genetics, Çanakkale Onsekiz Mart University,* [c]*Department of Biometry & Genetics, Ankara University,* [d]*Department of Child and Adolescent Psychiatry, Mersin University School of Medicine, and* [e]*Department of Public Health, Mersin University School of Medicine, Turkey*

We used a regression tree method (RTM) to determine risks of depression in children/adolescents. The survey records of 4,143 children/adolescents in a study based in Mersin, Turkey served as data in this study, and multi-step, stratified, and cluster sampling were used. Effects of 24 variables (sex, smoking, parental problems, *etc*.) were evaluated on depression scores. The Child Beck Depression Inventory (CBDI) was used to determine the level of depression. Subjects were into 12 different groups based on magnitudes of mean depression scores. The interactions among 7 variables determined to be risk factors are shown on a schema. The STATISTICA (ver.6.0) package program was used for all computations. Although traditional statistical methods have often been used for analysis in this field, such approaches are associated with certain disadvantages such as missing values, ignorance of interaction effects, or restriction of the shape of the distribution. To avoid such disadvantages, we therefore suggest the use of the RTM in studies involving numerical-based outcome variables and for the investigation of a large number of variables and it may be more effective than traditional statistical methods in epidemiological studies which determine risk factors.

Key words: children and adolescents, Beck depression inventory, classification and regression trees, cross-validation, diagnostic models

The diagnosis and subsequent classification of individuals, often referred to simply as "classification" and "regression," respectively, are important to medical studies, as they facilitate the determination of risk factors and the estimation of relevant parameters. In studies intended to identify risk factors, if an outcome variable is categorical or is transformed into a categorical variable, classification methods have traditionally been applied. If numerical variables are obtained from individuals by direct measurement or by a point-score evaluation system, then a regression model tends to be applied, provided that the significant effect on variation of the variable has been examined. Regression methods are typically also used to evaluate more than one variable simultaneously, and these variables are examined in multivariate analyses of groups [1-3].

In studies involving numerical-based outcome variables, the variables are transformed into categorical structures using a suitable cut-off value. In such cases, traditional statistical methods such as logistic regression,

Received April 9, 2004; accepted September 22, 2004.

*Corresponding author. Phone: +90-324-3412815; Fax: +90-324-3412400
E-mail: hcamdeviren@hotmail.com (H. Çamdeviren)

discriminate analyses, Pearson's chi-square test, and ANOVA models have been used to determine the risk factors [5-12]. However, transforming such variables into categorical structures leads to a loss of information, such that regression models, which can be a direct estimation of the numerical measure, provide much more detailed and accurate results [2, 13].

Tree-based models are among the most widely applied models recently used for diagnosis and data mining. These models are useful for modeling data and devising standard rules to make modeling decisions, *i.e.*, rules to be used in cases in which the comprehension of underlying processes and user confidence in the results are just as important as error minimization. For these reasons, tree models are often applied to investigate medical problems [1]. Currently, most statistical packages employ some form of tree-based modeling; in addition, there has been increasing interest in developing regression models for large datasets that are both accurate and easier to interpret than the more traditional statistical methods [15-18].

The regression tree method (RTM) is a tree-based model. When compared to traditional statistical methods, this method has both advantages and disadvantages. The regression tree method in particular is more useful than traditional methods when a data set is large, and when the number of variables is high. Moreover, the RTM method does not ignore interactions among factors, and it is not affected by high correlations between risk factors. In addition, the RTM method remains unaffected by missing values. The present method employs a surrogate variable to replace any variable that has a missing value. It should be noted that the results obtained by the RTM are based on visually presented data; this approach facilitates the interpretation of the results of an analysis [1, 19, 20]. On the other hand, determining the most suitable tree structure (optimal tree), as well as interpreting the results can be problematic.

This study aimed to determine risk factors that affect the scores of children and adolescents on the Beck Depression Inventory by using the RTM as an alternative method to traditional statistical methods.

Materials and Methods

Study Plan. This cross-sectional study considered the data obtained from 4,143 children and adolescents between ages of 11 and 20 who attended secondary and high schools in Mersin in 2002.

