Journal of Metabolic Syndrome

ISSN: 2167-0943

Research Article Open Access

Development and Validation of Metabolic Syndrome Prediction and Classification-Pathways using Decision Trees

Brian Miller[1]* and Mark Fridline[2]

[1]School of Sport Science & Wellness Education, The University of Akron, Akron, OH; Doctoral Student, Health Education and Promotion, School of Health Sciences, Kent State University, Kent, OH, USA
[2]Department of Statistics, The University of Akron, Akron, OH, USA

Abstract

Purpose: The purpose of the current investigation was to create, compare, and validate sex-specific decision tree models to classify metabolic syndrome.

Methods: Sex-specific Chi-Squared Automatic Interaction Detection, Exhaustive Chi-Squared Automatic Interaction Detection, and Classification and Regression Tree algorithms were run in duplicate using metabolic syndrome classification criteria, subject characteristics, and cardiovascular predictor variable from the National Health and Nutrition Examination Survey cohort data. Data from 1999-2012 were used (n=10,639; 1999-2010 cohorts for model creation and 2011-2012 cohort for model validation). Metabolic Syndrome was classified as the presence of 3 of 5 American Heart Association National Heart Lung and Blood Institute Metabolic Syndrome classification criteria. The first run was made with all predictor variables and the second run was made excluding metabolic syndrome classification predictor variables. Given that the included decision tree algorithms are non-parametric procedures, all decision tree models were compared to a logistic regression based model to provide a parametric comparison.

Results: The Classification and Regression Tree algorithm outperformed all other decision tree models and logistic regression with a specificity of 0.908 and 0.952, sensitivity of 0.896 and 0.848, and misclassification error of 0.096 and 0.080 for males and females, respectively. Only one predictor variable outside of the metabolic syndrome classification reached significance in the female model (age). All metabolic syndrome classification predictor variables reached significance in the male model. Waist circumference did not reach significance in the female model. Within each model, 5 female and 3 male pathways built off of <3 American Heart Association National Heart Lung and Blood Institute Metabolic Syndrome classification criteria resulted in an increased likelihood of presenting Metabolic Syndrome.

Conclusion: The proposed pathways show promise over other current metabolic syndrome classification models in identifying Metabolic Syndrome with <3 predictor variables, before current classification criteria.

Keywords: Cardiovascular disease; Metabolic syndrome; Decision trees; Diabetes; NHANES

Introduction

Metabolic syndrome (MetS) is a constellation of cardiometabolic predictor variables that when presented in tandem increases the risk of cardiovascular disease (CVD) and insulin resistance [1,2]. The prevalence of this classification affects approximately 1 in 3 adults in the United States [3]. Due to the high prevalence of this syndrome, proper identification of persons with MetS is imperative in order to prevent and/or modify the multiple predictor variables associated with CVD related morbidity and mortality as well as its high healthcare costs [1,2,4,5]. Furthermore, utilization of pathways for MetS classification could guide health education professional interventions before the onset of related morbidity and mortality. Using Decision Trees (DT) as a preliminary pre-metabolic syndrome classification criterion could improve outcomes associated with the development of MetS or could halt the progression of MetS and its relative consequences [6].

Classification of metabolic syndrome

Although there have been numerous attempts to harmonize classification models for MetS, there remains a lack of consensus amongst the leading organizations with particular disagreement based on predictor variable cut-off points as well as which predictor variables should be considered in making the MetS classification [1,7-9]. More recently, there has been support for MetS to be considered as a pre-morbid condition intended to inform health educators and clinicians on relative risk of developing CVD rather than a clinical diagnosis [6,10]. In lieu of a clinical diagnosis, MetS can provide a research framework for establishing a unified cardiometabolic pathophysiology, quantifying chronic disease risk, guiding clinical management decisions, and providing a concise methodology to inform public health and health education professionals of the relationship of clustering predictor variables [10].

Classification criteria based on the leading models from the national cholesterol

Education Adult Treatment Panel III (ATPIII), the International Diabetes Federation (IDF), the World Health Organization (WHO), and the American Heart Association National Heart Lung and Blood Heart Institute (AHA/NHLBI) risk models are limited in their usefulness because they classify MetS based on predictors with binary thresholds [1,2]. There currently exists limited evidence-based research that considers the severity of these MetS cardiometabolic predictor variables, their interactions with one another, and their relationship

***Corresponding author:** Brian Miller, School of Sport Science & Wellness Education, InfoCision Stadium 317, The University of Akron, Akron, OH, 44325-5103, USA, Tel: (216) 659-6985; E-mail: bm25@zips.uakron.edu

Received December 16, 2014; **Accepted** January 05, 2015; **Published** January 10, 2015

Citation: Miller B, Fridline M (2015) Development and Validation of Metabolic Syndrome Prediction and Classification-Pathways using Decision Trees. J Metabolic Synd 4: 173. doi:10.4172/2167-0943.1000173

to CVD. A major limitation within these models is the dichotomous nature of predictor variable identification [6,10]. However much like obesity, there are varied clinical implications based on the severity of predictor variables used to define MetS where the dichotomized cut-off points for each predictor variable might be clinically ambiguous. Furthermore, current MetS classification models lack consideration for established CVD predictor variables such as patient demographics (i.e. race/ethnicity and socioeconomic status), smoking [3] and previous cardiovascular events [11]. The creation of clinically feasible pathways for MetS classification that both stratifies each predictor variable based on its severity and then considers the interaction effect as predictor variable clusters could be invaluable for reducing risk of cardiovascular morbidity and mortality [5].

Decision trees

DT methodologies have been shown to be effective tools for the classification and prediction of cardiometabolic chronic disease such as MetS and insulin resistance [6,12-14]. However, with the exception of Miller, Fridline, Liu & Marino and Stern et al. other models have been based on international samples. To the best of our knowledge, no published pathways for MetS classification derived from DT methodologies have been built, validated, and implemented in clinical practice [6,14].

