

Research article

Open Access

## Binary classification of dyslipidemia from the waist-to-hip ratio and body mass index: a comparison of linear, logistic, and CART models

Michael C Costanza\*<sup>1</sup> and Fred Paccaud<sup>2</sup>

Address: <sup>1</sup>Division of Clinical Epidemiology, Geneva University Hospitals, Geneva, Switzerland and <sup>2</sup>Institute of Social and Preventive Medicine, University of Lausanne, Lausanne, Switzerland

Email: Michael C Costanza\* - Michael.Costanza@hcuge.ch; Fred Paccaud - Fred.Paccaud@inst.hospvd.ch

\* Corresponding author

Published: 06 April 2004

Received: 27 October 2003

BMC Medical Research Methodology 2004, 4:7

Accepted: 06 April 2004

This article is available from: <http://www.biomedcentral.com/1471-2288/4/7>

© 2004 Costanza and Paccaud; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** We sought to improve upon previously published statistical modeling strategies for binary classification of dyslipidemia for general population screening purposes based on the waist-to-hip circumference ratio and body mass index anthropometric measurements.

**Methods:** Study subjects were participants in WHO-MONICA population-based surveys conducted in two Swiss regions. Outcome variables were based on the total serum cholesterol to high density lipoprotein cholesterol ratio. The other potential predictor variables were gender, age, current cigarette smoking, and hypertension. The models investigated were: (i) linear regression; (ii) logistic classification; (iii) regression trees; (iv) classification trees (iii and iv are collectively known as "CART"). Binary classification performance of the region-specific models was externally validated by classifying the subjects from the other region.

**Results:** Waist-to-hip circumference ratio and body mass index remained modest predictors of dyslipidemia. Correct classification rates for all models were 60–80%, with marked gender differences. Gender-specific models provided only small gains in classification. The external validations provided assurance about the stability of the models.

**Conclusions:** There were no striking differences between either the algebraic (i, ii) vs. non-algebraic (iii, iv), or the regression (i, iii) vs. classification (ii, iv) modeling approaches. Anticipated advantages of the CART vs. simple additive linear and logistic models were less than expected in this particular application with a relatively small set of predictor variables. CART models may be more useful when considering main effects and interactions between larger sets of predictor variables.

### Background

Central adiposity is a predictor of cardiovascular disease (CVD) independently of other major risk factors, including body mass index (BMI) [1,2]. Part of the relationship between central adiposity and CVD is mediated by a mod-

ification of the metabolism of insulin and lipids [3]. Dyslipidemic individuals are more frequently "centrally obese" (*e.g.*, with a high waist-to-hip circumference ratio (WHR)) [4-6]. These observations have been made in a variety of populations from developed [7-9] and less

developed countries [9]. Apart from its interest for establishing a physiopathological causal link, this predictive association suggests the possibility of employing one or more anthropometric measurements of central adiposity as a first step in population screening for dyslipidemia [8,9]. Using inexpensive and readily obtainable anthropometric measurements instead of more costly and time-consuming wet- or even dry-chemistry laboratory cholesterol measurements is relevant even in developed countries where an emerging epidemic of CVD is occurring amidst rising health care costs.

One objective of the present study was to attempt to improve upon previous statistical strategies for detecting dyslipidemia in the general population, with specific focus on the predictive power of the anthropometric measurements WHR and BMI. A second objective was to compare the performance of four statistical modeling approaches that can be employed for binary classification: linear regression [10], logistic classification [11], and classification and regression trees (CART) [12,13]. By *can be employed* we mean: (a) with a modest amount of effort using commercially available software (we used SAS [14] and S-Plus [15]); and (b) that it is possible to apply classification-type methods for a binary outcome to the results of regression-type methods for a continuous outcome. We also wondered how well competing methods perform in practice, as opposed to how well they are supposed to perform in theory.

## Methods

### Study populations and samples

Subjects participated in the World Health Organization (WHO) MONICA (MONItoring trends and determinants in Cardiovascular disease) project described in detail elsewhere [16]. Participating regions included Vaud-Fribourg and Ticino in Switzerland. Vaud and Fribourg are adjacent French-speaking cantons in the west/southwest, while Ticino is an Italian-speaking canton in the southeast. These regions had similar distributions of and correlations between the predictor and outcome variables employed in the statistical models (see Results). Accordingly, the classification performance of region-specific models was estimated by external validation on data from the other region, as well as by (biased) resubstitution.

The third independent 1992–93 MONICA surveys were used. In Vaud-Fribourg, 3,299 individuals aged 25–74 years were invited to participate, and 1,742 (53%) did so. In Ticino, 2,000 individuals aged 35–64 years were invited and 1,510 (76%) participated. Analyses in the present study were restricted to the age range 35–64 years common to both regions (Vaud-Fribourg  $n = 1,182$ , Ticino  $n = 1,510$ ). In addition to WHR and BMI, the potential predictor variables examined were Gender, Age,

current cigarette Smoking, and high blood pressure (HBP: diastolic BP  $\geq 90$  mm Hg or under hypertension treated). Linear and logistic regression (but not CART) models require complete data on the study subjects, unless missing data imputation techniques are employed. For convenience, we excluded subjects with missing data on any of the predictor variables. This reduced the final sample sizes by 5% in Vaud/Fribourg ( $n = 1,120$ ) and by 6% in Ticino ( $n = 1,429$ ).

