Author's response to reviews

Title: Empirical Comparison of Linear, Logistic, and CART Models for Binary Classification of Dyslipidemia from Anthropometric Measurements

Authors:

Prof Michael C Costanza (Michael.Costanza@hcuge.ch)
Fred Paccaud (Fred.Paccaud@inst.hospvd.ch)

Version: 2 Date: 1 Feb 2004

RESPONSES TO REVIEWER (1)

Reviewer's report
Title: Empirical Comparison of Linear, Logistic, and CART Models for Binary Classification of Dyslipidemia from Anthropometric Measurements
Version: 1
Date: 17 January 2004
Reviewer: Stefan s Ma

Reviewer's report:

This is a well-written paper. The objectives of the paper are 1) to improve on previous statistical strategies for detecting dyslipidemia in the general population; and 2) to compare the performance of four statistical modeling approaches. The study was able to achieve these objectives supported by the empirical findings. However, since the performance of the four modeling approaches are very similar using empirical data, in terms classification performance criteria, i.e. sensitivity, specificity and both positive and negative predictive values, one lay person's concern is that what is the recommendation of choosing which modeling approach if they really like to apply it in their own setting. The question is that "is there a need to supplement the comparison by simulation study?" because we can simulate data for various situation in which the empirical data do not have.

What next?: Accept without revision
Level of interest: An article of limited interest
Quality of written English: Acceptable
Statistical review: No
Declaration of competing interests: none.

RESPONSE:

We think that a simulation study comparing the different modeling techniques would be useful in a situation where one had a (much) larger set of predictor variables, and thus a greater variety of model types could be feasibly entertained (e.g., with two- and higher-way interactions, or nonlinearity, etc.). However, our intention here was not that ambitious. Rather, we decided to stick with the relatively small set of predictors that had been used in previous studies and to confine our empirical comparisons within that simpler context. We thought that the corresponding loss in generality was balanced by the opportunity to compare the simplest additive logistic and linear models with the more comprehensive CART models. Incidentally, the end of the Conclusions now reads:

"At least for predicting dyslipidemia from WHR and BMI in the context of the small set of other predictor variables examined, we conclude that the simple additive logistic models obtained in
previous studies were about as effective as the more comprehensive statistical models examined 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 most value when much larger sets of predictor variables are considered. This may also be the case for other modeling techniques such as …”.

RESPONSES TO REVIEWER (2)

Reviewer's report
Title: Empirical Comparison of Linear, Logistic, and CART Models for Binary Classification of Dyslipidemia from Anthropometric Measurements
Version: 1
Date: 10 November 2003
Reviewer: Takamaru Ashikaga

Reviewer's report:

General

1-The manuscript represents a straight forward comparison of several competing approaches that attempt to explore interrelationships between a single outcome measure and multiple predictors in an attempt to explore if some items might be used as a preliminary screening set for dyslipidemia.

2-This is a well written manuscript that is easy to follow although several items are not presented. See item 7 and 8 below.

RESPONSE: We decided to keep the omitted items out; please see our responses to 7 & 8 below.

3-The conclusions are of value to those interested in general statistical methodology and those who are interested in two-stage types of population screening using correlates of a primary screening measure. The lack of any major differences in approaches is noteworthy. The bottom line comparison of classification into groups appears very reasonable as is the cross validation effort.

RESPONSE: Please see our response to 6 below.

Discretionary Revisions (which the author can choose to ignore)

4-The authors might comment on the availability of the data itself and if such might be made available on a web site if such is possible for other investigators.

RESPONSE: We think it best to leave that up to individual reader requests.

Minor Compulsory Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

5-The results seem to indicate that there are not any major differences between the methods examined. Could the authors comment on the argument that CART is really for use with many more potential predictors and that the data structure for the CART approach is very limited as far as a CART application is concerned. In addition while the model based approaches are very simple, all of the data are being used to a degree with these models.
The end of the Conclusions now reads:
"At least for predicting dyslipidemia from WHR and BMI in the context of the small set of other predictor variables examined, we conclude that the simple additive logistic models obtained in previous studies were about as effective as the more comprehensive statistical models examined 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 most value when much larger sets of predictor variables are considered. This may also be the case for other modeling techniques such as ...

In addition, the Abstract Conclusions now read:
"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. 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 larger sets of predictor variables."

Also, we had already said (now page 4) "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."