DTs are powerful classification and prediction techniques that analyze how both categorical and continuous predictor variables best combine to create pathways explaining the outcome of a given binary response variable according to statistical tests in tandem with "if-then" logic [6,14,15]. In DT algorithms, the data set is partitioned into two or more mutually exclusive subsets in each split with the goal of producing subsets of the data which are as homogeneous as possible with respect to the response variable. This nonparametric modeling technique shows promise over traditional regression techniques in that DT's make no assumptions about the underlying data including mutlicollinearity, are able to handle missing variables, are easily interpreted by non-statisticians, and consider the effects of variable clusters in relation to sample subsets unlike regression which considers the effect of each variable within the entire sample.

Chi-squared automatic interaction detection

The Chi-Squared Automatic Interaction Detection (CHAID) algorithm proposed by Kass operates using a series of merging, splitting, and stopping steps based on user-specified criteria as follows [16]. The merging step operates using each predictor variable where CHAID merges non-significant categories using the following algorithm: (1) Perform cross-tabulation of the predictor variable with the binary target variable. (2) If the predictor variable has only 2 categories, go to step 6. (3) $\chi 2$ -test for independence is performed for each pair of categories of the predictor variable in relation to the binary target variable using the $\chi 2$ distribution (*df*=1) with significance (α_{merge}) set at 0.05. For nonsignificant outcomes, those paired categories are merged. (4) For nonsignificant tests identified by α_{merge} >0.05, those paired categories are merged into a single category. For tests reaching significance identified by $\alpha_{merge} \leq 0.05$, the pairs are not merged. (5) If any category has less than the user-specified minimum subset size, that pair is merged with the most similar other category. (6) The p-values for the merged categories are adjusted using a Bonferroni correction to control for Type I error rate.

The splitting step occurs following the determination of all the possible merges for each predictor variable. This step selects which predictor is to be used to "best" split the node using the following algorithm: (1) $\chi 2$-test for independence using an adjusted p-value for each predictor. (2)The predictor with the smallest adjusted p-value (i.e., most statistically significant) is split if the p-value less than the user-specified significance split level ($\alpha split$) is set at 0.05; otherwise the node is not split and is then considered a terminal node.

The stopping step utilizes the following user-specified stopping rules to check if the tree growing process should stop: (1) If the current tree reached the maximum tree depth level, the tree process stops. (2) If the size of a node is less than the user-specified minimum node size, the node will not be split. (3) If the split of a node results in a child node whose node size is less than the user-specified minimum child node size value, the node will not be split. The parent node is the level where the data set divides into child nodes that can themselves become either parent nodes or end in a terminal or decision node. (4) The CHAID algorithm will continue until all the stopping rules are met.

Exhaustive CHAID (E-CHAID) proposed by Biggs, DeVille, and Suen uses the basic CHAID algorithm with more computationally intensive merging and testing of response variables [17]. In the E-CHAID algorithm, there is no reference to any α_{merge} value. Rather category merging continues until only two categories remain. Therefore, careful considerations should be made for over-fitting when the E-CHAID algorithm is used for large data sets with large amounts of continuous predictor variables.

Classification and regression trees

Unlike CHAID based algorithms, the Classification and Regression Tree (CART) algorithm proposed by Breiman, Freidman, Stone, and Olshen builds purely binary trees [18]. Therefore, CART pathways are easier to understand as parent nodes are always split into 2 child nodes that partition data to maximize homogeneity of each subset. In the CART procedure, the maximum tree is produced followed by tree pruning to avoid over-fitting.

The first step in the tree growing process is to find each predictor variables best split. In the CART algorithm, the splitting step employs a statistical calculation known as the Gini Impurity Function. This function is a measure of how often a randomly selected case would be incorrectly predicted; therefore it is used to determine the optimal binary split of the parent node into the child nodes. In the next step when the stopping rules are satisfied, the best possible split is chosen for the predictor variable when the impurity decreases the most from the parent node to the child nodes. This impurity decrease is quantified by the Gini Improvement Measure, which measures the decrease in impurity from the parent node to the child node. The parent node will be split when the change in impurity is maximized.

Logistic regression

Logistic Regression (LR) is a widely utilized statistical technique in binary response prediction [19]. However, LR output can be tedious to interpret and requires considerations for mutlicollinearity and missing values. These models are used when the response variable (y) is binary with the response variable taking the value of 1 with probability of success π or the value of 0 with probability of failure $1 - \pi$, and the predictor variables (xi) are either categorical or continuous values represented by the following equation:

$$ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = ln\left[\frac{P(y_i = 1/x_i)}{P(y_i = 0/x_i)}\right] = \beta + \sum_{i=1}^{p}\beta_i x_i$$

Where $\beta 0$ is a constant and βi are the coefficients of the predictor variables in the model. The LR equation, called the likelihood function, is used for estimating the regression model coefficients. The maximum likelihood estimation method uses an iterative procedure to find the model coefficients that best match the pattern of observations in the sample data. Interpretation of the model comes from transforming the LR coefficients for each predictor variable by taking the exponential of the coefficients $(e\beta i)$ to determine the influences of each predictor variable on the response variable in terms of the odds ratio. To determine if each model coefficient is statistically significant, the Wald statistic is used.

Purpose

The central hypothesis states that the decision tree pathways derived from DT algorithms using data from National Health and Nutrition Examination Survey (NHANES) cohorts would detect the presence of MetS in adults with <3 AHA/NHLBI MetS predictor variables. The current investigation had two aims. The first aim was to develop and validate sex-specific pathways for MetS classification using multiple DT derived methodologies. The second aim was to compare each DT model with and without MetS classification criteria.

Materials and Methods

Data management

The study sample was derived from National Health and Nutrition Examination Survey (NHANES) data made publically available by the Centers for Disease Control and Prevention (CDC). This included 7 cohorts from 1999-2012 collected in 2-year intervals. The data was arranged in a column-wise format with each subject given a sequence identifier. Data management was performed with dataset merging and data subset functions using SPSS version 22 (SPSS Inc., Chicago, IL). The final sample size for inclusion in model development was n=10,639 (male: n=5,474; female: n=5,165). The current investigation was approved by the Institutional Review Board.

