Health Measurement Scales: Methodological Issues

Demosthenes Panagiotakos*

Department of Nutrition Science - Dietetics, Harokopio University, Athens, Greece / 70 Eleftheriou Venizelou Str., 17 671, Athens, Greece

Abstract: Health scales or indices are composite tools aiming to measure a variety of clinical conditions, behaviors, attitudes and beliefs that are difficult to be measured quantitatively. During the past years, these tools have been extensively used in cardiovascular disease prevention. The already proposed scales have shown good ability in assessing individual characteristics, but had moderate predictive ability in relation to the development of chronic diseases and various other health outcomes. In this review, methodological issues for the development of health scales are discussed. Specifically, the selection of the appropriate number of components, the selection of classes for each component, the use of weights of scale components and the role of intra- or inter-correlation between components are discussed. Based on the current literature the use of components with large number of classes, as well as the use of specific weights for each scale component and the low-to-moderate inter-correlation rate between the components, is suggested in order to increase the diagnostic accuracy of the tool.

Keywords: Scale, risk, health, cardiovascular.

INTRODUCTION

Measurements are essential components of scientific research, whether in the natural, social or health sciences [1]. Health scales (called also indices or scores) are composite tools aiming to measure a variety of conditions or characteristics that are, usually difficult to measure quantitatively. Specifically, in bio-medical sciences there is a variety of clinical conditions (e.g. severity of a disease, health-related quality of life, sense of pain), psychological behaviors (e.g. depression, anxiety, stress), dietary behaviors, as well as attitudes and beliefs that are very difficult to measure quantitatively and with accuracy, since there is no "mechanical" scale to be used. In an attempt to define and quantify such attributes, specific composite tools called "scales" have been proposed. Generally, a scale is a set of items (usually questions called components) that each express a different dimension of an attribute [1, 2]. These components in statistical science are considered as discrete or continuous random variables that are scored using, usually arbitrary rules and summed in order to develop a total score that describes the individuals' characteristics. For example, a scale for the assessment of depression is a combination of questions that are related to this disorder. According to the responses given by an individual, scores are assigned to each answer (e.g. score 0 for never having this symptom to score 5 for having this symptom every day). The total score assists in classifying an individual to none, moderate, or sever depressive symptoms. In addition, an index for dietary evaluation is a combination of questions regarding the frequency of consumption of several foods (e.g. cereals, fruits, vegetables, etc) [3-5]. According to the responses given and the rationale for healthy eating, scores are assigned to each response. Particularly, an increasing monotonic function could be assigned to the responses of eating fruits and vegetables that are essential for good health (i.e. 0 for never, 1 for rare to 5 for every day). On the other hand, a decreasing monotonic function could be assigned to the responses of eating meat and products that are not essential for health (i.e. 5 for never, 4 for rare to 0 for every day). The total score of the scale will give the information of the level of individual "healthy eating" behavior [5]. Finally, there are scales for the assessment of risk for a hard health outcome, like a cardiac event that are based on biochemical and clinical characteristics, like the Framingham Heart Study risk charts or the European Society of Cardiology SCORE project [6, 7].

The use of composite scales is not only attractive, but also mandatory in order to address problems in statistical analysis and inferences caused by the synergistic effects or interaction between several characteristics, which express different dimensions of an attribute. For example, it has been strongly suggested that food consumption may act interdependently in the development of a chronic disease, like a dietary pattern that includes reduced fat intake which is usually correlated with increased antioxidant consumption. Thus, it is not clear whether fats or vitamins play a role in the prevention of the disease. As a result, a "single indicator" analysis is considered inadequate to evaluate possible effect modification among isolated components of the scale, as well as confounding by the effect of other variables. Furthermore, entering together in a statistical model a number of highly correlated variables (i.e. usually the individual components of a scale) may lead to the co-linearity phenomenon,

*Address correspondence to this author at the Department of Nutrition Science - Dietetics, Harokopio University, Athens, Greece / 70 Eleftheriou Venizelou Str., 17 671, Athens, Greece; Tel: +30-210-9549332; Fax: +30-210-9600719; E-mail: dbpanag@hua.gr

1874-1924/09

resulting in less robust estimations of the coefficients and less accurate predictions [8].