### Statistical models

Although the total serum cholesterol to high density lipoprotein cholesterol (TC/HDL-C) ratio is a continuous variable, we assumed that assessing the dyslipidemia classification performance of a predictive model would *ultimately* require comparing predicted *binary* values of dyslipidemia status. We applied five modeling approaches (Strategies 0–4) which reflected: no model (0); algebraically specified (1, 2) *vs.* unspecified (3, 4) models; and regression-(1, 3) *vs.* classification-based (2, 4) models. Strategies 1–4 were expected to outperform the minimal benchmark Strategy 0.

#### Strategy 0: modal regional prevalence of dyslipidemia (no model)

Individuals in a given region were classified as dyslipidemic or not dyslipidemic, depending on the observed modal (most frequent) dyslipidemia category either in the whole region or stratified by gender. Strategy 0 represented a "no model" approach in the sense that the additional predictor variables were ignored.

#### Strategy 1: linear regression

Additive linear models,

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e,$$

where  $Y = \text{TC/HDL-C ratio}$ ,  $\{X_1, X_2, \dots, X_k\}$  ( $k \leq 6$ ) = a subset of the predictor variables {WHR, BMI, Gender, Age, Smoking, HBP}, and  $e = \text{Gaussian error with constant variance}$ , were fitted. TC/HDL-C, WHR, BMI, and Age were analyzed as continuous variables, while Gender, Smoking, and HBP were analyzed as binary variables. An individual with estimated  $Y \geq 5.0$  was classified as dyslipidemic, or classified as not dyslipidemic otherwise.

Including all the predictor variables was termed the *full* model, while including only {WHR, BMI, Gender} was termed the *reduced* model. Both types of model were fitted separately by region. In addition, for women and men separately, {WHR, BMI, Age, Smoking, HBP} "full" and {WHR, BMI} "reduced" models also were fitted.

No formal predictor variable selection procedures, nor models with predictor variable product-interactions were

employed. We simply wished to magnify any differences and facilitate comparisons between the algebraic linear regression (Strategy 1) *vs.* non-algebraic regression tree models (Strategy 3).

### Strategy 2: logistic classification

For the same predictor variables as in Strategy 1, but with *binary*  $Y = 1$  if  $TC/HDL-C \geq 5.0$ ,  $Y = 0$  otherwise, additive logistic models

$$\log[p/(1-p)] = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e,$$

where  $p$  = probability that  $Y = 1$  for given values of the predictors and  $e$  = binomial error term, were fitted. This model assumes the relationship between  $\log[p/(1-p)]$  and the predictor variables is linear. An individual with estimated  $p \geq 0.50$  was classified as dyslipidemic, or classified as not dyslipidemic otherwise.

As in Strategy 1, neither predictor variable selection nor specification of predictor variable product interactions were employed to magnify differences and facilitate comparisons between the algebraic logistic classification (Strategy 2) *vs.* non-algebraic classification tree models (Strategy 4).

### Strategy 3: regression trees

For the same predictor variables and *continuous*  $Y$ s as in Strategy 1, regression tree models also were fitted. At each one-step-look-ahead of the "full" tree-growing process, the  $Y$ s were examined within all possible binary splits of each predictor variable to select the best single split for creating homogeneous groups with maximal between-group mean-squared errors. This process was continued until "optimality" of the groups at the final nodes ("leaves") of the tree was achieved. In practice, the full tree tends to be overly complex and idiosyncratic with respect to the data employed to "grow" it. Thus, a common recommendation [*e.g.*, [17]] is to "prune" the full tree backwards through further criteria based on both maximal within-leaf homogeneity of the  $Y$ s and minimal tree size in order to produce a smaller pruned tree that is less subject to these drawbacks. It is also recommended [17] that the process be internally cross-validated, *e.g.*, by randomly dividing the data into tenths, performing the pruning on the full tree grown with nine tenths and evaluating it on the remaining tenth of the data, and averaging the classification performance criteria (see below) from all ten 9:1 partitions of the data.

After following these recommendations, the estimated value of  $Y$  at each pruned tree leaf was taken to be the mean among those subjects comprising the leaf. All individuals in the leaf were classified as dyslipidemic if the

estimated  $Y \geq 5.0$ , or classified as not dyslipidemic otherwise.

### Strategy 4: classification trees

For the same predictor variables and *binary*  $Y$ s as in Strategy 2, classification tree models also were fitted. The rationale, algorithms, and recommendations employed were similar to those for regression trees, with one important difference. An appealing recommendation [17] to employ both minimal misclassification rate (instead of maximal within-leaf homogeneity of the  $Y$ 's) and minimal tree size optimality criteria to prune the full tree backwards was followed and internally cross-validated as described in Strategy 3.

The estimated value of  $Y$  at each pruned tree leaf was taken to be the modal category (dyslipidemic or not) among those subjects comprising the leaf. All individuals in a leaf were then classified in accord with the modal category.

### Classification performance criteria

The classification performance of all models were compared in terms of five measures: (1) overall correct classification (total % agreement between observed and model-classified dyslipidemia status); (2) sensitivity (% with observed  $TC/HDL-C \geq 5.0$  and classified as such); (3) specificity (% with observed  $TC/HDL-C < 5.0$  and classified as such); (4) positive predictive value (PPV, % classified as  $TC/HDL-C \geq 5.0$  and observed as such); (5) negative predictive value (NPV, % classified as  $TC/HDL-C < 5.0$  and observed as such). For the Vaud-Fribourg and Ticino region-specific models, all five classification performance measures were estimated by resubstitution of the data from the same region as well as by external validation on the subjects from the other region.