Sampling Procedure. There were 86 secondary and 36 high schools included in the study (total: 81,676 students). If a depression prevalence of 12% (± 1 standard deviation) is considered, then the minimum sample size would be 3,865 for ages 11-20. On the other hand, a suitable sample size was determined to be 4,500 students, *i.e.*, 5.5% of the population. The present study sample was composed of 4,256 students (94.6% of 4,500). During the data quality control process, 113 children/adolescents were excluded from the study due to missing or unreadable answers, and thus the final analysis included 4,143 students (92.1% of 4,500 students).

The following types of sampling were used in the survey study: multi-step, stratified, and cluster sampling. In the present study, we obtained data from a questionnaire. The schools were classified in terms of their socio-economic status as follows: good, satisfactory, and poor. According to the weight of each group, we randomly selected 12 secondary and 6 high schools as the study area. Classes were chosen by random selection according to the number of students attending that particular school.

Preparing of Survey. During the in-school interview, all children were administered a detailed, structured questionnaire and the Child Beck Depression Inventory (CBDI). The CBDI is a 26-item scale that was developed to measure depressive symptomatology in children and adolescents. The focus here was on how each subject had been feeling for the past 2 weeks, and some items emphasized symptom intensity over frequency or persistence [21]. The questionnaire included the demographic, the clinical characteristics, and the risk factors for depression in children and adolescents (*i.e.*, substance use such as cigarettes, bully, and alcohol; the break-up of a romantic relationship; failure at school; and loss of a loved one). The characteristics and descriptive statistics of this questionnaire are given in Table 1.

Statistical Analysis. We used a regression tree method, which was found to be a more useful method than other, more traditional statistical methods for the determination of risks for depression.

The RTM steps followed in this study can be summarized as described below.

The learning sample (L) was indicated as n paired (y, x). Here, x variables were risk factors or covariates. These variables formed both the numerical and categorical variables. The risk factors used in this study are described in Table 1. According to the RTM used here, y

Table 1 Questions in questionnaire and descriptive statistics

Question number	Questions	Categories	Frequences as n (%) and Mean ± SD
Question 1	Sex	Boys	2252 (54.4)
		Girls	1891 (45.6)
Question 2	Have you ever failed?	Yes	399 (9.6)
		No	3744 (90.4)
Question 3	Has a problem with mother or father?	Yes	735 (17.7)
		No	3408 (82.3)
Question 4	Living with a stepparent?	No	4038 (97.5)
		Yes	105 (2.5)
Question 5	Is there a somatic finding (s)?	No	1067 (25.8)
		Yes	3076 (74.2)
Question 6	Did you leave one of your close friends within the last year?	Yes	1972 (47.6)
		No	2171 (52.4)
Question 7	Any fall in examination grades within the last year?	Yes	1778 (42.9)
		No	2365 (57.1)
Question 8	Have you had a traffic accident within the last year?	Yes	336 (8.1)
		No	3807 (91.9)
Question 9	Any problems with classmates within the last year?	Yes	2014 (48.6)
		No	2129 (51.4)
Question 10	Any decrease in attention within the last two weeks?	Yes	1626 (39.2)
		No	2517 (60.8)
Question 11	Feeling tired, exhausted or weak within the last two weeks?	Yes	2379 (57.4)
		No	1764 (42.6)
Question 12	Have you ever felt guilty or worthless within the last two weeks?	Yes	1022 (24.7)
		No	3121 (75.3)
Question 13	Any sleeping problem within the last two weeks?	Yes	1731 (41.8)
		No	2412 (58.2)
Question 14	Have you ever overslept within the last two weeks?	Yes	760 (18.3)
		No	3383 (81.7)
Question 15	Lack of satisfaction from things which were satisfactory before?	Yes	1235 (29.8)
		No	2908 (70.2)
Question 16	Any increase in weight gain or appetite within the last two weeks?	Yes	1035 (25.0)
		No	3108 (75.0)
Question 17	Any loss in weight or appetite within the last two weeks?	Yes	952 (23.0)
		No	3191 (77.0)
Question 18	Have you ever felt upset or alone?	Yes	1756 (42.4)
		No	2387 (57.6)
Question 19	Have you ever felt pessimistic within the last two weeks?	Yes	1981 (47.8)
		No	2162 (52.2)
Question 20	Do you get punishment at home?	Yes	1294 (31.2)
		No	2849 (68.8)
Question 21	Are you humiliated by teachers at school?	Yes	2143 (51.7)
		No	2000 (48.3)
Question 22	Have you ever smoked cigarette?	Yes	1056 (25.5)
		No	3087 (74.5)
Question 23	Dou you take alcohol?	None	3126 (75.5)
		At least one a week	97 (2.3)
		One a month	255 (6.2)
		Rarely	581 (14.0)
		Give up	84 (2.0)
Question 24	Age of children		14.53 ± 1.89
Question 25	Depression score		11.23 ± 6.44

