Collaborative evaluation and management of students’ health-related physical fitness: applications of cluster analysis and the classification tree

Jou-An Chen¹, Chi-Chuan Shih², Pay-Fan Lin³, Jin-Jong Chen⁴ and Kuan-Chia Lin⁵,*

¹National Siluo Agriculturial Industrial High School, Yunlin, Taiwan
²National Nantou Commercial High School, Nantou, Taiwan
³School of Nursing, National Taipei University of Nursing and Health Science, Taipei, Taiwan
⁴Department of Physical Therapy and Assistive Technology, National Yang-Ming University, Taipei, Taiwan

Abstract

Health-related physical fitness has decreased with age; this is of immense concern to adolescents. School-based health intervention programs can be classified as either population-wide or high-risk approach. Although the population-wide and risk-based approaches adopt different healthcare angles, they all need to focus resources on risk evaluation. In this paper, we describe an exploratory application of cluster analysis and the tree model to collaborative evaluation of students’ health-related physical fitness from a high school sample in Taiwan (n=742). Cluster analysis show that physical fitness can be divided into relatively good, moderate and poor subgroups. There are significant differences in biochemical measurements among these three groups. For the tree model, we used 2004 school-year students as an experimental group and 2005 school-year students as a validation group. The results indicate that if sit-and-reach is shorter than 33 cm, BMI is >25.46 kg/m², and 1600 m run/walk is >534 s, the predicted probability for the number of metabolic risk factors ≥2 is 100% and the population is 41, both results are the highest. From the risk-based healthcare viewpoint, the cluster analysis can sort out students’ physical fitness data in a short time and then narrow down the scope to recognize the subgroups. A classification tree model specifically shows the discrimination paths between the measurements of physical fitness for metabolic risk and would be helpful for self-management or proper healthcare education targeting different groups. Applying both methods to specific adolescents’ health issues could provide different angles in planning health promotion projects.

Keywords: classification tree; cluster analysis; health-related physical fitness.

Introduction

Adolescent health problems, lack of exercise and poor physical capability not only impede adolescents physical and mental development but also their learning efficiency, living habits, interpersonal interaction and is an important issue of public health (1–5). Previous studies have shown that health-related physical fitness has decreased with age, this is of immense concern to adolescents; therefore, educating and promoting adolescents to keep a regular exercising habit during their school life is an important issue at present (6). School health and hygiene education has a long-term influence on adolescent physical fitness. Adolescence is an important period for physical development and it is the critical period before entering into adulthood, it is also the vital phase of physiological, psychological and social development in the life development cycle, which is an important stage for building ones self-concept and life style. Therefore, how to promote health-related physical fitness in school education has a profound effect.

School-based health intervention programs can be classified as either population-wide or high-risk approach (7). Population (or classroom-based) interventions include those programs delivered school-wide to all students, whereas high-risk programs target specifically those who experience certain conditions, e.g., are overweight or those who have poor physical fitness (7). Although the population-wide and risk-based approaches adopt different healthcare angles, they all need to focus resources on risk evaluation. As for school hygiene or health promotion plans, researchers usually use regression models to analyze risk factors in health related issues, such as student obesity or physical fitness, and to develop intervention programs to target adolescent health. Yet linear regression analysis usually has the following limitations (8–11). First, though the models can detect the interactions between independent and dependent variables, it is very difficult to interpret the results when examining the interactions among three or more variables simultaneously. Second, linear regression models cannot target desired subgroups. Health promotion interventions are designed based on results of population mean; therefore they usually fail to cater to the special needs of subgroups in the population and ignore the differences between groups. Third, regression models must be in accordance with the normal distribution hypothesis, also, independent and dependent variables must be in linear relation; prediction and
classification errors may occur if the two prerequisites are not complied with when applying regression models.

Academically, the application of cluster analysis and classification tree analysis has been emphasized in medicine and the field of hygiene. Based on previous studies and the literature, the application of cluster analysis is good for health behavior categorizing from the health promotion viewpoint (12–14) or when grouping the study of disease prevention strategies (14) depending on the various purposes of studies. The classification tree has been applied for discussing clinical mortality and morbidity and for presenting their clinical paths in recent years (9, 11, 15, 16). The tree model methodology builds a decision tree structure that classifies subjects into different risk groups and determines optimal cut-off points for a more efficient execution of the health plan. It also can effectively discriminate among the whole population and gradually show its significance in areas, such as reducing inequalities in health and health intervention (9, 11, 15, 16). However, to our knowledge thus far, the utilization of cluster analysis and classification trees in physical fitness or in school hygiene is less emphasized, particularly those that use regression trees as targeted interventions. Such information could be useful in further understanding and developing of health promotion programs tailored to deal with student’s health related issues.