The inclusion criteria were based on the following parameters: Age range of 18-59 years, 12 hour fasting protocol for laboratory values, abstinence from alcohol and/or tobacco use prior to laboratories, and a negative exam for pregnancy for females. The age criteria was chosen based on Ford, Li, and Zhao [3] where the highest prevalence of MetS was exhibited after 59 years of age. This decision was made in order to create pathways to detect MetS before onset of MetS with traditional classification criteria based on the high prevalence of MetS beyond age 59. Participants with missing data based on the MetS classification criteria were excluded due to the inability/uncertainty in making a complete MetS classification. The 1999-2010 cohorts were reserved for model creation (training) and the 2011-2012 cohort was reserved for model validation. Both of the training and validation sets were separated by sex. The distributions of all parameters were the same between training and validation sets. Blood pressure readings were the average of 4 blood pressure collections per subject. An indicator of cardiovascular events was built off of the presence of 1 of 5 cardiovascular events including congestive heart failure, coronary heart disease, angina, heart attack, and/or stroke.

Metabolic syndrome classification

The MetS classification was defined as the presence of 3 of 5 predictor variables based on the clinical classification model proposed by the AHA/NHLBI, see Table 1 [1].

Statistical analysis

The DT models were developed using CHAID, E-CHAID, and CART algorithm analysis using SPSS version 22 (SPSS Inc., Chicago, IL). Each DT analysis was run in duplicate with parent nodes defined at 250 subjects, child node defined at 100 subjects, and significance for all statistical tests within each DT set at ≤ 0.05. Maximum tree depth was user specified at 5 levels. The NHANES cohort data was divided by sex to create sex-specific models for MetS classification with the 2011-2012 cohort reserved for model validation. Each DT algorithm was run twice with the first model including all possible predictor variables and the second without any AHA/NHLBI MetS classification criteria. Predictor variables included the AHA/NHLBI MetS classification criteria in addition to binary smoking status, American Heart Association Blood Pressure Classification, anthropometrics [height (cm), weight (kg), Body Mass Index (BMI) (kg/m^2), and weight classification)], marital status, socioeconomic status measured via Family Poverty to Income ratio (PIR) (a measure of adjusted family income to relative poverty threshold), and race/ethnicity. Each DT was assessed using classification specificity, sensitivity, and classification error expressed as proportions. Sensitivity quantifies the proportion of correctly classified MetS and specificity gives the proportion of correctly classified non-MetS.

Within the CART algorithm, DT predictor variables were ranked by level of importance related to MetS. The best DT model was chosen and described for each node using the total proportion of MetS and no-MetS classification and a MetS Index describing the estimated probability of MetS compared to the overall prevalence of MetS in the NHANES cohort. For both the training and validation sets, MetS classification threshold was set at the current MetS prevalence within the NHANES cohort in accordance with Stern et al. who used DT models to explain insulin resistance. In this study the classification threshold of the response variable was set at the response variable's prevalence within the study cohort [14]. Instead of maintaining the 50% classification threshold for the response variable, the optimal classification cut-off point was set to maximize the sum of theoretical sensitivity and specificity, as determined from the cohort data. This decision was made to increase the number or correctly classified cases of MetS.

Stepwise Forward Logistic Regression (LR) was performed on the predictor variables used to define MetS as a parametric classification comparison. This procedure was used to approximate the predictive power of the DT techniques. The classification threshold was set at the current prevalence of MetS within the NHANES cohort as mentioned previously. The final LR model was corrected for multicollinearity problems between the predictor variables by removing highly correlated predictor variables. Within LR, severe multicollinearity can cause instability in the model coefficients when highly correlated variables are included in the model. Variables with large amounts of missing data were excluded.

Measure	Defining Cut-off Points
Elevated Waist Circumference^1	
Male	>94 cm
Female	>80 cm
Elevated Triglycerides^2	≥ 150 mg/dl
HDL Cholesterol^2	
Male	<40 mg/dl
Female	<50 mg/dl
Blood Pressure^2	≥130 mmHg Systolic and/or ≥ 80 mmHg Diastolic
Fasting Plasma Glucose^2	≥100 mg/dl

^{1}Values based on lowered AHA/NHLBI Guidelines [1]
^{2}Drug therapy for dyslipidemia, hypertension, and/or hyperglycemia were alternate indicators meeting the criteria for MetS for that risk factor

Table 1: National Health Lung & Blood Institute Metabolic Syndrome Classification Criteria.

Results

Model performance

The average prevalence of MetS within the NHANES cohort was 33.1%. Subject characteristics are displayed in Table 2. The best performing models based on specificity and sensitivity for both males and females (Table 3) were the CART models considering all study parameters as contenders for inclusion. The classification error of each of the best performing models were also the lowest of the DT and LR models at 0.096 and 0.080 for the male and female model, respectively.

Best performing female model

The first split within the DT was based on Triglycerides (TG) which corroborates with the ranked order of importance in Figure 1. The second level was on splits based on either High Density Lipoprotein Cholesterol (HDL-C) or Fasting Plasma Glucose (FPG). All MetS classification risk-factors were present in the model with the exception of Waist Circumference (WC). The only non-MetS predictor variable that the algorithm identified as statistically significant was age for the female cohort (Table 4 and Figure 2) with age greater than 46 years were 6.3 times more likely to be classified with MetS. However, this predictor variable was present in the lowest level within the model. Within the female cohort, all the terminal nodes with significant risk of Mets (MetS Index>1) were based on <3 MetS classification criteria. Within the female model the terminal node with the highest likelihood of presenting with MetS using <3 AHA/NHLBI MetS classification criteria is interpreted as a female patient presenting with TG<150 mg/dl, FPG100 mg/dl, and HDL<50 mg/dl. The probability of MetS for this pathway is 0.969 which results in being 2.910 times more likely to than the average likelihood of presenting with MetS (Table 4, Terminal Node 9).