is a numerical variable considered as the dependent variable. In this study, y indicates the Beck depression score. The Beck depression scores were then grouped as 2 homogenous sub-groups, recursively. In each separation, we used a risk factor and its cut-off value.

For each risk factor, there were S values that served as candidate values for splitting the nodes. Among these values, S* indicated the best split criterion. Moreover, S* formed the best homogeneous sub-group [22, 23].

The beginning node, *i.e.*, the "root node," was the most heterogeneous. Other homogenous sub-groups were referred to as "terminal nodes," and still other sub-groups were "child nodes".

At the initial stage, the building of a regression tree began with a root node, which contained all of the subjects; then, a series of yes/no questions generated descendant nodes. Beginning with the first node, the regression tree found the best possible variable to split the root node into 2 child nodes. In order to find the best variable, the software selected all possible splitting variables (called "splitters"), as well as all possible values of the variable that could be used to split the node. In choosing the best splitter, the program sought to maximize the average "purity" of the 2 child nodes. These nodes were more homogenous than the root node. If the splitting variable value of a subject was smaller than the determined cut-off value, the subject was allocated to the left child node; if this value was equal to or greater than the cut-off value determined, the subject was allocated to the right child node. The least squared deviation (LSD) method was used as a measure of the homogeneity of the nodes [24, 25].

After the subsequent nodes were further classified as either a left node or a right node, a decrease in variance was obtained as follows.

$$\begin{aligned}\phi(t) = {} & \frac{1}{N(t)}\sum_{i\in t}[y_i - \overline{y}(t)]^2 - \\ & p_L\frac{1}{N(t_L)}\sum_{i\in t_L}[y_i - \overline{y}(t_L)]^2 - \\ & p_R\frac{1}{N(t_L)}\sum_{i\in t_R}[y_i - \overline{y}(t_R)]^2\end{aligned}$$

Here, p_L is the proportion of cases in parent node t classified in the left child node t_L; p_R is the proportion of cases in parent node t classified in the right child node t_R; y_i is the value of the dependent variable for the experimental case i; $\overline{y}(t)$ is the average value of parent node t; $\overline{y}(t_x)$ is the average value of child node tx; and $N(t_x)$ is the number of cases classified in child node tx.

The tree building proceeded until continuation became impossible. The process was stopped under one of the following conditions: (a) there was only one observation in each of the child nodes; or (b) all observations within each child node had an identical distribution of predictor variables, leading to splitting. The maximum tree value was obtained after the tree reached a maximum dimension. An important issue in choosing the size of regression trees is the intended use; larger trees provide greater accuracy for predicting sites in which the response has not yet been measured. On the other hand, smaller trees may be more appropriate for understanding and interpreting relationships among the data, because nodes accounting for less deviance provide less information [17].

For this reason, backward pruning was applied in order to eliminate problems with over-fitting. The cost-complexity parameter was taken into account in the pruning process. The definition of the cost-complexity measure was characterized by a complexity parameter, namely, including a penalty for additional terminal nodes. Such parameters were expressed in terms of the decrease in impurity [23, 26].