## Results

### Descriptive comparisons of the two study samples

The predictor and outcome variables in the Vaud-Fribourg and Ticino MONICA study samples are summarized in Table 1. Switzerland has a relatively high prevalence of dyslipidemia (especially among men) compared to other countries [18]. The Ticino subjects were on average two years older, had a slightly higher  $TC/HDL-C$  ratio and thus a higher prevalence of dyslipidemia, and had more current cigarette smokers (predominantly among men) than the Vaud-Fribourg subjects. On the other hand, the distributions of WHR and BMI were similar in both regions.

The correlation matrices for both regions indicated that the bivariate relationship patterns also were similar (Table 2). WHR, BMI, and Gender had the highest correlations with the  $TC/HDL-C$  ratio (continuous or binary), with noticeable attenuation of the gender-specific correlations between WHR and  $TC/HDL-C$ . Further, the correlations

**Table 1: Comparisons of Swiss MONICA samples (ages 35–64 yrs).**

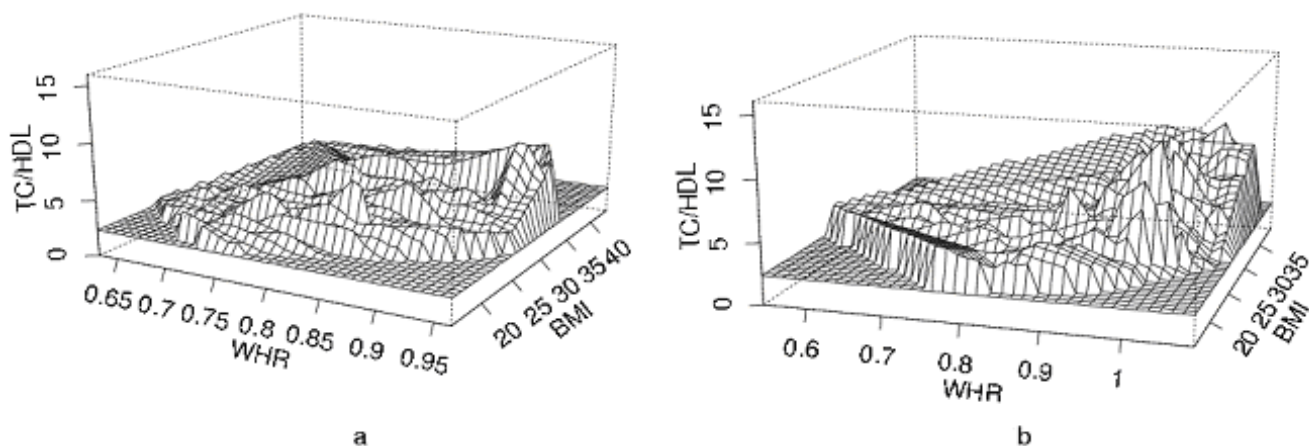
Study Variable	Vaud-Fribourg <sup>a</sup>		Ticino <sup>b</sup>	
Male Gender	48.9%		48.2%	
Age (yrs) <sup>c</sup>	47.8 ± 8.5		49.5 ± 8.2	
	Women	47.9 ± 8.5		49.7 ± 8.3
	Men	47.8 ± 8.4		49.2 ± 8.1
TC/HDL-C ratio <sup>c</sup>	4.9 ± 1.7		5.1 ± 1.8	
	Women	4.2 ± 1.3		4.4 ± 1.6
	Men	5.7 ± 1.8		5.8 ± 1.9
Dyslipidemia <sup>d</sup>	41.6%		44.4%	
	Women	22.4%		25.9%
	Men	61.7%		64.4%
WHR <sup>c</sup>	0.85 ± 0.09		0.85 ± 0.08	
	Women	0.78 ± 0.05		0.80 ± 0.06
	Men	0.92 ± 0.06		0.91 ± 0.05
BMI (kg/m <sup>2</sup> ) <sup>c</sup>	25.6 ± 4.0		26.0 ± 4.3	
	Women	24.6 ± 4.2		25.4 ± 4.9
	Men	26.5 ± 2.6		26.6 ± 3.4
Current Cigarette Smoking	25.6 %		31.1 %	
	Women	24.7%		26.5%
	Men	26.6%		36.2%
Hypertension <sup>e</sup>	21.7%		22.9%	
	Women	14.7%		17.0%
	Men	29.0%		29.2%

<sup>a</sup> n = 1,120 (572 Women, 548 Men) <sup>b</sup> n = 1,429 (741 Women, 688 Men) <sup>c</sup> Mean ± SD. <sup>d</sup> TC/HDL-C ratio ≥ 5.0. <sup>e</sup> Diastolic blood pressure > 90 mmHg and/or treated hypertension.