Best performing male model

The first split within the DT was based on TG which corroborates with the ranked order of importance in Figure 3. All second level splits were based on WC. Considering the risk-factors ranked by importance,

	Male			Female		
Parameter	Total	MetS	No Mets	Total	MetS	No Mets
Age at Screening (year)	36 ± 13	42 ± 11	33 ± 12	37 ± 12	43 ± 11	34 ± 12
Family PIR	2.60 ± 1.65	2.69 ± 1.67	2.55 ± 1.63	2.48 ± 1.67	2.35 ± 1.64	2.53 ± 1.68
Weight (kg)	86.0 ± 19.9	97.2 ± 19.9	79.8 ± 17.0	75.3 ± 20.3	86.9 ± 21.7	70.1 ± 17.3
Standing Height (cm)	175.7 ± 7.7	175.8 ± 7.7	175.6 ± 7.8	162.2 ± 7.0	161.7 ± 7.0	162.4 ± 7.0
Body Mass Index (kg/m²)	27.8 ± 5.8	31.4 ± 5.6	25.8 ± 4.8	28.6 ± 7.3	33.1 ± 7.6	26.5 ± 6.2
Waist Circumference (cm)	96.8 ± 15.8	107.9 ± 13.8	90.8 ± 13.3	93.4 ± 16.6	104.9 ± 15.4	88.2 ± 14.3
Total cholesterol (mg/dl)	193 ± 42	205 ± 45	186 ± 39	192 ± 41	204 ± 45	186 ± 38
LDL-cholesterol (mg/dl)	117 ± 35	124 ± 36	114 ± 35	113 ± 34	122 ± 37	109 ± 32
HDL-cholesterol (mg/dl)	48 ± 13	41 ± 11	52 ± 13	56 ± 15	48 ± 13	60 ± 15
Triglyceride (mg/dl)	147 ± 149	225 ± 212	105 ± 68	115 ± 97	179 ± 144	87 ± 41
Systolic Blood Pres (mmHg)	121 ± 14	127 ± 15	118 ± 12	115 ± 15	123 ± 18	111 ± 12
Diastolic Blood Pres (mmHg)	72 ± 12	77 ± 12	69 ± 11	69 ± 11	74 ± 11	67 ± 10
Fasting Glucose (mg/dl)	104 ± 34	118 ± 48	97 ± 18	99 ± 29	115 ± 44	92 ± 14

Based on Collective Sum of 1999-2012 NHANES Cohorts

Table 2: Sample Characteristics.

		Training			Validation		
Sex	Model	Specificity[1]	Sensitivity[2]	Risk[3]	Specificity[1]	Sensitivity[2]	Risk[3]
Male	CHAID	0.815	0.866	0.167	0.850	0.870	0.172
	E-CHAID	0.814	0.834	0.179	0.796	0.835	0.191
	CART	**0.908**	**0.896**	**0.096**	**0.900**	**0.856**	**0.115**
	LR	0.843	0.878	0.144	0.827	0.853	0.164
	CHAID-MetS	0.654	0.865	0.271	0.597	0.839	0.319
	E-CHAID-MetS	0.718	0.807	0.250	0.686	0.785	0.289
	CART-MetS	0.608	0.898	0.289	0.558	0.881	0.330
	LR-MetS	0.746	0.776	0.244	0.738	0.811	0.239
Female	CHAID	0.889	0.822	0.132	0.870	0.756	0.160
	ECHAID	0.838	0.837	0.162	0.838	0.782	0.179
	CART	**0.952**	**0.848**	**0.080**	**0.939**	**0.811**	**0.100**
	LR	0.881	0.854	0.127	0.878	0.835	0.134
	CHAID-MetS	0.722	0.754	0.268	0.689	0.786	0.282
	ECHAID-MetS	0.718	0.755	0.271	0.698	0.777	0.284
	CART-MetS	0.767	0.732	0.244	0.753	0.731	0.254
	LR-MetS	0.770	0.706	0.249	0.733	0.724	0.270

CHAID = X^2 Automatic Interaction Detection, E-CHAID = Exhaustive X^2 Automatic Interaction Detection, CART = Classification and Regression Tree, LR = Logistic Regression

Bold values indicate best performing models

[1] Specificity = Proportion of correctly classified non-MetS cases

[2] Sensitivity = Proportion of correctly classified MetS cases

[3] Risk = MetS misclassification defined as total proportion of MetS misclassification

Table 3: Model Performance.

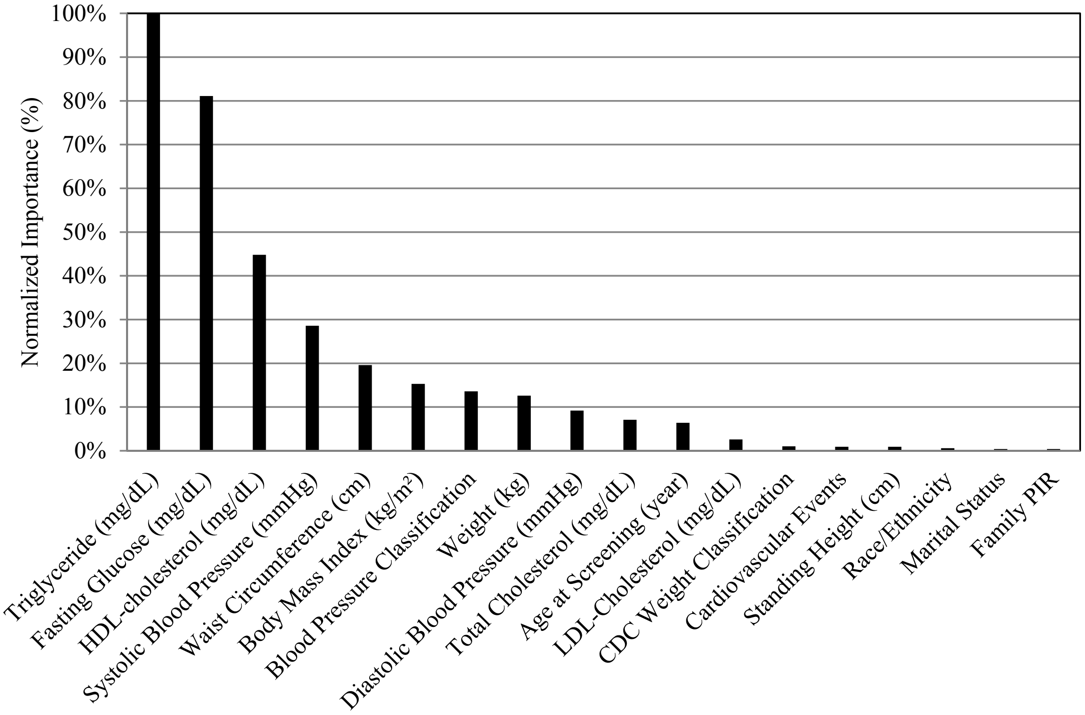

Figure 1: Female CART Decision Tree Ranked Order of Normalized Importance.