All these problems can be addressed using scales that can measure complex concepts, more effectively than single indicators and are more capable in handling multiple items. Moreover, use of scales capture the extremes in attitudes, behaviours, pre-empts confounding, and possible effect modification among individual variables through the same scaling procedure and they do not tend to be biased. However, discussions regarding issues of measurement were noticeably absent in medical research [1]. Particularly, there are several unresolved issues regarding: (a) the use of the appropriate scoring system (i.e. monotonic or not-monotonic, with small or large range), (b) the use of weights in a scale components, (c) the level of inter-correlation between the components of a scale, and, (d) the number of components used for developing a scale. All the aforementioned issues may play an important role in improving accuracy of the scales and they are discussed here, with examples from clinical or theoretical practice.

METHODOLOGICAL ISSUES IN DEVELOPING HEALTH MEASUREMENT SCALES

As mentioned above the use of the appropriate scoring system (i.e. monotonic or not-monotonic, small or large range), the use of weights in a scale's components, the level of inter-correlation between the components and the optimal number of components used for developing a scale, are discussed. Particularly, the use of scoring system is of major importance since some behaviors are not linearly related to a health outcome; like alcohol drinking and cardiovascular disease risk where a parabolic trend has been consistently reported in the literature. Thus, the use of non-monotonic scoring is considered essential to better evaluate the role of this particular behavior on the investigated outcome. Moreover, in a typical Likert type scoring system (i.e., strongly disagree, disagree, neither disagree nor agree, agree and strongly agree) some of the potential answers may not have the same impact to the outcome as others. Thus, the use of weights in a particular answer based on the existence knowledge may improve the diagnostic ability of the tool. Similarly, in composite tools, some components may have greater impact of the investigated outcome compared with others; thus, assigning weights in these components may influence the accuracy of the composite scale in predicting the outcome. Moreover, many of the scale items may have a level of inter-correlation since they aim to evaluate the same characteristic. The level of inter-correlation or the number of items used may also have a significant impact in developing a scale and on its accuracy in predicting the outcome that it has been designed to do.

The use of the Appropriate Scoring System

Based on an extensive literature search, the majority of scales that have been constructed using 1, 2 or 3 thresholds for each component and assigning scores of 1, 2, 3 when the attitude is towards or not a healthy behavior (or vice versa). In a recent publication [9] we investigated whether the number of classes of scale components influences the diagnostic accuracy of the tool. The accuracy was measured through the area under the Receiver Operating Characteristics (ROC) curve (i.e. the Area Under Curve, AUC). Particularly, based on simulated data, a scale with infinity number of components ($I\infty$) from the Normal distribution was initially developed. Afterwards, 10 other new scales were developed from the previous one using 100-quantiles, 50-quantiles, 15-quantiles, 10-quantiles, 8-quantiles, 6-quantiles, quintiles, quartiles, tertiles and the median of $I\infty$ scale. A positive association between the number of classes and the diagnostic accuracy of the scales (measured using the sensitivity and AUC) was observed (Fig. **1**). This finding was also confirmed by the choice of natural logarithm as the function that best describes the relationship between the number of classes of each scale component and the measures of diagnostic accuracy of the tool. The previous findings lead to the conclusion that the maximum diagnostic accuracy of a scale is achieved when the maximum number of classes in each component is used.

The latter association remained stable when some components were un-correlated with the outcome, while it was not detected when all index components were un-correlated with the outcome.

The finding of the aforementioned work is important for health practice and research, since it strongly suggests using as many classes as it can be used for a scale's components. For example, in order to measure the level of anxiety (a potential cardiovascular risk factor) one may use a scale with many possible responses (like not at all, 3-5 per month, 1-2 per week, daily etc) instead of using a simple "yes/no" answer. Taking also into account that most clinicians wish to have a diagnostic tool that accurately predicts the truly diseased individuals, this conclusion may substantially reduce the level of misclassification in screening procedures for high risk people. However, one may claim that scales constructed using small number of classes may be more comprehensive and easier applied in daily clinical practice. As discussed above they result in a tool with low diagnostic accuracy. The latter could be an explanation for the lack of significant findings from some studies that used small-range scales [10, 11]. Moreover, one may also argue that large-range components may lead to lack of reproducibility, which is of importance for assessing etiology in medical research. The latter consideration is still under investigation.

Nevertheless, it should be mentioned that the use of continuous instead of discrete components in order to develop a composite scale is always preferable. However, in many cases the use of continuous components is not feasible for various practical reasons, like the inability of a person to quantitatively measure a feeling or behaviour.