Thus, this study is aimed at using different thinking and statistical methods on the issue of student’s health-related physical fitness in order to group data from the population and to test whether correlations exist between subgroups in biochemical measurements; thereafter to provide specific grouping discrimination paths and then analyze the differences in attributes and hidden meanings among groups. Front the results of the study we expect to provide an analytical method to understand the individual differences of students’ physical fitness data and eventually improve health promotion and management efficiency to meet this subpopulation’s needs.

Methods

Study subjects

The physical fitness data of all male freshmen enrolled in 2004 and 2005 school-years in a high school in Yulin country, Taiwan were collected as the study sample. A total of 380 students in 2005 school-year and 357 students in 2004 school-year were surveyed, they were all aged from 16 to 17 years. The data collection for blood biochemical parameters were performed by medical technologists of the responsible regional hospital. Blood samples were taken using the standard procedure. Three consecutive blood pressure readings at least 5 min apart were taken from the right arm with the person seated and diastolic blood pressure was measured at the 5th phase. Physical fitness tests were conducted by qualified physical education teachers based on standard procedure recommended by Ministry of Education.

Study variables

Four health-related fitness tests were then carried out during physical education lessons in school (17, 18). They were:

(a) Muscular strength and muscular endurance: muscle strength is the amount of force produced when a muscle is resisting an external force. Muscular endurance is the ability of a muscle to contract repeatedly or sustain a continuous contraction with less than maximum force. One minute sit-ups were used as the test item for this study.
(b) Flexibility: the ability of a joint to move through its range of motion, it is mainly affected by joints, muscles and tendons. Sit-and-reach was used as the test item for this study.
(c) Explosive power: the ability of a very short outburst of energy at a fast rate. Standing long jump was used as the test item for this study.
(d) Metabolic ability: the functions of heart, lung, and circulatory system of blood vessels. Sixteen-hundred meter running was used as the test item for this study.
(e) Body mass index (BMI): the measurement of obesity level of bodies. BMI = Mass (kg)/height (m²).

Biochemical measurements included systolic blood pressure (SBP), diastolic blood pressure (DBP), uric acid (UA) and total cholesterol. The hyperuricemia was defined as uric acid ≥7.0 mg/dL (19). Hypertension was defined if the average of the three readings was ≥140/90 mm Hg (20). The criterion for high cholesterol was 240 mg/dL according to the recommendation of the National Cholesterol Education Program (21). The numbers of selected metabolic risk factors ≥2 were then used as the dependent variables for the classification tree in order to understand the goodness of fit of the tree model construction.

Data analysis

SPSS 18.0 (SPSS, Inc., Chicago, IL) software was used to run the cluster analysis, next, Ward’s method of agglomerative was used to obtain the cluster number, and then we used Euclidean theory to detect the distance between observation samples, grouped the closest two into one group, and then cluster all the observation samples being tested into one big group. Furthermore, Ward’s method was applied to compute variance within group, the deviance variance magnitude worked as the basis for deciding a better quantity for clusters. Thereafter, we used the non-hierarchical K-means cluster method to run second phase analysis to generate the cluster numbers and recognize the attributes of each groups. Subsequently, we analyzed the group differences on biochemical measurements (uric acid, total cholesterol, SBP and DBP).

The classification trees in this study were run by S-Plus 6.2 (MathSoft, Inc., Seattle, WA) software, and the growing, stopping and pruning of the trees were determined by the Gini improvement measure. In the preliminary examination of the data, in order not to miss any possible key information and explanation of deviance, the stop condition sets the observed value of the minimum node and the minimum contact point as 10 and 5 and the deviance of the minimum node as 0.1 to allow complete growth of the trees. However, this results in extremely large and complex tree structures that are unexplainable in practical application. Therefore, minimum cost-complexity was applied to look for the most appropriate tree size. Additionally, the first cross-validation was conducted to randomly divide the data into a model-building set (90% of the study group) and a validation set (10% of the study group) (22). The model-building set was then used to establish a tree that was pruned using the validation set to achieve an estimation of the most appropriate tree through the minimum cost-complexity test. In the end, to confirm the stability and predictability, the 2004 school-year tree was regarded as experimental sample group while 2005 school-year tree served as the validation sample group (22).
the tree model was built from the experimental sample group, and then the second cross-validation was conducted using the validation sample group. The prediction accuracy was based on the five indices: (a) total correct rate, the ratio of number of samples classified correctly to total number of samples; (b) sensitivity, the ratio of actual positives which are correctly identified as such; (c) specificity, the ration of negatives which are correctly identified; (d) positive prediction value, the ratio of samples with positive traits who are correctly identified; (e) negative prediction value, the ratio of samples without positive traits who are correctly diagnosed.