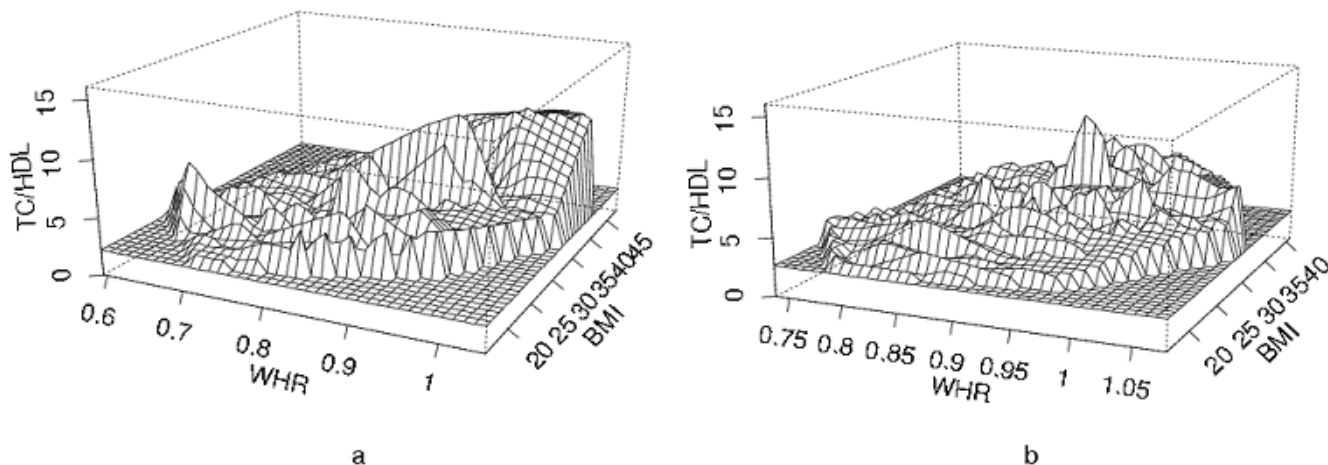
**Table 2: Correlations among study variables in two Swiss MONICA samples (ages 35–64 yrs).**

	WHR	BMI	Age	Current Smoking	Hypertension <sup>d</sup>	Gender
TC/HDL-C ratio	0.53 <sup>a</sup> /0.49 <sup>b</sup>	0.41/0.36	0.14/0.09	0.11/0.13	0.19/0.20	0.43/0.38
Women	0.37/0.42	0.36/0.41	0.27/0.30	0.15/0.14	0.20/0.23	-
Men	0.32/0.27	0.36/0.34	0.06/-0.07	0.08/0.07	0.09/0.09	-
Dyslipidemia <sup>c</sup>	0.46/0.48	0.36/0.35	0.13/0.13	0.09/0.10	0.15/0.19	0.40/0.39
Women	0.37/0.35	0.34/0.35	0.24/0.29	0.13/0.08	0.18/0.22	-
Men	0.21/0.27	0.27/0.32	0.07/0.03	0.06/0.05	0.02/0.09	-
WHR		0.53/0.49	0.19/0.21	0.03/0.10	0.24/0.24	0.77/0.72
Women		0.51/0.52	0.29/0.32	0.05/0.02	0.20/0.28	-
Men		0.61/0.53	0.31/0.34	0.00/0.05	0.13/0.12	-
BMI			0.23/0.23	-0.08/-0.07	0.27/0.26	0.24/0.13
Women			0.31/0.31	-0.08/-0.11	0.28/0.33	-
Men			0.15/0.13	-0.10/-0.04	0.21/0.17	-
Age				-0.12/-0.09	0.18/0.19	-0.00/-0.03
Women				-0.15/-0.08	0.20/0.26	-
Men				-0.09/-0.10	0.17/0.15	-
Current Smoking <sup>d</sup>					-0.02/-0.04	0.02/0.11
Women					-0.03/-0.06	-
Men					-0.01/-0.06	-

<sup>a, b</sup> Pearson correlations (r) in Vaud-Fribourg <sup>a</sup>/Ticino <sup>b</sup>. <sup>c</sup> I = (TC/HDL-C ratio ≥ 5.0), 0 = otherwise. <sup>d</sup> I = Yes, 0 = No.



**Figure 1**  
3-D perspective plots of TC/HDL-C ratio vs. WHR and BMI. **a:** Vaud-Fribourg women (n = 572). **b:** Vaud-Fribourg men (n = 548).



**Figure 2**  
3-D perspective plots of TC/HDL-C ratio vs. WHR and BMI. **a:** Ticino women (n = 741). **b:** Ticino men (n = 688).

between TC/HDL-C and Age, Smoking, and HBP were markedly stronger (albeit still low) among women than men. The highest correlation ( $r > 0.7$ ) among the predictor variables was between WHR and Gender (see also Table 1). The next highest was between WHR and BMI ( $r \geq 0.49$ , overall and gender-specific). These results indicated that WHR, BMI, and Gender would probably be the most important of the predictor variables examined.

Accordingly, 3-D perspective plots of TC/HDL-C ratio vs. WHR and BMI were obtained to visualize what the

anthropometric measures were expected to predict (Figures 1, 2). The irregularities in the figures are striking; *i.e.*, the surfaces are not very "smooth". Hence, smooth predictive functions for binary classification such as the additive, algebraically specified linear regression or logistic classification models might not have been expected to perform so well. On the other hand, the non-additive, non-algebraically specified CART models might have been expected to perform relatively better.

**Table 3: Classification performance of overall (both genders) reduced {WHR, BMI, Gender} models for Vaud-Fribourg, with cross-validation on Ticino subjects.**