The findings discussed suggest the use of large-range components. These considerations are of clinical as well as methodological importance since their application in practice will provide researchers with a methodological framework to develop more accurate predicting tools of a health related outcome.

The use of Weights

The contribution of each component to the total score of a composite scale is a major, unresolved issue. The majority of the presented scales have been developed assigning the same weight to all components. However, it is widely ac-

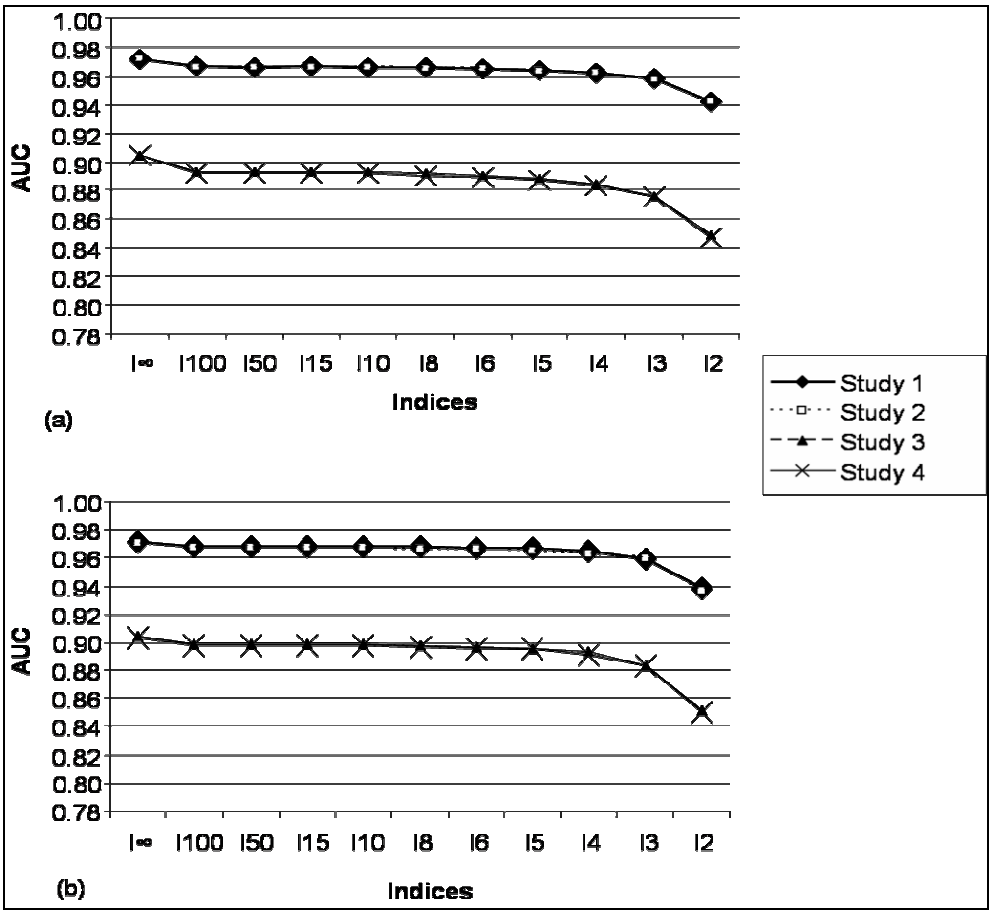

Fig. (1). Area under the ROC curve of several scales, with diseased to non-diseased ratio is 1:3 (**a**) or 1:1 (**b**) (adopted from Kourlaba G and Panagiotakos DB, 2009 (9)).
Study 1: The sample size of each simulated data set was 1000 and each scale was constructed summing 10 components.
Study 2: The sample size of each simulated data set was 100 and each scale was constructed summing 10 components.
Study 3: The sample size of each simulated data set was 1000 and each scale was constructed summing 5 components.
Study 4: The sample size of each simulated data set was 100 and each scale was constructed summing 5 components.

cepted that in the majority of the scales the components do not have the same relationship to a health outcome. For example, consider a composite scale that aims to evaluate the risk for developing cardiovascular disease, which is based on arterial blood pressure, lipids and glucose levels. It is rational to believe, based on the evidence, that all the aforementioned components do not have the same effect on the development of cardiovascular disease. Moreover, several studies examining the relationship between dietary scales and disease outcomes have reported no or moderate associations [12, 13]. The latter could also be a result of the use of un-weighted scales to evaluate the food-health relationships. One may claim that using unequally distributed scores of a scale's components is an indirect weighting method. Specifically, some investigators have suggested using different scaling distributions between the components (e.g. in the example given before one may assign score 1 in those who have hypertension, or abnormal lipids levels, but score 2 for those reporting diabetes, since the later has stronger associations with the disease; in those without these conditions score 0 would be assigned). However, the previous "weighting" system is arbitrary, lacks of scientific basis and therefore, cannot be provided as a methodological guideline.