Branches were cut in cases when the complexity parameter achieved a minimum value; then, after pruning, the most appropriate tree construction was considered as the optimal tree rendered by pruning. The mean and variance of y (*i.e.*, the Beck depression score) was estimated from n measurements that were in nodes of the optimum tree.

In addition, we used 10-fold cross-validation as an error estimation method; this method is known to be the most acceptable method of estimation in such cases. Cross-validation, or a test sample, was used to provide estimates of the future prediction error for each sub-tree [27, 28].

Results

The descriptive statistics are given in Table 1 as Mean ± SD and frequencies (count and percentage). At the first step of the computations, we separately evaluated the effect of each of 24 questions on the scores of children and the adolescents on the Child Beck Depression Inventory. The importance of each question was calculated, and the results are given in Table 2. Because question 12 reflected the highest score, it was determined to be the most important variable, and question 12 was followed by

Table 2 Variable importance values

Questions (risk factors)	Variable Rank	Importance
Question 1	8	0.08
Question 2	15	0.15
Question 3	73	0.73
Question 4	1	0.01
Question 5	4	0.04
Question 6	39	0.39
Question 7	65	0.65
Question 8	12	0.12
Question 9	58	0.58
Question 10	31	0.31
Question 11	46	0.46
Question 12	100	1.00
Question 13	17	0.17
Question 14	10	0.10
Question 15	39	0.39
Question 16	1	0.01
Question 17	37	0.37
Question 18	63	0.63
Question 19	77	0.77
Question 20	32	0.32
Question 21	43	0.43
Question 22	41	0.41
Question 23	11	0.11
Question 24	13	0.13

question 19 in this regard (Table 2). Then, questions 12 and 19 were used to classify the Beck depression scores. However, when Fig. 1 was examined, the first division began with the variable from question 18; the variables from questions 3 and 19 were used for the second division. These conditions resulted from the relationships between the predictor variables.

In order to obtain the optimum tree, we created the maximum tree. The Cross Validation (CV) cost and the resubstitution cost values of the optimum tree were calculated as being 32.41 ± 0.99 and 31.11, respectively. These values reveal the estimation of error variance of the tree with respect to 2 different methods. In the 3 that was selected for this study, these error variances were balanced and achieved the lowest value. In general, the minimization of error variance is an indicator of how close this prediction model is to the actual value. A balance between 2 error variances indicates that the validation of the tree structure is high and is an indicator for the use of this tree structure for the future analysis of other data sets. The cost-complexity parameter, which was a measure of complexity, was 0.15 for the selected tree (*i.e.*, the optimal tree). The variance in the root node and in the terminal nodes (total) was estimated at 41.44 and