Results

Table 1 presents the characteristics and distribution of fitness variables, 357 students in 2004 school-year and 380 students in 2005 school-year. Table 2 shows the attributes distribution of cluster analysis in both school-years. First, we checked the deviation changes of variance between groups and the scree plot, the populations of both years could be divided into three subgroups. In the 2004 school-year group, the first subgroup accounted for 43.42% (n=155), the z score of every fitness test was the best, thus this group was named the good subgroup. The second subgroup accounted for 34.47% (n=131), except for the sit-and-reach, every fitness test came out as moderate. The third subgroup accounted for 22.36% (n=85), except for the sit-and-reach which came out as medium, the rest of the test results were the worst, hence this group was named the poor subgroup. We further examined the 2005 school-year group; the traits of each subgroup were as follows. The first subgroup accounted for 43.16% (n=164), every fitness test appeared the best, thus this group was named the good subgroup. The second subgroup accounted for 34.47% (n=131), except for the sit-and-reach, every fitness test came out as moderate. The third subgroup accounted for 22.36% (n=85), except for the sit-and-reach which came out as medium, the rest of the test results were the worst, so this group was named the poor subgroup. From the data of the two school-years, though the number of people and the ratio difference, the group structures had consistency to a certain degree under cluster analysis. Based on the subgroup structures, we did further comparative analysis (Table 2) using biochemical tests (uric acid, cholesterol, SBP, DBP) and the results showed that there were significant differences in all biochemical measurements between the subgroups.

Figure 1 presents the results of the 2004 school-year classification tree analysis, according to the diagnosis of minimum cost-complexity. The optimal tree size for 2004 school-year fitness categorization consists of five terminal nodes. The combination paths and interrelationships are described as follows. The terminal node 3 represents when the sit-and-reach is shorter than 33 cm, BMI is >25.46 kg/m², and 1600 m run/walk is >534 s, the predicted probability for number of metabolic risk factors ≥2 is 100% and the population is 41. The terminal node 5 indicates when the sit-and-reach is longer than 33 cm but BMI is >23.34 and the 1 min sit-ups <36 times, the predicted probability for number of metabolic risk factors ≥2 is 77.4% and the population is 31. By contrast, the terminal node 4 shows that when the sit-and-reach is >33 cm and BMI is lower than 23.34, the predicted probability for number of metabolic risk factors <2 is 93.9% and the population

Table 1  Descriptive statistics of variables of health-related physical fitness.

<table>
<thead>
<tr>
<th>Variables</th>
<th>2004 school-year</th>
<th>2005 school-year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean±standard deviation (n=357)</td>
<td>Mean±standard deviation (n=380)</td>
</tr>
<tr>
<td>Body mass index, kg/m²</td>
<td>22.04±4.7</td>
<td>22.46±4.65</td>
</tr>
<tr>
<td>1600 m running, s</td>
<td>556.08±77.53</td>
<td>530.36±100.16</td>
</tr>
<tr>
<td>1 min sit-ups, times</td>
<td>38.48±7.05</td>
<td>36.59±7.23</td>
</tr>
<tr>
<td>Sit-and-reach, cm</td>
<td>32.43±6.47</td>
<td>31.31±7.24</td>
</tr>
<tr>
<td>Standing long jump, m</td>
<td>2.09±0.28</td>
<td>2.14±0.27</td>
</tr>
</tbody>
</table>

Table 2  (A) Cluster analysis results; (B) biochemical measurements among three subgroups.

<table>
<thead>
<tr>
<th>Variables</th>
<th>2004 school-year</th>
<th>2005 school-year</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Good) (n=155)</td>
<td>(Moderate) (n=139)</td>
<td>(Poor) (n=63)</td>
</tr>
<tr>
<td>Body mass index</td>
<td>-0.432</td>
<td>-0.271</td>
</tr>
<tr>
<td>1600 m running</td>
<td>0.162</td>
<td>0.420</td>
</tr>
<tr>
<td>1 min sit-ups</td>
<td>0.723</td>
<td>0.170</td>
</tr>
<tr>
<td>Sit-and-reach</td>
<td>0.749</td>
<td>-0.571</td>
</tr>
<tr>
<td>Standing long jump</td>
<td>0.041</td>
<td>-0.006</td>
</tr>
</tbody>
</table>