6-The cross validation effort conducted does differ to a degree from other possible validation efforts given two available data sets. The authors should be commended for attempting such an effort overall. They may wish to comment upon some alternatives such as sample splitting from both regions to develop the initial results and then conducting validation of the derived equations and algorithms using the remaining half-samples from both regions. This approach is clearly not possible when the models are developed prior to their application in another geographic area.

RESPONSE:
Please note that we have added two recent references [19, 20] which emphasize the importance of external validation (which term now replaces "cross-validation" throughout the text). Accordingly, we think that sample splitting is just not as good as external validation, so we have not explicitly mentioned it. On the other hand, within the CART models, internal 9:1cross-validation was used because it was included in the particular programs we used (see relevant Methods subsections).

7-The contrast between a linear regression approach and a logistic regression approach seems very natural. However, the issue of developing a model based upon a continuous outcome and then using predicted values to formulate ordered categories or a dichotomy does not appear to be as direct as formulating the categories using the dependent measure initially and then using a multinomial or basic logistic regression model based approach. Thus the use of an estimated multiple linear regression model based value to categorize the dependent measures adds the issue of multiple linear regression prediction accuracy into the discussion. This area can be addressed with the addition of some of the usual measures of goodness of fit for the models provided. This would include the usual multiple correlation coefficient for the multiple linear regression model as well as the Hosmer-Lemeshow type goodness of fit or McFaden's rho square for the logistic model. However the multiple correlation coefficient may not tell the full story about the ability to predict values of the TC/HDL-C ratio in the region of 5.0 unless it is very near unity. In addition to the goodness of fit measures, it would be nice to see the regression coefficients and their standard errors since these are also usually provided in when these models are estimated.

RESPONSE:
Forming the dyslipidemia groups from the continuous outcome first and then applying a logistic approach is essentially what was done in Strategy 2. Whereas in Strategy 1, we simply wanted to see what would happen if a (by definition) more comprehensive linear regression prediction model was THEN used for binary classification. Although more detail on the regression coefficients and
correlations might be of some interest, we were mostly interested in the end result: i.e., did it work better ultimately? In the interest of brevity, we decided not to include those additional details.

8-Some of the results are not presented. Since there are no space limits, these should be included.

RESPONSE:
In our judgement, nothing would be gained by presenting the other results because they were so similar. Hence, we decided to keep them out.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

What next?: Accept after minor compulsory revisions
Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable
Statistical review: No
Declaration of competing interests: None
RESPONSES TO REVIEWER (3)

Reviewer's report
Title: Empirical Comparison of Linear, Logistic, and CART Models for Binary Classification of Dyslipidemia from Anthropometric Measurements
Version: 1
Date: 3 December 2003
Reviewer: Tommy Visscher

Reviewer's report:

General

The first objective of Costanza and Paccaud was to study whether there are better strategies to predict levels of cholesterol in a general population than previous statistical strategies, with specific focus on waist-hip ratio (WHR) and body mass index (BMI). The second objective was to compare the performance of four statistical modeling approaches with regard to a binary distribution of cholesterol. Please note that this reviewer is an epidemiologist rather than a statistician. I apologize at forehand, if I make wrong interpretations.

The authors state that predicting levels of cholesterol is important because impaired cholesterol levels determine cardiovascular risk. Of course there is no doubt that impaired cholesterol levels are important in predicting cardiovascular risk. However, this reviewer feels that it is not useful to predict cholesterol levels from BMI and WHR. In practice, cardiovascular risk is estimated from levels of body weight and cholesterol, next to smoking and blood pressure levels. Predicting 'cholesterol levels' is, to my opinion, not relevant when cardiovascular risk is already predicted from the same measures that are available. 'Predicted' cholesterol levels will then not add to a better predicted cardiovascular risk.

RESPONSE:
The (partial) statement that "In practice, cardiovascular risk is estimated from levels of body weight and cholesterol..." seems to beg the question. In addition, it is not perfectly clear that '"Predicted' cholesterol levels will then not add to a better predicted cardiovascular risk". In any case, we have rephrased last sentence of the first paragraph of the Background as follows to address this issue: "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”.
Also, we were thinking more in terms of an application to a population-wide screening program, for
which even the relatively cheap cost of dry-chemistry cholesterol measurements (e.g., using a
Reflotron), as well as the (short but not zero) waiting times for the results, are important issues.

Further, if a clinician would be interested in cholesterol levels per se in specific situations, for
instance when cholesterol levels are especially impaired without the other cardiovascular risk factors
being impaired, it seems to be important enough to measure cholesterol levels, rather than
predicting them from WHR or BMI.

This reviewer is sorry to report that for this reason the paper does not seem relevant for general
readers or readers with interest for cardiovascular diseases.