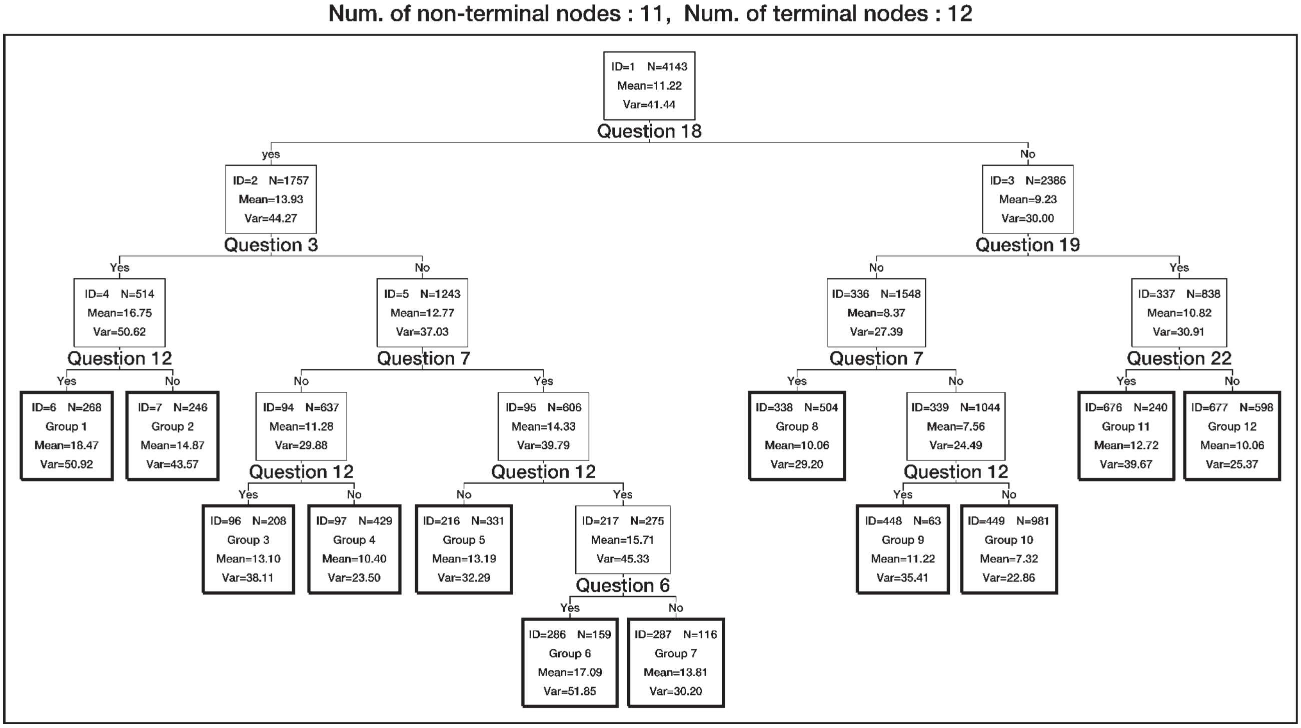
Num. of non-terminal nodes : 11, Num. of terminal nodes : 12
ID=1 N=4143 Mean=11.22 Var=41.44
Question 18
yes
No
ID=2 N=1757 Mean=13.93 Var=44.27
Question 3
ID=3 N=2386 Mean=9.23 Var=30.00
Question 19
Yes
No
ID=4 N=514 Mean=16.75 Var=50.62
Question 12
ID=5 N=1243 Mean=12.77 Var=37.03
Question 7
No
Yes
ID=336 N=1548 Mean=8.37 Var=27.39
Question 7
ID=337 N=838 Mean=10.82 Var=30.91
Question 22
Yes
No
ID=6 N=268 Group 1 Mean=18.47 Var=50.92
ID=7 N=246 Group 2 Mean=14.87 Var=43.57
No
Yes
ID=94 N=637 Mean=11.28 Var=29.88
Question 12
ID=95 N=606 Mean=14.33 Var=39.79
Question 12
Yes
No
ID=338 N=504 Group 8 Mean=10.06 Var=29.20
ID=339 N=1044 Mean=7.56 Var=24.49
Question 12
Yes
No
ID=676 N=240 Group 11 Mean=12.72 Var=39.67
ID=677 N=598 Group 12 Mean=10.06 Var=25.37
Yes
No
ID=96 N=208 Group 3 Mean=13.10 Var=38.11
ID=97 N=429 Group 4 Mean=10.40 Var=23.50
No
Yes
ID=216 N=331 Group 5 Mean=13.19 Var=32.29
ID=217 N=275 Mean=15.71 Var=45.33
Question 6
Yes
No
ID=448 N=63 Group 9 Mean=11.22 Var=35.41
ID=449 N=981 Group 10 Mean=7.32 Var=22.86
Yes
No
ID=286 N=159 Group 6 Mean=17.09 Var=51.85
ID=287 N=116 Group 7 Mean=13.81 Var=30.20

Fig. 1 Structure of optimal tree.