(A) Health-related physical fitness tests (z score)

(B) Biochemical measurements

| Uric acid, mg/dL                | 6.12             | 6.27             | 7.13             | 0.002            | 6.29             | 6.37             | 7.31             | <0.001           |
| Cholesterol, mg/dL              | 157.26           | 163.75           | 183.21           | <0.001           | 153.38           | 156.49           | 172.77           | <0.001           |
| SBP, mm Hg                      | 121.30           | 121.65           | 132.60           | <0.001           | 124.64           | 124.87           | 133.75           | 0.004            |
| DBP, mm Hg                      | 72.21            | 72.68            | 77.86            | <0.001           | 67.23            | 68.89            | 75.41            | 0.001            |
is 133. And when the sit-and-reach is >33 cm, BMI is >23.34 while 1 min sit-ups >36 times, this combination path leads to metabolic risk factors <2, is 100% and the population is 18 (terminal node 6). Finally, the node 1 also presents when the sit-and-reach is smaller than 33 cm but BMI is lower than 25.46, this combination path leads to metabolic risk factors <2, is 87.1% and the population is 117.

Table 3 is the result of cross-validation of the classification tree (the experimental sample group was used to build the tree and then the validation sample group was used to make cross confirmation). As a result, classification accuracy of the experimental sample group is 89.6% while that of the validation sample group is 83.4%; the categorization accuracy for both sample groups is approximate. Further, we compared sensitivity, positive predictive value, specificity, and negative predictive value, the results of the four predictive indices are as follow: the sensitivity of the validation group is 73.6%, specificity is 89.7%, positive predictive value is 80.8% and negative predictive value is 87.5%.

**Discussion**

The current study is regarding a practical analysis on student physical fitness utilizing two subject-oriented analytic methods. As a result, cluster analysis effectively divides the population into subgroups; while the classification tree establishes a simple, stable and precise physical fitness evaluation system which specifies the discrimination paths and interrelationships among different metabolic risk groups.

Health promotion is a fairly important section in hygiene and healthcare field in schools; when we can instill knowledge, concepts, and skills about healthcare and disease prevention into individuals’ life through facilitating health promotion. Traditionally, the variable-oriented statistic methods (e.g.,

![Figure 1](A classification tree for number of selected metabolic risk factors ≥ 2 among 357 male high school students in Yunlin County, Taiwan.)

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Results of cross-validation in classification tree.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Total correct</td>
</tr>
<tr>
<td>(1) Experimental sample group (2004 school-year)</td>
<td>89.6</td>
</tr>
<tr>
<td>(2) Validation sample group (2005 school-year)</td>
<td>83.4</td>
</tr>
</tbody>
</table>
regression models) are frequently used to identify the risk clues for designing health promotion interventions. The primary function of statistical regression models is to test the correlation between independent variables and dependent variables when interfering factors are controls. Regression models check the correlation between dependent variables and independent variables using the mean of the population, therefore the special needs of subgroups in the population are often neglected when interventions are developed based on the population mean generated from the regression models method (9, 10, 22, 23). However, though regression models can test the interactions between independent and dependent variables, it is difficult to interpret the results when interactions are assessed among three or more variables simultaneously (8, 9, 16).

The utilization of cluster analysis has already been used in various fields and studies. For example, it is used to sort health behaviors in the medical and health science fields, to classify species in the biology field, and to categorize customers according to consumer behaviors in the marketing field (12–14, 24, 25). In these studies, the database is grouped, named according to attributes of different clusters, and then is given customized marketing concepts and health intervention suggestions based on the characteristics and contexts of the clusters. By contrast, studies about the use of cluster analysis on the issue of health promotion are limited in students’ health-related physical fitness. After practical proving, our results of the cluster analysis from two school-year students are quite similar and it also shows obvious statistical differences between physical measurements comparisons. The outcome reflected that cluster analysis has a certain reliability in categorizing students’ health-related physical fitness. Besides, our findings of strong associations between different clusters (subgroups) and metabolic risk factors were in agreement with previous research evidence (3, 26–30). Therefore, from the risk-based interventional point of view, the cluster analysis that targets high-risk populations can sort students’ physical fitness data in a short time and then narrow down the scope to recognize subgroups and is familiar to school health managers who can incorporate the findings into their practice (7, 31, 32). It is also important to encourage healthy physical fitness in adolescents as it may play an important role in biochemical parameters associated with increased risk of chronic disease. In a review article of the major school-based obesity prevention programs, Resnicow concluded that high-risk, and to a lesser degree school-wide interventions, can significantly reduce the prevalence of pediatric obesity (7).