(Strategy) Fitted Model	Total % Correct	Sensitivity	Specificity	+ Predictive Value (PPV)	- Predictive Value (NPV)
<i>Classifications of both genders</i>					
(0) No Model <sup>a</sup>	58 <sup>c</sup> (56) <sup>d</sup>	0 (0)	100 (100)	0 (0)	58 (56)
(1) Linear Regression	71 (72)	73 (78)	69 (68)	63 (66)	78 (79)
(2) Logistic Classification	71 (72)	63 (67)	77 (77)	66 (70)	74 (74)
(3) 2-Node Reg. Tree <sup>e</sup>	72 (70)	58 (56)	82 (82)	69 (71)	73 (70)
(4) 7-Node Class. Tree <sup>f</sup>	74 (71)	70 (68)	77 (73)	69 (67)	78 (74)
<i>Classifications of women only</i>					
(0) No Model <sup>a</sup>	78 <sup>c</sup> (74) <sup>d</sup>	0 (0)	100 (100)	0 (0)	78 (74)
(1) Linear Regression	78 (75)	26 (39)	94 (88)	54 (53)	81 (80)
(2) Logistic Classification	78 (76)	13 (27)	96 (94)	52 (60)	79 (79)
(3) 2-Node Reg. Tree <sup>e</sup>	78 (76)	8 (16)	98 (97)	59 (65)	79 (77)
(4) 7-Node Class. Tree <sup>f</sup>	81 (76)	41 (48)	93 (85)	63 (53)	84 (82)
<i>Classifications of men only</i>					
(0) No Model <sup>b</sup>	62 <sup>c</sup> (64) <sup>d</sup>	100 (100)	0 (0)	62 (64)	0 (0)
(1) Linear Regression	63 (69)	91 (95)	18 (22)	64 (69)	54 (70)
(2) Logistic Classification	64 (68)	81 (84)	35 (38)	67 (71)	54 (57)
(3) 2-Node Reg. Tree <sup>e</sup>	65 (65)	77 (74)	46 (48)	70 (72)	55 (50)
(4) 7-Node Class. Tree <sup>f</sup>	67 (65)	81 (77)	44 (44)	70 (71)	59 (51)

<sup>a</sup> All classified as non-dyslipidemic (modal category). <sup>b</sup> All classified as dyslipidemic (modal category). <sup>c</sup> Resubstitution estimate for Vaud-Fribourg data (n = 1,120 (572 women, 548 men)). <sup>d</sup> (Cross-validation estimate based on Ticino data (n = 1,429 (741 women, 688 men))). <sup>e</sup> Used (WHR) only; same classifications as 4-node, 5-node, 6-node, 7-node, and 9-node regression trees, which used (WHR, BMI), and same variable and classifications as 3-node regression tree. (Also same variable and classifications as for 2-node, full model regression tree.) <sup>f</sup> Used (WHR, BMI) only. (Also same variables and classifications as for 7-node, full model classification tree.)

**Overall classification models**

The classification performance of the overall(both genders) models which included Gender as a predictor is summarized in Table 3. Each pruned regression and classification tree model listed was the smallest whose classification performance was equivalent to that of any larger tree. There were only minor differences in the predictor variables retained and the numbers of leaves between the

CART models selected for the Vaud-Fribourg and Ticino samples (not shown). Likewise, the rankings of the predictor variables by their relative (nominal) statistical significance in the linear and logistic regression models differed slightly for two samples and between model types (not shown). As expected, WHR and BMI were among the two or three most important predictor variables in all models. On the whole, the classification results for all

**Table 4: Classification performance of gender-specific reduced {WHR, BMI} predictive models.**

(Strategy) Fitted Model	Total % Correct	Sensitivity	Specificity	+ Predictive Value (PPV)	- Predictive Value (NPV)
<i>Model based on Vaud-Fribourg women (n = 572), cross-validated on Ticino women (n = 741).</i>					
(0) No	78 <sup>c</sup>	0	100	0	78
Model <sup>a</sup>	(74) <sup>d</sup>	(0)	(100)	(0)	(74)
(1) Linear	78	19	95	53	80
Regression	(76)	(33)	(91)	(55)	(79)
(2) Logistic	78	19	95	53	80
Classification	(75)	(32)	(91)	(54)	(79)
(3) 3-Node	80	40	92	59	84
Reg. Tree <sup>d</sup>	(75)	(45)	(86)	(52)	(82)
(4) 3-Node	81	38	93	62	84
Class. Tree <sup>e</sup>	(75)	(44)	(86)	(53)	(82)
<i>Model based on Vaud-Fribourg men (n = 548), cross-validated on Ticino men (n = 688)</i>					
(0) No	62 <sup>c</sup>	100	0	62	0
Model <sup>b</sup>	(64) <sup>d</sup>	(100)	(0)	(64)	(0)
(1) Linear	63	88	24	65	55
Regression	(69)	(91)	(30)	(70)	(65)
(2) Logistic	64	86	29	66	55
Classification	(68)	(88)	(33)	(70)	(60)
(3) 3-Node	65	78	45	70	56
Reg. Tree <sup>d</sup>	(66)	(76)	(47)	(72)	(52)
(4) 5-Node	68	78	51	72	59
Class. Tree <sup>f</sup>	(67)	(78)	(47)	(73)	(54)

<sup>a</sup> All women classified as non-dyslipidemic (modal category). <sup>b</sup> All men classified as dyslipidemic (modal category). <sup>c</sup> Resubstitution estimate. <sup>d</sup> (Cross-validation estimate). <sup>d</sup> Same variables and classifications as 3-node, full model regression tree. <sup>e</sup> Same variables and classifications as 3-node, full model classification tree. <sup>f</sup> Same variables and classifications as 5-node, full model classification tree.

models were consistent between the two regions. Thus for brevity, only the resubstitution results for the Vaud-Fribourg models with external validation on the Ticino subjects are shown.

For both genders combined, regardless of measure, classification performance was a modest 60–80% for all models, and no clear preference among different models was discernible. Moreover, the reduced models performed nearly as well as the full models. Again for brevity, only results for the reduced models are shown. Kappa measures of agreement were also calculated, indicating 70–80% classification concordance between models, with a slight tendency for the linear and logistic models on the one hand, vs. CART models on the other, to agree more among themselves (75–80%) than with models of the other type (70%) (not shown otherwise). This tendency was not evident for the regression-*per se* vs. classification-*per se* models.