31.11, respectively, for the optimal tree. Thus, the decrease in the variance of the optimal tree was 10.33, which was found to be significant ($P < 0.05$). Therefore, it can be said that the Beck depression scores were homogenous in the terminal nodes of the optimal tree. The total variance in the terminal nodes was assessed as the resubstitution cost. In addition, the correlation coefficient between the actual Beck depression score and the predicted score was found to be 0.50, according to the tree structure ($P < 0.01$). This result indicated that the optimal tree was successful at predicting the point score for depression.

The structure of the optimum tree is given in Fig. 1. In Fig. 1, nodes shown with a bold black square line represent terminal nodes, and the other nodes are root and child nodes. For these nodes, the value that is located in the upper right corner is the number of observations in each group (nodes), and Mu and Var are the mean and variance of depression scores of these groups, respectively. The number located in the upper left corner represents the identification number (ID) of these groups.

Among the 24 questions that were taken into consideration in order to investigate the effect of their corresponding answers on the scores of children and adolescents on the Beck Depression Inventory, only 7 were selected (questions 3, 6, 7, 12, 18, 19, and 22). These selected questions were more effective than the others at estimating the Beck depression scores. From these 7 questions, we formed 12 terminal nodes. Each of these nodes was considered as a homogenous group. The characteristics of these homogenous groups are summarized below.

Group 1. Among the children who felt upset or alone and had a problem with their parents within the last 2 weeks, those who felt guilty or worthless within the last 2 weeks.

Group 2. Among the children who felt upset or alone and had a problem with their parents within the last 2 weeks, those who did not feel guilty or worthless within the last 2 weeks.

Group 3. Among the children who felt upset or alone and did not have a problem with their parents within the last 2 weeks, those who did not experience a drop in grades within the last year and who did feel guilty or worthless within the last 2 weeks.

Group 4. Among the children who felt upset or alone and had no problem with their parents within the last 2 weeks, those who did not experience a drop in grades within the last year and who did not feel guilty or worthless within the last 2 weeks.

Group 5. Among the children who felt upset or alone and had no problem with their parents within the last 2 weeks, those who did experience a drop in grades within the last year and who did not feel guilty or worthless within the last 2 weeks.

Group 6. Among the children who felt themselves upset or alone and had no problem with their parents within the last 2 weeks, those who did not experience a drop in grades within the last year and who felt guilty or worthless within the last 2 weeks, and among these children, those who lost one of their close friends within the last year.

Group 7. Among the children who felt upset or alone and had no problem with their parents within the last 2 weeks, those who did experience a drop in grades within the last year and felt guilty or worthless within the last 2 weeks, and among these children, those who did not lose one of their close friends within the last year.

Group 8. Among the children who did not feel upset or alone and pessimistic within the last 2 weeks, those who experienced a drop in grades within the last year.

Group 9. Among the children who did not feel upset or alone and pessimistic within the last 2 weeks, those who did not experience a drop in grades within the last year and who did feel guilty or worthless within the last 2 weeks.

Group 10. Among the children who did not feel upset or alone and pessimistic within the last 2 weeks, those who did not experience a drop in grades within the last year and did not feel guilty or worthless within the last 2 weeks.

Group 11. Among the children who did not feel upset or alone and pessimistic within the last 2 weeks, those who had ever smoked and/or continued smoking.

Group 12. Among the children who did not feel upset or alone and pessimistic within the last 2 weeks, those who had never smoked.

The Beck depression score means for the above groups were compared by ANOVA. The results of this test indicated that the group means differences were highly significant ($P < 0.01$). To assess risk in the different groups, Tukey's multiple comparison test was used. The highest depression score mean was found for groups 1 (Mean ± SD; 18.48 ± 7.13) and 6 (Mean ± SD; 17.10 ± 7.20). However, the difference between the 2 group means was not statistically significant. Therefore, Group 1 and Group 6 were combined. It was observed that the

Beck depression scores for the children were significantly increased in the following cases: when the children felt upset or down during the last 2 weeks, when they had a problem with their parents, when their grades were low during the last year, when they felt guilty during the last 2 weeks, and when they had lost a close friend in the past year. The lowest average Beck depression score (CBDI), 7.33 ± 4.77, was obtained from Group 10. There were significant differences between the average of this group's score and the average score of the remaining 11 groups. When Group 10 was examined, it became apparent that this group of children did not feel upset or alone and pessimistic within the last 2 weeks, and also had not experienced a drop in grades within the last year, nor felt guilty or worthless within the last 2 weeks.