For decision tree analysis, we found that classification tree analysis is particularly suitable for describing the relationship between data and can indicate the whole spectrum in a student population. The point is to find a group of potentially high risk or relatively healthy subjects, and through the interpretation of visual charts find the combination path of physical fitness, rather than finding individual risk factors. This then advocates the self-management plan or proper healthcare education and includes activities targeting different groups (in relative healthy and unhealthy conditions), which would be an immense help for school health intervention programs (8–11, 15, 16, 22, 23). Providing the utilization of regression tree for all of the students may have greater benefit in the peer group effect, self-awareness and in avoiding the possibility of an intervention targeting only those at high risk. In fact, previous studies have demonstrated that the population or classroom-based intervention, targeting all students, are at least as effective in improving the CVD risk profile of children with multiple risk factors as the risk-based approach (7, 31, 32).

In addition, through learning and executing, we find the advantages of classification trees models. First, it is a non-parametric analytical method which can overlook the hypothesis requirements for the data function format. Second, classification trees finds the best cut-points when processing independent continuous variables, this feature concretes the statistical analysis of variables. In the current study, we found that sit-and-reach and BMI are the key discriminator factors of the metabolic risk. Then the sit-and-reach and BMI plus 1600 m run/walk and 1 min sit-ups are predictors which specifically show the discrimination paths and interactions between the risk and normal groups.

Interpreting the results of the classification trees, sit-and-reach result is with greatest variation, it is possibly because male students in this age stage are adventurous, and fond of speedy and exciting exercise; also male students are in the phase of peer relationship establishment, hence they like group activities and exercise that can express high school student vitality, such as baseball and basketball. Yet sit-and-reach is mainly designed to test flexibility which is the stretching exercise for muscles and joints, such as gymnastic exercises (17, 18, 33, 34), the nature of these exercises are more moderate, less cumulative, and can be carried out by individuals, therefore they are less attractive to high school and vocational school students. In addition, to flexibility, our study also found that the BMI is another key discriminator factor. This was consistent with evidence that higher BMI scores were generally associated with lower fitness (17, 35, 36). A healthier body composition in childhood and adolescence is associated with a healthier cardiovascular profile later in life (18). In fact, previous studies have also shown that for youth and adolescents, boys performed significantly poorer in sit-and-reach than girls. And BMI was only correlated with sit-and-reach tests in boys (36). It has been shown there were interactions between fatness and fitness on CVD risk factors in Asian youth (37). Therefore, based on the tree structure point of view, our study results may imply that the combination of poor flexibility and body composition in male adolescent may worsen and contribute to the metabolic risk.

Limitations

This report suffers from a number of limitations, which we will briefly discuss. First, in view of the limitations of biochemical measurements, these findings do not consider additional metabolic factors [i.e., glucose, high density lipoprotein (HDL-C)] known to influence the development of cardiovascular disease. Despite that, the current exploratory analyses do provide sufficient findings to discuss the importance of the application of cluster analysis and tree models to collaborative management in students’ health-
related physical fitness. Future studies may be improved by examining a comprehensive set of metabolic risk factors. Second, the data of standing long jump is not included in the parameters of constructing the tree; it does not mean this fitness test is not important; however the homogeneity of the test results among the students tested means that it was not chosen. Yet it may have various results under classification trees for different age-level students or other populations in a community, so it is worth further study. Our study results were applied specifically to a Chinese male student population, so studies in female or other ethnic populations are necessary.

**Implications for health professionals**

This study uses two subject-oriented analytical methods in practical analysis of student fitness issues, and the result shows that cluster analysis can sort and categorize data in a short time, consequently health professionals can decrease the scope to recognize subgroups at high risk when promoting health prevention projects in a target population; while the classification tree can build a simple, stable and precise fitness evaluation system to point out and advocate the discrimination paths and interrelationships between different groups specifically to certain health conditions which would be helpful for self-management or proper healthcare education targeting different groups. Applying both methods to specific adolescents’ health situations could provide health professionals with different angles and prospects to plan further health promotion projects.

**Acknowledgments**

We would like to recognize the utilization of the student fitness database of National Siluo Agricultural Industrial High School for this study, the findings of the research are merely for health promotion references, and do not represent the position of National Siluo Agricultural Industrial High School. This study was approved by the institutional review boards of Department of Nursing, National Taipei College of Nursing (ID: 95A054).

**References**