Terminal Node	Level 1	Level 2	Level 3	Level 4	Level 5	MetS¹	MetS Index²
9	TG<150	**FPG ≥ 100**	**HDL-C<50**	*	*	0.969	2.910
5	**TG ≥ 150**	**HDL-C<50**	*	*	*	0.962	2.889
12	**TG ≥ 150**	HDL-C ≥ 50	**FPG ≥ 100**	*	*	0.956	2.871
14	TG<150	**FPG ≥ 100**	HDL-C ≥ 50	**SBP ≥ 130**	*	0.897	2.694
16	TG<150	**FPG ≥ 100**	HDL-C ≥ 50	SBP<130	Age ≥ 46	0.391	1.174
11	**TG ≥ 150**	HDL-C ≥ 50	FPG<100	*	*	0.330	0.991
8	TG<150	FPG<100	**SBP ≥ 131**	*	*	0.304	0.913
15	TG<150	**FPG ≥ 100**	HDL-C ≥ 50	SBP<130	Age<46	0.062	0.186
7	TG<150	FPG<100	SBP<131	*	*	0.024	0.072

TG = Triglycerides (mg/dl), WC = Waist Circumference (cm), FPG = Fasting Plasma Glucose (mg/dl), SBP = Systolic Blood Pressure (mmHg), HDL-C = High Density Lipoprotein Cholesterol (mg/dl), Age (years)
Bold values indicate that the predictor reached the MetS threshold within the AHA/NHLBI MetS Classification Criteria
* Indicates no further node splits within level
¹Probability of MetS Classification
² MetS Index = Estimated probability of MetS compared to the overall prevalence of MetS in the NHANES cohort (33.1%)

Table 4: Female CART Decision Tree Model Performance.

BMI was in the top predictor variables. However this predictor variable did not appear in the model. Within the male DT, there were 3 pathways that resulted in a significant increase in likelihood of MetS (MetS Index>1) was based on <3 MetS criteria (Table 5 and Figure 4). Within the male model the terminal node with the highest likelihood of presenting with MetS using <3 AHA/NHLBI MetS classification criteria is interpreted as a male patient presenting with TG ≥ 150 mg/dl, WC<94 cm, and a FPG>100 mg/dl. The probability of MetS for this pathway is 0.655 which results in being 1.967 times more likely to than the average likelihood of presenting with MetS (Table 5, Terminal Node 10).

Discussion

The purpose of the current investigation was to create, compare, and validate sex-specific DT models to classify MetS. DT models were derived using CHAID, E-CHAID, and CART algorithms based on the presence of MetS as the response variable and the MetS classification criteria, predictor variables from cardiovascular risk model and subject characteristics as the predictor variables whose values were obtained from 1999-2012 NHANES data [3,6,10,11,13]. MetS is classified by the presence of 3 of 5 criteria defined by AHA/NHLBI classification guidelines [1].

This study has multiple novelties. First, these models are based on large amounts of data that is representative of adults in the United States. Second, the pathways derived from this model show promise in accurately classifying sex-specific MetS using fewer measurements than traditional classification criteria. Third, unlike traditional MetS classification models, the pathways of the current investigation do not provide universal cutoffs for each predictor variable. Rather, these pathways consider the clustering and multilevel interactions among predictor variables to identify stepwise pathways to classify MetS. Finally, each pathway describes the likelihood of developing MetS.