In a very recent work, 3 weighting methods for scale components were proposed and tested using both simulated and empirical data [14]. In particular, the odds ratios obtained from unadjusted logistic regression that evaluated the effect of individual scale components on a hypothesized health outcome were suggested as specific weights. Although the use of these specific weights was associated with improvements in the diagnostic ability of the scales, this method shares some important limitations. Specifically, the suggested weights are based on the relative relationship of each scale's item with a particular, simulated outcome. Therefore, even for the same component of a scale, different weights could be proposed when applying the scale to different health outcomes. In addition, the influence of the intercorrelations and synergistic effect of the scale items, which is common in real practice, are ignored using the odds ratios obtained from unadjusted models as potential weights. Thus, the use of odds ratios obtained from multiple logistic regression models may improve the accuracy of the scale.

Then, and under the concept of the same work, the weights of the scales' components were derived based on odds ratios obtained from multiple logistic regression models

when each index component was entered as independent variable and a total score of the rest of components (i.e. summation of the initial scores of these components) was entered in the model. The latter was made to control for the potential confounding effect of the other components. It should be noted that a score of the components was preferred instead of all components separately, because it is very common that scale items are highly correlated, since they aimed to describe the same characteristic. Afterwards, another scoring procedure was also suggested. Particularly, the weights of scale components were obtained by multiplying the aforementioned weights with the Deviance, which is a measure of the importance of the component in predicting outcome.

Finally, weights were suggested by multiplying the weights obtained from the odds ratios with the inverse of the variance of the specific odds ratio, which represents the effect size of the association. The current findings highlight that the predictive capacity of the weighted indices constructed using the weights mentioned above is higher compared to that of the un-weighted index (Fig. **2**). However, no meaningful differences were observed between the different scoring procedures suggested in the referenced work. The previous findings were confirmed when empirical data were used from an epidemiological study [14].

Although the use of the suggested weights seems to improve the diagnostic accuracy of health measurement scales, some important issues of using them should be further discussed. At first, the proposed weights are based on the relative relationship of each scale item with the specific outcome. Thus, different weights should be proposed for different health outcomes for the same components. For example, let assume a scale that is based on food group consumption. Based on the literature all foods included influence both the development of hyperlipidemia, hypertension and diabetes; however, they do not have the same impact on several health outcomes, such as coronary heart disease or stroke. Therefore, different weights should be used for specific foods according to their relationship with coronary disease or stroke. In other words, several food-based scales should be developed based on what the scale intends to predict. Another limitation of the aforementioned methods of scoring is that specific weights for scale components should be given based on specific datasets. Thus, it is not certain that weights derived from 1 dataset will improve the predictive ability of corresponding indices developed from another dataset. In other words the issue of lack of generalization arises. A solution to the aforementioned consideration is using the effect size estimates (i.e., odds ratios) from available metanalyses of studies investigating the same topic.

Thus, despite the clarity and consistency of the weights presented in previous works, further research is needed in order to examine whether other effect size measures may be used as weights of a scale components.

The Number of Components and the Level of Inter-correlation between the Components

In the above paragraphs the range of values and the use of weights of the scale components were discussed. However, the number of components used or the correlation structure of the components remains an important, undiscovered topic for the development of an accurate health measurement scale. In one of our recent studies in this field we showed that the diagnostic accuracy of a scale increases as the number of components increases, if and only if all components are related to the outcome that the scale intends to evaluate [15]. Moreover, the diagnostic accuracy is improved when non- or low inter-correlated components are used.