The overall classification rates in Table 3 were not uniform by gender. For Vaud-Fribourg women, the models had

higher specificity and NPV, but lower sensitivity and PPV; for Vaud-Fribourg men these tendencies were reversed. Apparently, this "interaction" by gender was not "automatically detected" consistently nor particularly well by the overall tree-based models, none of which retained the Gender variable.

**Gender-specific classification models**

Classification performance for models fitted separately to each gender is shown in Table 4. The differences in classification rates relative to those of the corresponding overall models were at best uneven. The "improvements" of the 3-node, reduced model regression tree over the 2-node, reduced model regression tree (Table 3) for Vaud-Fribourg women notwithstanding, on balance any small to moderate gains in classification here (e.g., in sensitivity) were met by losses there (e.g., in specificity) for all types of model for both regions.

There were more inconsistencies in the predictor variables retained by the gender-specific CART models compared to the overall CART models between the two regions,

especially for men (not shown). These inconsistencies were due in part to the necessarily smaller gender-specific sample sizes, as well as to idiosyncrasies in the observed sample data for the two regions (Figures 1, 2).

## Discussion

In another study comparing Swiss and Seychelles Islands populations [9], several indicators of central adiposity (*i.e.*, waist circumference and WHR) worked reasonably well when employed in logistic regression models for predicting dyslipidemia, either as individual predictors or in conjunction with other variables such as those employed in the present study. The predictive value of WHR for the Swiss populations served to corroborate the findings of Reeder *et al.* [8] in a Canadian population in the sense that similar variables and logistic models were employed in both studies.

Both of the latter studies attempted to quantify the predictive power of anthropometric measurements as first stage population screening indicators of dyslipidemia. However, neither study was particularly thorough in choosing the statistical methodology for the predictive models. For example, the (main) dependent variable, TC/HDL-C, although continuous, was analyzed as a binary variable with additive logistic regression models. Likewise, WHR and BMI, also continuous, were coded and employed in the logistic models as so-called "action level" dichotomies [1] (*e.g.*, WHR  $\geq 0.90$  for men or WHR  $\geq 0.80$  for women was coded as "high" WHR by gender, BMI  $\geq 27$  was coded as "high" BMI for both genders, and "high" was contrasted with "low" WHR or BMI in the models). Thus, we wondered if more comprehensive statistical models would have led to improved classification.

The present findings are based on juxtaposing the results for the very simplest additive, algebraic, linear and logistic regression *vs.* the non-additive, non-algebraic CART models based on the relatively small set of predictor variables examined. They serve to some degree to indicate the limits of predictability of dyslipidemia by first stage population screening programs based on statistical models which focus primarily or exclusively on anthropometric measurements such as WHR and BMI. The observed relationships between the latter and our TC/HDL-C ratio-based dyslipidemia continuous or binary variables were at best moderately strong, hence dyslipidemia was only moderately predictable therefrom. Nonetheless, although their predictive power is far from perfect, even the models for first stage population screening purposes such as those studied here could lead to potential cost savings. This conclusion did not seem to depend on the TC/HDL-C  $\geq 5.0$  cutpoint we employed to define dyslipidemia, as the data suggested that the relationship is stable within the limits of a reasonable change.

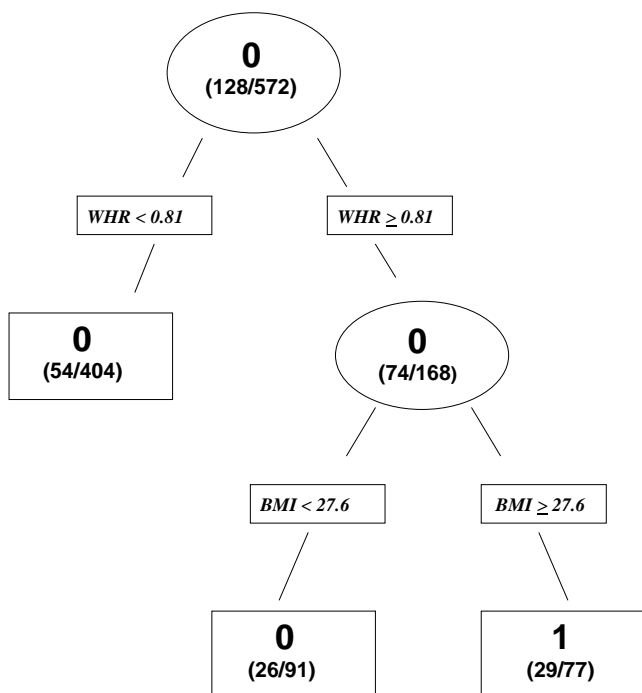
Our reliance on the composite WHR and BMI measures in our models instead of the individual waist, hip, weight, and height measurements may not have optimally or even adequately captured the relationship between the latter variables and the TC/HDL-C ratio. However, our rationale was to investigate and attempt to improve upon the types of classification rules intended for use in population dyslipidemia screening that have been obtained in previous studies employing similar but more limited analytical approaches. BMI and WHR are routinely employed because they are directly related to clinical entities (*i.e.* peripheral overweight, central obesity, etc.). Moreover, the issue of partial relationship was addressed by examining models using waist circumference alone instead of WHR, but we found little difference in the results (not shown). These potential limitations notwithstanding, external validation has recently been shown to be crucial for judging the merits of any predictive model [19,20]. The external validations of the various models estimated from the two different Swiss MONICA samples did provide some evidence of their predictive stability in these populations.