These conditions revealed that the most positive combination of risk factors were taken into consideration, *i.e.*, the depression scores would be expected to be low under those conditions.

The minimum depression score mean of the rest of Group 1, Group 6, and Group 10 was found to be 10.10 in Group 8, and the maximum depression score mean was 14.88 in Group 2; the difference between the 2 means was statistically significant. According to these results, it is likely that Group 2 was the group at highest risk. When the reason for this risk was investigated, it was determined that a parental problem was an important reason for risk among children who felt that they were upset or alone. The reason for risk in Group 8 could be explained by a drop in examination grades within the last year.

The following results were obtained after comparisons of the remaining 7 groups with each other were carried out. No meaningful difference was found among Group 3 (Mean = 13.11), Group 5 (Mean = 13.19), and Group 7 (Mean = 13.82) in terms of Beck depression points. Similarly, the difference between Group 4 (Mean = 10.41) and Group 12 (Mean = 10.10) was not statistically significant. However, the mean Beck depression scores of Groups 3, 5, and 7 were found to be higher than those of the Beck depression scores of Groups 4 and 12. Although the mean depression score point score of Group 9 (Mean = 11.22) was similar to those of the other 6 groups, the mean of Group 11 (Mean = 12.73) was higher than those of Groups 4 and 12.

Discussion

In this study, it was found that children or adolescents who had a problem with their parents, a reduction in grades in the past year, the loss of a close friend in the past year, and those who had smoked showed significant increases in their scores on the Beck Depression Inventory.

Some previous studies have shown that the loss of a parent or a loved one [5], the break-up of a romantic relationship [6], learning disorders [7], poor social competence [8], school failure, and family difficulties are indicated risk factors for depression in children/adolescents [9, 12]. In addition, cigarette smoking has been found to increase the risk of developing an episode of major depression [10–12]. In these studies, Pearson's chi-square test, logistic regression models, Pearson's correlation and linear regression analysis, or ANOVA type-models were employed as statistical methods. Either univariate models or multivariate models, including main effects without interactions, were used. In addition, in studies using logistic regression or chi-square analyses, depression point scores which are regarded as outcome variables, have been investigated after transforming the variables into a categorical structure with a suitable cut-off value. However, this process leads to a loss of information. In addition, one-way ANOVA models are commonly used, such as ANOVA-type models. Actual biological variation of the data is not examined accurately in such cases, since factorial ANOVA models are not applied. However, in cases involving a high number of variables, it is not possible to include all of the variables in the factorial ANOVA model or to determine interactions beforehand. The deficiencies of these models were avoided by use of the RTM and a schema which includes interactions between risk factors.

In addition, multiple regression models are frequently used as alternatives to the RTM, but single multiple regression models do not provide any means of grouping motifs that may work together. Regression trees separate independent variables that, together, change the dependent variable and create multiple groupings to explain the data. This feature of the RTM provides a distinct advantage over multiple regression methods [29].

In view of the present findings, the RTM appears to provide more reliable results in this context than do the traditional approaches; hence, this new statistical method has been widely used in recent years, especially in studies involving numerical-based outcome variables and for the investigation of a large number of variables. However, care should be taken to choose an optimal algorithm in order to determine the optimal tree structure, as well as to form and prune that structure.