Specifically, based on simulated data we showed that the diagnostic accuracy of a scale developed using non- or low inter-correlated components is higher compared with that of a scale that was developed using moderate to high inter-correlated components. The latter reveal the issue of the explained variability of the components that are high correlated. In addition to the previous findings, the diagnostic accuracy of a scale increases as the number of components used increases, too. Moreover, all scale components should be related to the outcome that the scale aims to predict, since the simulated and empirical data of this work showed that the diagnostic accuracy of a scale developed using both related and un-related to a specific outcome components is lower compared with a scale developed using only the com-

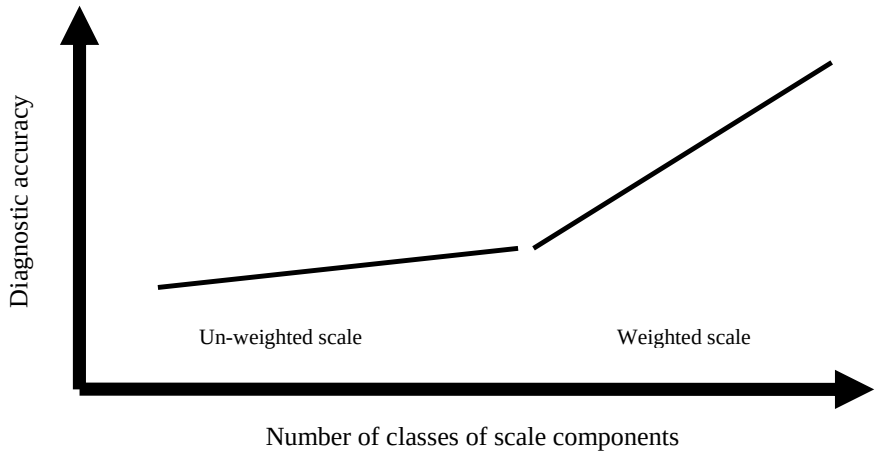

Fig. (2). Use of weights in a scale components and its diagnostic accuracy.

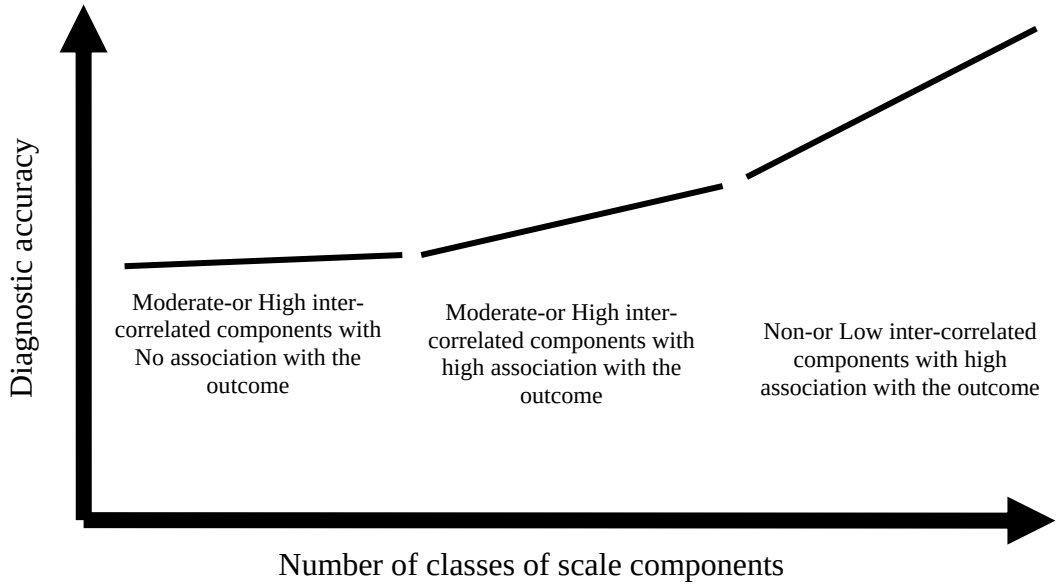

Fig. (3). The role of correlation structure in a scale components and its diagnostic accuracy.

ponents correlated with the outcome. This could be an explanation why several health scale measurements, especially from the nutrition epidemiology field, have failed to observe significant associations with health outcomes [13]. The majority of these dietary scales have developed based on general and not disease-specific dietary recommendations. Furthermore, in the aforementioned work we suggested that the diagnostic accuracy of a single component is higher than of the accuracy of a scale that has been developed using inter-correlated components and with only some of them associated with the outcome.

Finally, the correlation structure of the components does not affect the already reported association between the number of classes and the diagnostic accuracy of the scale. (Fig. **3**). In particular, it was observed that the diagnostic accuracy increases as the number of classes increases, irrespective of the correlation structure of the components. All the previous findings were also confirmed when empirical data were used [15].