The overall (both genders) sensitivities and specificities of the various predictive models for the Swiss samples in this study were comparable to those obtained using only logistic regression models with WHR and BMI as coded by Reeder *et al.* [8] in Canadian samples, and by Paccaud *et al.* [9] in samples from Switzerland and the Seychelles. However, discrepancies in these measures and reversals by gender were more pronounced in the present study. It may be that our use of continuous versions of these variables in the models led to these differences.

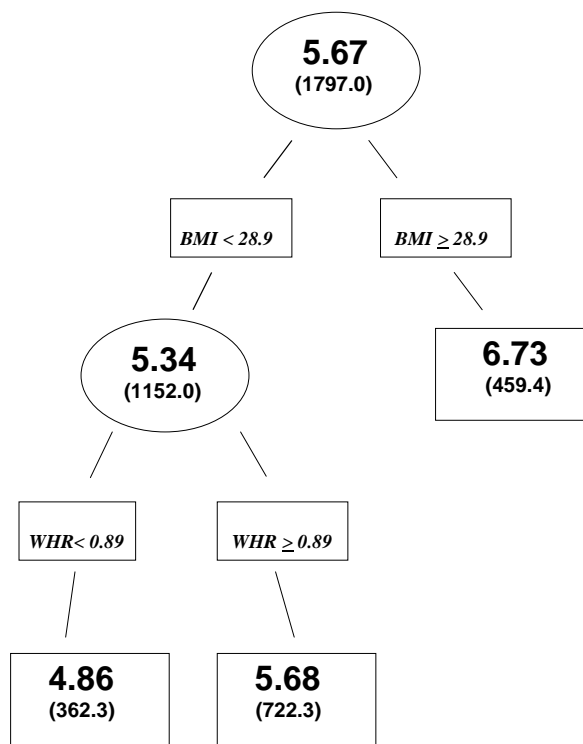
The forward (backward) variable selection process inherent in full (pruned) CART modeling differs in an important way from the stepwise selection procedures that are commonly used with linear and logistic models. That is, a predictor variable selected for binary splitting at a given step may be "re-selected" at subsequent steps, or even "re-removed" as at previous steps. In essence, this difference is what makes tree-based models so-called "automatic interaction detectors" [21], and also why it is difficult to pre- or even post-specify tree-based models algebraically, but fortunately (perhaps) it is not necessary to do so to apply them in practice. A major feature of this approach is that no assumption of linearity between *Y* and the predictor variables (which can be categorical (binary or polychotomous), ordinal, or continuous) is required. Tree-based models are obviously appealing because of these features.

Despite the expected advantages of CART models over their linear and logistic counterparts (also see [22]), as well as the evidently modest ability of WHR and BMI to





**Figure 3**  
3-node classification tree for Vaud-Fribourg women (n = 572) (gender-specific reduced {WHR, BMI} model in Table 4). Ovals: interior nodes; rectangles: terminal nodes (leaves). Numbers inside nodes are predicted values (+ corresponding misclassification rates). Binary classification rule: 1: predict dyslipidemia; 0: predict no dyslipidemia.



**Figure 4**  
3-node regression tree for Vaud-Fribourg men (n = 548) (gender-specific reduced {WHR, BMI} model in Table 4). Ovals: interior nodes; rectangles: terminal nodes (leaves). Numbers inside nodes are estimated mean values of TC/HDL-C (+sums of squares about the mean values). Binary classification rule: TC/HDL-C ≥ 5.0, predict dyslipidemia; TC/HDL-C < 5.0, predict no dyslipidemia.

predict dyslipidemia, we were somewhat disappointed with the comparative classification performance of the CART models for these particular data, especially because we had deliberately "handicapped" the linear and logistic modeling strategies by not applying any formal predictor variable selection methodology and by considering only strictly additive models.

On the other hand, the CART models did provide some corroboration of and further insights regarding the above-mentioned "action levels" for WHR and BMI employed in the logistic models of Reeder *et al.* [8] and Paccaud *et al.* [9]. For example, consider the 3-node classification tree for Vaud-Fribourg women shown in Figure 3, and the 3-node regression tree for Vaud-Fribourg men shown in Figure 4. A woman whose WHR ≥ 0.81 and (then) whose BMI ≥ 27.6 would be classified as dyslipidemic (*i.e.*, estimated  $Y = 1$ ). A man whose BMI ≥ 28.9 would immediately be classified as dyslipidemic (*i.e.*, predicted  $Y = 6.73 \geq 5.0$ ), while a man whose BMI < 28.9 but (then) whose WHR ≥ 0.89 would also be classified as dyslipidemic (*i.e.*,

predicted  $Y = 5.68 \geq 5.0$ ). The cutpoints in these 3-node CART models are similar to the previous "action-levels", but are employed a bit differently for classification purposes depending on gender. Such details were much less apparent in the linear and logistic models.

Some additional improvements might have been obtained by incorporating differential costs of misclassification into the classification-tree (also logistic) models. However, these costs are not always easy to specify. This issue can alternatively be addressed indirectly by changing the (usual default) classification cut-off point from 0.50 to (say)  $p_s$  = sample prevalence of dyslipidemia, and (in effect) classifying an individual as dyslipidemic only if their model-estimated *posterior* probability of being dyslipidemic exceeds their *prior* probability of being dyslipidemic (*i.e.*,  $p_s$ ). This latter approach was examined in the present study, but on balance the corresponding classification performance results were not much different from

those based on the usual 0.50 cut-off point (not shown otherwise). This was due at least in part to the fact that the observed values of  $p_g$  (see Table 1) were not close to the extremes of 0 or 1. Of course, changing the cut-off point in this manner simply implies trade-offs between sensitivity and specificity, which may or may not be warranted depending on the actual costs of misclassification.