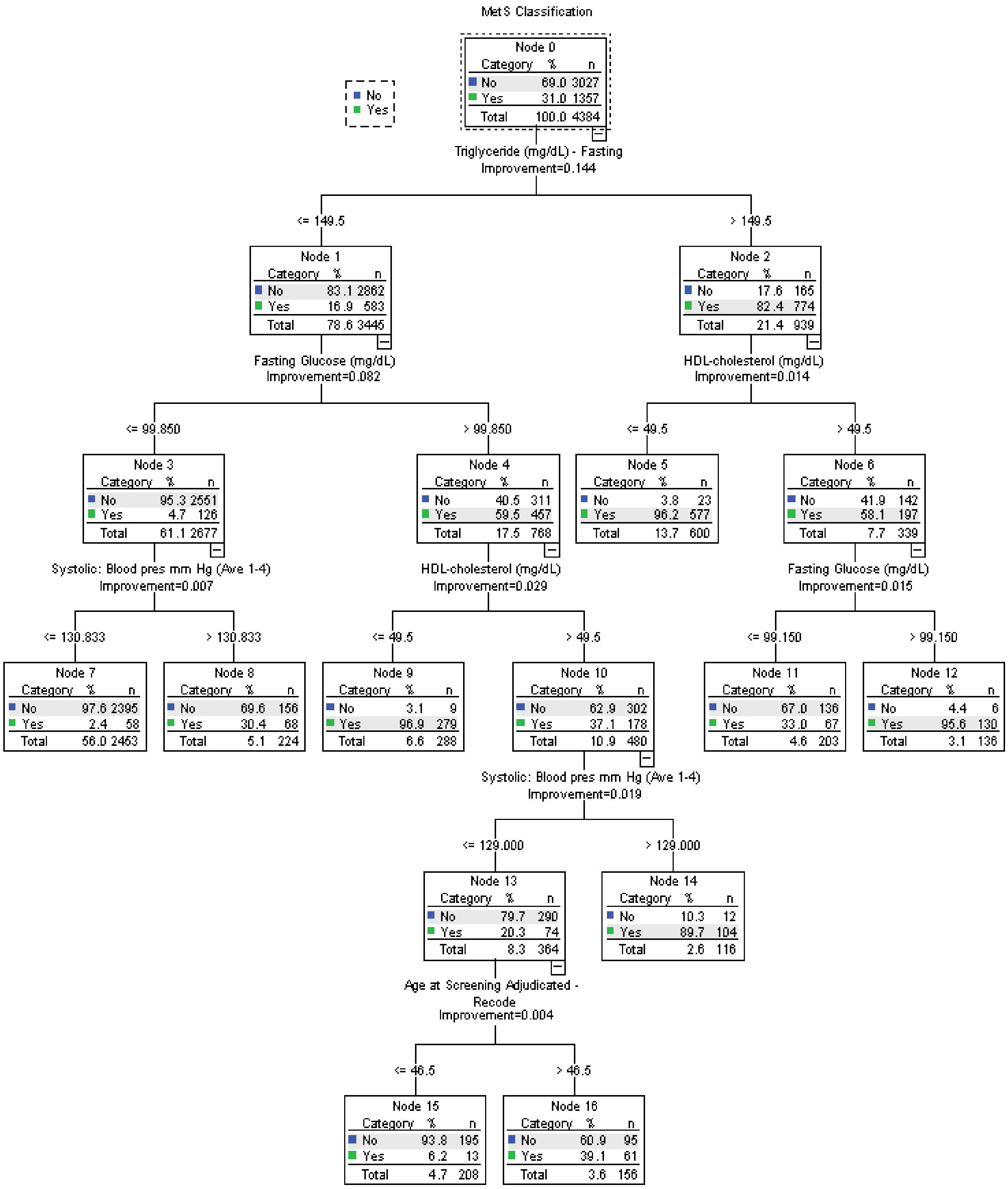

Figure 2: Female MetS Classification Decision Tree. Tree Growth Method = CART. All study parameters were contenders for inclusion on CART model. The absence of a parameter indicates that it did not reach significance for inclusion in the model.

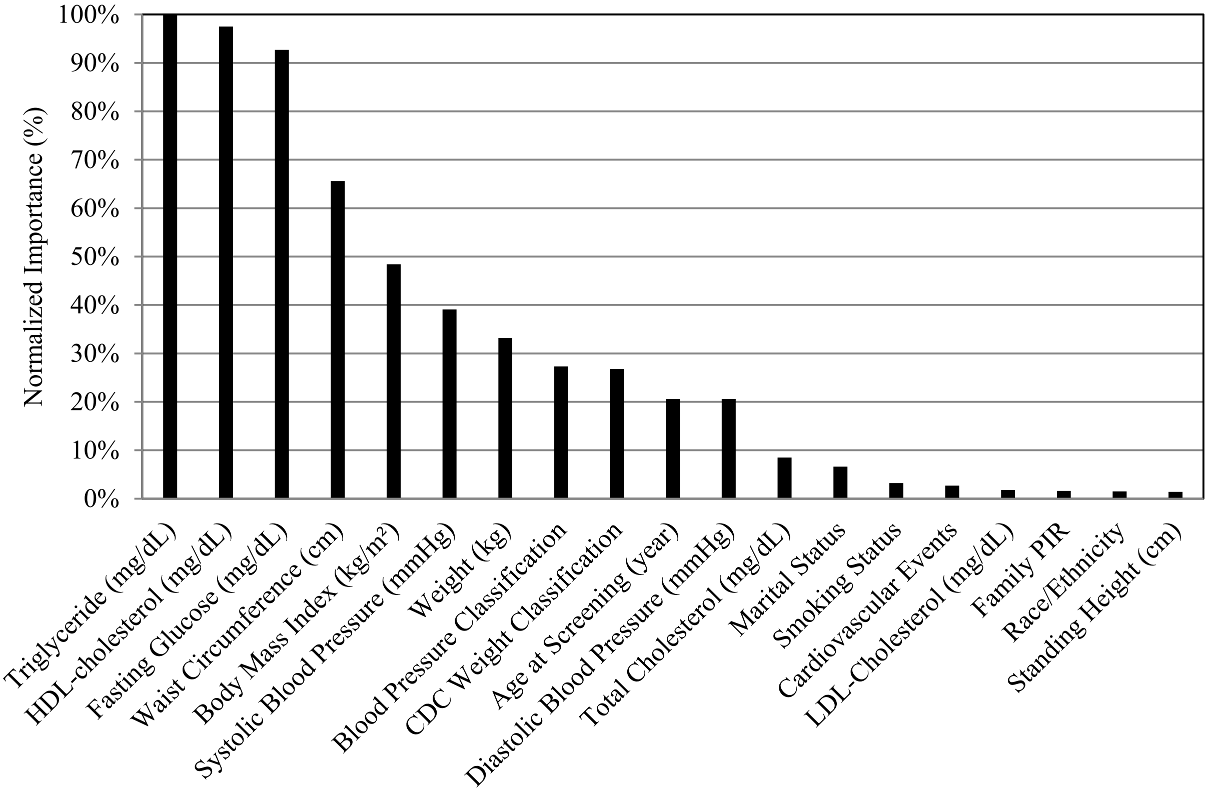

Figure 3: Male CART Decision Tree Ranked Order of Normalized Importance.

In this study the prevalence of MetS within the NHANES cohort, a representative sample of the United States adult population, was 33.1% which approximates Ford, Li, & Zhao's study that found the prevalence of MetS within the NHANES cohort to be 34.3% [3].

The first level split indicates the risk-factor with the highest association with MetS. The first level split was based on TG which corroborates with Worachartcheewan et al. who used CART to classify MetS in a sample of Thai men and women [13]. The results of this study also corroborate with Miller, Fridline, Liu, & Marino who used the CHAID algorithm to classify MetS in a sample of young adults using NHANES data. The best performing model in this study was built as a user-specified first level split on WC [6]. When the algorithm was not user-specified, the CHAID algorithm identified TG as the first level split. Interestingly in this study, the proposed CHAID model with the user-specified first split on WC outperformed the CHAID algorithm without first-level split specification and the logistic regression model in both overall sensitivity and classification accuracy for MetS.

There were notable differences between the male and female models. The first was that WC was present in the male model but not the female model. This phenomenon might be based on the body fat distribution of women prior to menopause that occurs in women at or near the age split identified by the DT model [20]. This suggests that a moderate increase in adiposity would not result in a significant increase in central adiposity. Therefore the WC measurement might not be warranted in women. Conversely for men, the body fat distribution would contribute to increases in central adiposity as body fat increases. This finding corroborates with Hari et al. who compared sex-specific differences between multiple MetS classification models and found that measures of central adiposity, specifically WC, were more profound for males than females [21]. Future investigation regarding this phenomenon is warranted considering that physicians and health professionals recommend WC measurements for both sexes.