The aforementioned findings, suggest the use of low- or non- inter-correlated components with a high level of association with the investigated outcome, in order to obtain an accurate scale. The number of components used for the development of a scale is important only when low- or non- inter-correlated components are used; in this case the diagnostic accuracy increases as the number of components increases.

CONCLUSIONS

Health measurement scales are important tools in evaluating an individuals characteristics that cannot be measured directly. During the past years health scales have become firmly established as a routine part of evaluating interventions and in planning health care [16]. However, although a large number of scales have been proposed and are widely used in scientific research several methodological issues have not been entirely appreciated and understood. These issues could be possible explanations for the lack of associations of the existing scales with various chronic diseases in the vast majority of populations that they have been applied (e.g. adults, post-menopausal women, elderly, etc). During the past few years, some newer and more sophisticated techniques for scale development have been suggested [16]. In the present review, some issues regarding the development of a more accurate scale were presented and discussed. In conclusion, the use of components consisted of large number of classes, as well as the use of specific weights for each scale component, and the low-to-moderate inter-correlation rate between the components, is evident from our simulated and empirical studies. Nevertheless, further work is needed, regarding the repertoire of health measurement scales, including the replacement of some outdated methods with newer and more accurate ones.

ACKNOWLEDGEMENT

I would like to acknowledge the contribution of Georgia Kourlaba, PhD student in my Department, for helping me in exploring this scientific field.

REFERENCES

[1] Streiner DL. Norman GF. Introduction. Health Measurement Scales, 4th ed. USA: Oxford University Press USA 2008; pp. 1-4.

[2] Kant AK. Indexes of overall diet quality: a review. J Am Diet Assoc 1996; 96: 785-91.

[3] Trichopoulou A, Costacou T, Bamia C, Trichopoulos D. Adherence to a mediterranean survival in a greek population. N Engl J Med 2003; 348: 2599-608.

[4] Martinez-Gonzalez MA, Fermandez-Jarne E, Serrano-Martinez M, Wright M, Gomez-Gracia E. Development of a short dietary intake questionnaire for the quantitative estimation of adherence to a cardioprotective Mediterranean diet. Eur J Clin Nutr 2004; 58: 1550-2.

[5] Panagiotakos DB, Pitsavos C, Stefanadis C. Dietary patterns: a Mediterranean diet score and its relation to cardiovascular disease risk, clinical and biological markers. Nutr Metab Cardiovase Dis 2006; 16: 559-68

[6] Kannel WB, McGee DL, Gordon T. A general cardiovascular risk profile: the Framingham study. Am J Cardiol 1976; 38: 46-51.

[7] Conroy RM, Pyorala K, Fitzgerald AP. SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003; 24: 987-1003.

[8] Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidiol 2002; 13: 3-9.

[9] Kourlaba G, Panagiotakos DB. The diagnostic accuracy of a composite index increases as the number of partitions of the components increases and when specific weights are assigned to each component. J Appl Stat 2009; (in press).

[10] Patterson RE, Haines PS, Popkin BM. Diet quality index: capturing a multidimensional behavior. J Am Diet Assoc 1994; 94: 57-64.

[11] Kennedy ET, Ohls J, Carlson S, Fleming K. The healthy eating index: design and applications. J Am Diet Assoc 1995; 95: 1103-8.

[12] Harnack L, Nicodemus K, Jacobs DR, Jr., Folsom AR. An evaluation of the dietary guidelines for Americans in relation to cancer occurrence. Am J Clin Nutr 2002; 76: 889-96.

[13] McCullough ML, Feskanich D, Stampfer MJ, *et al*. Adherence to the dietary guidelines for Americans and risk of major chronic disease in women. Am J Clin Nutr 2000; 72: 1214-22.

[14] Kourlaba G, Panagiotakos D. Use of weights in the items of a composite index increases the accuracy in prediction: an application to diet and health-related outcome. Proc Hellen Stat Inst 2007.

[15] Kourlaba G, Panagiotakos DB. The number of index components affects the diagnostic accuracy of a diet quality index: the role of intra- and inter-correlation structure of the components. Ann Epidemiol 2009; 19: 692-700.

[16] McDowell I. Health Measurement Scales. Encyclopedia of Public Health. The Gale Group Inc. 2002. [Accessed on July 24, 2009]. Available from: http://www.encyclopedia.com/doc/1G2-3404000406.html

Received: October 09, 2009 Revised: October 23, 2009 Accepted: October 26, 2009