## Conclusions

At least for binary prediction of dyslipidemia from waist-to-hip ratio and body mass index in the context of the relatively small set of other predictor variables examined, the simple additive logistic models obtained in previous studies were about as effective as the more comprehensive statistical models investigated here. Indeed, for the data at hand, perhaps even an old standby such as linear discriminant analysis [23], the forerunner of logistic classification, would have sufficed. In all fairness, CART models may be of more value when much larger sets of predictor variable main effects and interactions than the one considered in this study are considered in the statistical modeling process.

## Competing interests

None declared.

## Authors' contributions

FP was responsible for obtaining the data. MCC analyzed the data and drafted the manuscript. Both authors designed the comparative study, interpreted the results, and wrote the final paper.

## References

- Han TS, Van Leer EM, Seidell JC, Lean MJ: **Waist circumference action levels in the identification of cardiovascular risk factors: prevalence study in a random sample.** *BMJ* 1995, **311**:1401-1405.0.
- Reeder BA, Senthiselvan A, Despres JP, Angel A, Liu L, Wand H, Rabin SW: **The association of cardiovascular disease risk factors with abdominal obesity in Canada. Canadian Heart Health Surveys Research Group.** *CMAJ* 1997, **157**:S39-S45.
- Shetterly SM, Marshall JA, Baxter J, Hamman RF: **Waist-hip-ratio measurement location influences associations with measures of glucose and lipid metabolism. The San Luis Valley Diabetes Study.** *Ann Epidemiol* 1993, **3**:295-299.
- Bjorntorp P: **Regional patterns of fat distribution.** *Ann Int Med* 1985, **103**:994-995.
- Seidell JC, Cigolini M, Charzewska J, Ellsinger BM, di Base G: **Fat distribution in European women: a comparison of anthropometric measurements in relation to cardiovascular risk factors.** *Int J Epidemiol* 1990, **19**:303-308.
- Pouliot MC, Després JP, Lemieux S, Moorjani S, Bouchard C, Tremblay A, Nadeau A, Lupien PJ: **Waist circumference and abdominal sagittal diameter: best simple anthropometric indexes of abdominal visceral adipose tissue accumulation and related cardiovascular risk in men and women.** *Am J Cardiol* 1994, **73**:460-468.
- Houmard JA, Wheeler WS, McCammon MR, Holbert D, Israel RG, Barakat HA, Wells JM, Truitt N, Hamad SF: **An evaluation of waist to hip ratio measurement methods in relation to lipid and carbohydrate metabolism in men.** *Int J Obes* 1991, **15**:181-188.
- Reeder BA, Liu L, Horlick L: **Selective screening for dyslipidemia in a Canadian population.** *J Clin Epidemiol* 1996, **49**:217-222.

- Paccaud F, Schlüter-Fasmeyer V, Wietlisbach V, Bovet P: **Dyslipidemia and abdominal obesity: An assessment in three general populations.** *J Clin Epidemiol* 2000, **53**:393-400.
- Chambers JM: **Linear models.** In *Statistical Models in S Volume 4*. Edited by: Chambers JM, Hastie TJ. Wadsworth & Brooks/Cole, Pacific Grove, CA; 1992:95-144.
- Hastie TJ, Pregibon D: **Generalized linear models.** In *Statistical Models in S Volume 6*. Edited by: Chambers JM, Hastie TJ. Wadsworth & Brooks/Cole, Pacific Grove, CA; 1992:195-248.
- Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees* Wadsworth, Belmont, CA; 1984.
- Clark LA, Pregibon D: **Tree-based models.** In *Statistical Models in S Volume 9*. Edited by: Chambers JM, Hastie TJ. Wadsworth & Brooks/Cole, Pacific Grove, CA; 1992:377-420.
- SAS Institute Inc: *SAS OnlineDoc®, Version 8, Cary, North Carolina, USA* 1999.
- Insightful Corp: *S-PLUS 2000 Guide to Statistics*, Seattle, WA 1999.
- World Health Organization MONICA Project Principal Investigators: **The MONICA Project (MONitoring trends and determinants in Cardiovascular disease): a major international collaboration.** *J Clin Epidemiol* 1988, **41**:105-114.
- Venables WN, Ripley BD: *Modern Applied Statistics with S-Plus* 3rd edition. Springer, NY; 1999:327.
- Wietlisbach V, Paccaud F, Rickenbach M, Gutzwiller F: **Trends in cardiovascular risk factors (1984-1993) in a Swiss region: results of the three population surveys.** *Prev Med* 1997, **26**:523-533.
- Terrin N, Schmid CH, Griffith JL, D'Agostino RB Sr, Selker HP: **External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks.** *J Clin Epidemiol* 2003, **56**:721-729.
- Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, Moons KGM: **External validation is necessary in prediction research: A clinical example.** *J Clin Epidemiol* 2003, **56**:826-832.
- Sonquist JA, Morgan JN: **The detection of interaction effects: A report on a computer program for the selection of optimal combinations of explanatory variables.** *Monograph 35, University of Michigan, Ann Arbor: Survey Research Center Institute for Social Research* 1964.
- Cook EF, Goldman L: **Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis.** *J Chron Dis* 1984, **37**:721-731.
- Fisher RA: **The use of multiple measurements in taxonomic problems.** *Ann Eugenics* 1936, **7**:179-188.

## Pre-publication history

The pre-publication history for this paper can be accessed here:

<http://www.biomedcentral.com/1471-2288/4/7/prepub>

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