Also notable was the close relationship between WC and BMI based on the normalized order of importance in Figures 1 and 4. Both WC and BMI have been shown to be a strong proxy of visceral adiposity [22]. However, BMI only considers the relationship of weight to height and does not consider actual body composition and girth measurements. Central adiposity has been identified as a strong predictor of MetS and a strong contributor to BMI and Despres et al. demonstrated a strong correlation between BMI and WC which suggests the interchangeability of measures [22]. Given that WC was more significantly associated with MetS than BMI, the inclusion of WC most likely diminished the effect of BMI in the DT models. Therefore WC seems to be a more sensitive predictor of MetS than BMI.

Also interesting in the female model was the inclusion of a non-MetS classification criterion parameter, age. Although this factor was not present in a high-risk MetS pathway (MetS Index>1), age ≥ 46 years were 6.3 times likely to present with MetS than females within this pathway with an age<46 years. One suggestion for the split based on age was at 46 years relates to the cardiometabolic changes related to menopause. However, a review by Barret-Conner of menopause in relation to CVD risk in women delineated the direct relationship between menopause and CVD risk [23]. The methodology of the current investigation was unable to explain the inclusion of this predictor. Further investigation exploring the relationship between central adiposity and likelihood of presenting with MetS for women by age and pre, peri, and post-menopause is warranted.

A successful improvement in current methodologies using the

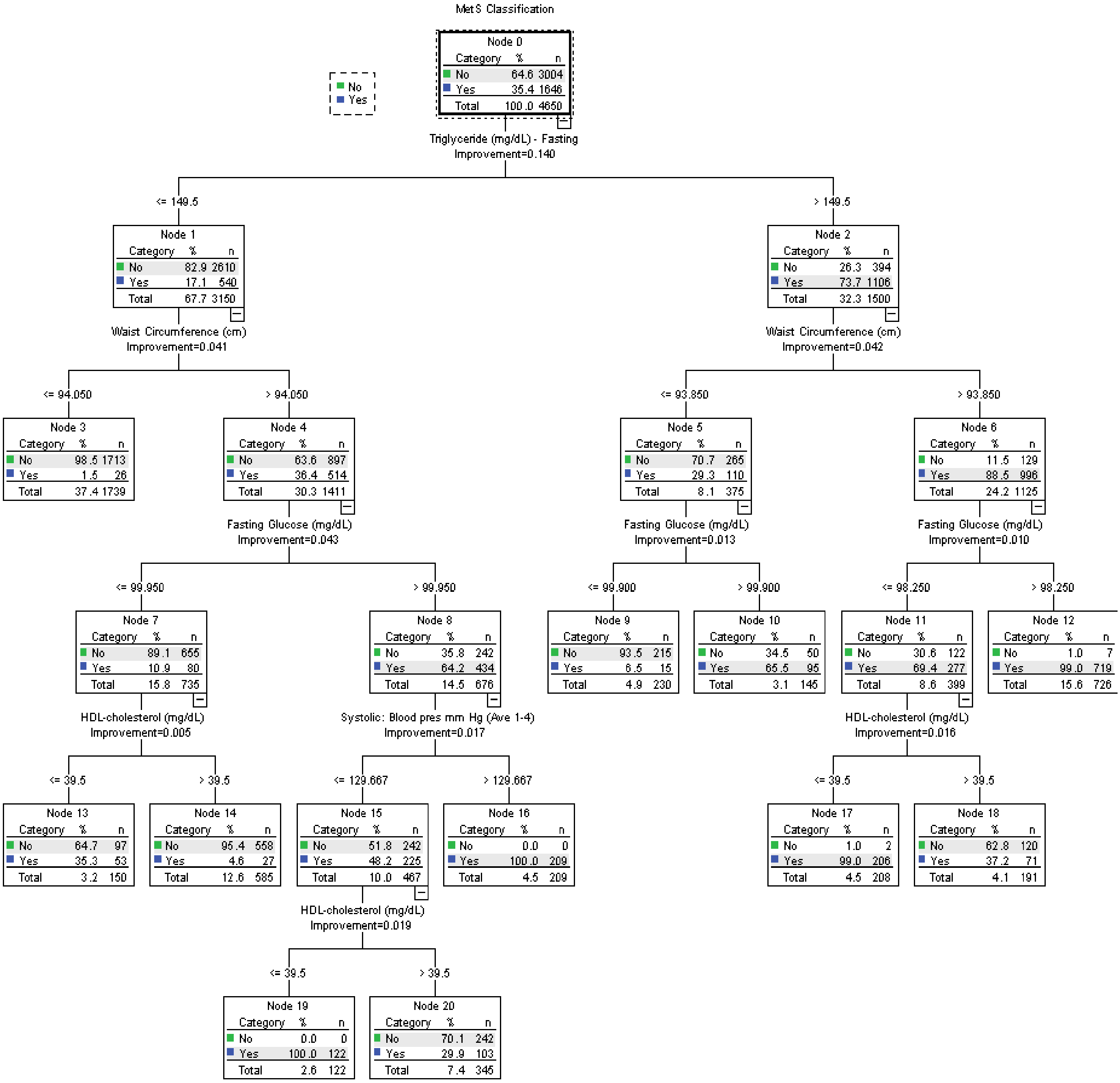

Figure 4: Male MetS Classification Decision Tree. Tree Growth Method = CART. All study parameters were contenders for inclusion on CART model. The absence of a parameter indicates that it did not reach significance for inclusion in the model.

models developed in the current investigation in relation to other classification models would be the classification of MetS with less than 3 risk-factors and/or identify the MetS risk of multiple clustering combinations of predictor variables. In the female model all of the pathways leading to increased risk of MetS were based on less than 3 predictor variables. However, in the male model only one pathway required less than 3 predictor variables for MetS classification. Clinical application of these pathways can inform health educators and/or clinicians identifying high risk pathways and focusing on interventions that could shift a patient to a lower risk pathway.