References

1. Breiman L, Friedman JH, Olshen RA and Stone CJ: Classification and Regression Trees. Wadsworth, Belmont (1984) pp 358.
2. Wang Y and Witten IH: Inducing model trees for continuous classes. Proc of Poster Papers, 9 th European Conference on Machine Learning, Prague, Czech, April (1997).
3. Segal MR and Bloch DA: A comparison of estimated proportional hazards models and regression trees. Stat Med (1989) 8: 539–550.
4. Wells VE, Deykin EY and Klerman GL: Risk factors for depression in adolescence. Psychiatr Deve (1985) 3: 83–108.
5. Monroe SM, Rohde P, Seeley JR and Lewinsohn PM: Life events and depression in adolescence: relationship loss as a prospective risk factor for first onset of major depressive disorder. J Abnorm Psychol (1999) 108: 606–614.
6. Spencer T, Biederman J and Wilens T: Attention-deficit/hyperactivity disorder and comorbidity. Pediatr Clin North Am (1999) 46: 915–927.
7. Renouf AG, Kovacs M and Mukerji P: Relationship of depressive, conduct, and comorbid disorders and social functioning in childhood. J Am Acad Child Adolesc Psychiatry (1997) 36: 998–1004.
8. Reinherz HZ, Giaconia RM, Hauf AM, Wasserman MS and Paradis AD: General and specific childhood risk factors for depression and drug disorders by early adulthood. J Am Acad Child Adolesc Psychiatry (2000) 39: 223–231.
9. Brown RA, Lewinsohn PM, Seeley JR and Wagner EF: Cigarette smoking, major depression, and other psychiatric disorders among adolescents. J Am Acad Child Adolesc Psychiatry (1996) 35: 1602–1610.
10. Prochazka AV, Weaver MJ, Keller RT, Fryer GE, Licari PA and Lofaso D: A randomized trial of nortriptyline for smoking cessation. Arch Intern Med (1998) 158: 2035–2039.
11. Lewinsohn PM, Rohde P and Seeley JR: Major depressive disorder in older adolescents: prevalence, risk factors, and clinical implications. Clin Psychol Rev (1998) 18: 765–794.
12. Drapper N and Smith H: Applied Regression Analysis. 3 rd Ed, Wiley, New York (1998) pp 436.
13. Robnik-Sikonja M, Cukjati D and Kononenko I: Comprehensible evaluation of prognostic factors and prediction of wound healing. Artif Intell Med (2003) 29: 25–38.
14. White D and Sifneos JC: Regression tree cartography. Journal of Computational and Graphical Statistics (2002) 11: 600–614.
15. Clark LA and Pregibon D: Tree-based models; in Statistical Models in S, Chambers JM and Hastie TJ eds, Chapman and Hall, London (1992) pp 377–419.
16. Quinlan J: Induction of Decision Trees. Machine Learning, Kluwer Academic (1986) 1: 81–106.
17. Dykes JF: Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv, The Statistician (1998) 47: 485–497.
18. Unwin A and Hofmann H: New interactive graphics tools for exploratory analysis of spatial data; in Innovations in GIS *5*, Carver S ed, Taylor & Francis Ltd., London (1998) pp 46–55.
19. Kovacs: Rating scale to assess depression in school aged children. Acta Paedopsychiat (1981) 46: 305–315.
20. Talmon JL: A multiclass nonparametric partitioning algorithm. Pattern Recognition Letters, Mathematical Statistics (1986) 4: 31–38.
21. Cappelli C, Mola F and Siciliano R: A statistical approach to growing a reliable honest tree. Comput Stat & Data Anal (2002) 38: 285–299.
22. Bevilacqua M, Braglia M and Montanari R: The classification and regression tree approach to pump failure rate analysis. Reliab Eng Syst Safety (2003) 79: 59–67.
23. Karalic A: Linear regression in regression tree leaves; in International School for Synthesis of Expert Knowledge, Bled, Slovenia (1992).
24. Torgo L: A comparative study of reliable error estimators for pruning regression trees; in Proceeding of the Iberoamerican Conference on Artificial Intelligence, Coelho H ed, Springer-Verlag, Porto (1998).
25. Dietterich TG: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput (1998) 10: 1895–1923.
26. Kuhnerta PM, Dob KA and McClurec R: Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Comput Stat & Data Anal (2000) 34: 371–386.
27. Honeycutt E and Gibson G: Use of regression methods to identify motifs that modulate germline transcription in *Drosophila melanogaster.* Genet Res (2004) 83: 177–188.