Conclusions

In summary, the current investigations findings suggest that DT-based pathways to classify MetS and likelihood of presenting with MetS could detect MetS before other classification models. Within the female model, waist circumference measures did not reach significance as a predictor variable. However, age did reach significance for inclusion in the female model. Five of the pathways with increased likelihood of MetS in the female model were built using ≤ 2 MetS AHA/NHBLI classification criteria. Three of the pathways with increased likelihood of MetS in the male model were built using ≤ 2 MetS AHA/NHLI

Terminal Node	Level 1	Level 2	Level 3	Level 4	Level 5	MetS[1]	MetS Index[2]
16	TG<150	**WC > 94**	**FPG ≥ 100**	**SBP > 130**	*	1.000	3.003
19	TG<150	**WC > 94**	**FPG ≥ 100**	SBP<130	**HDL-C<40**	1.000	3.003
12	**TG ≥ 150**	**WC ≥ 94**	FPG > 98	*	*	0.990	2.973
17	**TG ≥ 150**	**WC ≥ 94**	FPG ≤ 98	**HDL-C<40**	*	0.990	2.973
10	**TG ≥ 150**	WC<94	**FPG ≥ 100**	*	*	0.655	1.967
9	**TG ≥ 150**	WC<94	FPG<100	*	*	0.650	1.952
18	**TG ≥ 150**	**WC ≥ 94**	FPG ≤ 98	HDL-C ≥ 40	*	0.372	1.117
13	TG<150	WC ≤ 94	FPG<100	**HDL-C<40**	*	0.353	1.060
20	TG<150	**WC>94**	**FPG ≥ 100**	SBP<130	HDL-C ≥ 40	0.299	0.898
14	TG<150	WC ≤ 94	**FPG ≥ 100**	HDL-C ≥ 40	*	0.046	0.138
3	TG<150	WC ≤ 94	*	*	*	0.015	0.045

TG = Triglycerides (mg/dl), WC = Waist Circumference (cm), FPG = Fasting Plasma Glucose (mg/dl), SBP = Systolic Blood Pressure (mmHg), HDL-C = High Density Lipoprotein Cholesterol (mg/dl), Age (years)

Bold values indicate that the predictor reached the MetS threshold within the AHA/NHLBI MetS Classification Criteria

* Indicates no further node splits within level

[1] Probability of MetS Classification

[2] MetS Index = Estimated probability of MetS compared to the overall prevalence of MetS in the NHANES cohort (33.1%)

Table 5: Male CART Decision Tree Model Performance.

classification criteria. Future research warrants the implementation and further validation of these pathways using a clinical sample. There still remains no clinically established criterion for pre-metabolic syndrome. These pathways show promise in developing a preliminary pre-metabolic syndrome classification tool to guide intervention before the onset of MetS using current models.

References

1. Alberti KGMM, Eckel RH, Grundy SM, Paul ZZ, James IC, et al. (2009) Harmonizing the metabolic syndrome: A joint interim statement of the international diabetes federation task force on epidemiology and prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation 120: 1640-1645.
2. Grundy SM (2011) The metabolic syndrome. Atlas of atherosclerosis and metabolic syndrome 1: 1-26.
3. Ford ES, Li C, Zhao G (2010) Prevalence and correlates of metabolic syndrome based on a harmonious definition among adults in the US. Journal of Diabetes 2: 180-193.
4. Birnbaum HG, Mattson ME, Kashima S, Williamson TE (2011) Prevalence rates and costs of metabolic syndrome and associated predictor variables using employees' integrated laboratory data and health care claims. J Occup Environ Med 53: 27-33.
5. Boudreau D, Malone D, Raebel M (2009) Health care utilization and costs by metabolic syndrome predictor variables. Metabolic syndrome and related disorders 7: 305-314.
6. Miller B, Fridline M, Liu P, Marino D (2014) Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. Computational and mathematical methods in medicine.
7. Alberti KGM, Zimmet P, Shaw J (2005) The metabolic syndrome—a new worldwide definition. The Lancet 366: 1059-1062.
8. Kassi E, Pervanidou P, Kaltsas G, Chrousos G (2011) Metabolic syndrome: Definitions and controversies. BMC Med 9: 48.
9. Strazzullo P, Barbato A, Siani A, Cappuccio FP, Versiero M, et al. (2008) Diagnostic criteria for metabolic syndrome: A comparative analysis in an unselected sample of adult male population. Metab Clin Exp 57: 355-361.
10. Simmons R, Alberti K, Gale E, Colagiuri S, Tuomilehto Q, et al. (2010) The metabolic syndrome e: Useful concept or clinical tool? Report of a WHO expert consultation. Diabetologia 53: 600-605.
11. Liu J, Grundy SM, Wang W, Smith SC, Vega GL, et al. (2007) Ten-year risk of cardiovascular incidence related to diabetes, prediabetes, and the metabolic syndrome. Am Heart J 153: 552-558.
12. Edelenyi FS, Goumidi L, Bertrais S, Phillips C, Macmanus R, et al. (2008) Prediction of the metabolic syndrome status based on dietary and genetic parameters, using random forest. Genes & nutrition 3: 173-176.
13. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V (2010) Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pract 90: e15-e18.
14. Stern SE, Williams K, Ferrannini E, DeFronzo RA, Bogardus C, et al. (2005) Identification of individuals with insulin resistance using routine clinical measurements. Diabetes 54: 333-339.
15. Gandomi AH, Fridline MM, Roke DA (2013) Decision tree approach for soil liquefaction assessment. The Scientific World Journal 8.
16. Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Applied statistics 119-127.
17. Biggs D, De Ville B, Suen E (1991) A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics 18: 49-62.
18. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press.
19. Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. John Wiley & Sons.
20. Camhi SM, Bray GA, Bouchard C, Greenway FL, Johnson WD, et al. (2011) The relationship of waist circumference and BMI to visceral, subcutaneous, and total body fat: Sex and race differences. Obesity 19: 402-408.
21. Hari P, Nerusu K, Veeranna V, Sudhakar R, Zalawadiya S, et al. (2012) A gender-stratified comparative analysis of various definitions of metabolic syndrome and cardiovascular risk in a multiethnic US population. Metabolic syndrome and related disorders 10: 47-55.
22. Despres JP, Lemieux I, Bergeron J, Pibarot P, Mathieu P, et al. (2008) Abdominal obesity and the metabolic syndrome: Contribution to global cardiometabolic risk. Arterioscler Thromb Vasc Biol 28: 1039-1049.
23. Connor BE (2013) Menopause, atherosclerosis, and coronary artery disease. Current opinion in pharmacology 13: 186-191.