RESPONSE:
We reiterate that we were thinking more in terms of applications to population-wide screening
programs, for which even the relatively cheap cost of dry-chemistry cholesterol measurements (e.g.,
using a Reflotron), as well as the (short but not zero) waiting times for the results, are important
issues. In that sense it is only after the initial screening that more careful measurement of an
individual subject’s cholesterol may be required.

This reviewer has doubts on whether the second research aim has been met to study four modeling
approaches regard binary levels of cholesterol. If I understand the methods section well, the authors
assessed four models of which only two on a binary distribution of cholesterol (< 5.0 vs >5.0) and two
on continuous levels. If the latter is true, the title is wrong. ‘for binary classification of dyslipidaemia’. Page 5, strategy 2 (‘for the same Xi predictors in strategy 1, but with binary Y) are compatible with
my idea that not all models were on a binary distribution of cholesterol.

RESPONSE:
In the sense (as explained several times in the text) that the ultimate use of the models, whether
based initially on a binary outcome or on a continuous one, was assumed to be binary prediction, the
title is correct and the second research aim has been met. We are thinking here in the "dichotomous
spirit" of whether or not a physician decides to treat a patient, regardless of the dosage.

If all models were indeed on binary distributions of cholesterol, the research focus would be even
less interesting. A physician should be interested in the absolute level of cholesterol, not only in
whether cholesterol levels are impaired or not.

RESPONSE:
We tried to account for this viewpoint by using the continuous measures more comprehensively than
they had been in the previous studies cited. The previously published models [1-9] were based on
binary outcomes alone.

The conclusion in the last sentence that 'even an odd standby such as.......would have sufficed' is
disappointing. This reviewer appreciates that the outcome of a study should not influence the
decision on publication. However, the question rises why the authors hypothesized that the studied
models would be better. This is not clear. ‘Other’ statistical strategies are only mentioned in the very
last sentence of this paper.

RESPONSE:
We realize that we may have seemed to have been unfair to the "automatic interaction detection"
CART models which, in theory, ARE more comprehensive, because our example was relatively
simple. To redress this oversight, we have rewritten the last part of the Conclusions as follows:
“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 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 > 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 > 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.

At least for predicting dyslipidemia from WHR and BMI in the context of the small set of other predictor variables examined, we conclude that the simple additive logistic models obtained in previous studies were about as effective as the more comprehensive statistical models examined 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 most value when much larger sets of predictor variables are considered. This may also be the case for other modeling techniques such as generalized additive models [24], multivariate adaptive regression splines [25], neural networks [26], optimal discriminant analysis [27], and hierarchically optimal classification tree analysis [28]. However, because the latter were either not as readily available or as user-friendly as they could be, we excluded them from our study”.

Discretionary Revisions (which the author can choose to ignore)

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

Page 4: Smoking (smoke): readers are not interested how the authors named their variables.

RESPONSE:
This oversight has been changed throughout.

As an epidemiologist, I would prefer the use of 'independent variables', or related terminology to 'Xi predictors'.

RESPONSE:
Except in the model equations, we now use the term "predictor variables" throughout..

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

I have some other remarks:

Page 3: 'central obesity is studied by use of the WHR.' It should be realized that central obesity is assessed by levels of waist circumference alone. Fat distribution is assessed by WHR. Indeed, WHR levels may indicate cardiovascular risk better than waist circumference alone. This issue may
deserve attention in the discussion, as both large waist circumferences and low hip circumference seem to affect disease risk.

RESPONSE:
On page 3 we say: "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]."

In the Conclusions we now say: "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).

Page 4: Two MONICA samples are used, one with age-range 25-74, one with age-range 35-64. For comparability reasons analyses were restricted to age-range 35-64. If the two samples are comparable then, why are the analyses performed for the samples separately?

RESPONSE:
This is a matter of semantics. Yes, the two samples are "comparable" (e.g., in the senses of Tables 1 & 2), but they are also independent and from different regions, hence also "different" enough to merit being analyzed separately.

Page 4: We applied five modeling approaches minimal benchmarks strategy.: This is unclear for non-statisticians without reading the explanations. I would advice changing the text into: five models were tested followed by the text starting from Stragey 0.

RESPONSE:
We say: "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." We think it is clear by the designation "0" that the first strategy was a just a "fall-guy". And yes, we recommend that the reader actually read what we wrote here!

Page 4: ' No model'. Non-statisticians could be mislead by this terminology.

RESPONSE:
The key word here is "could". We think that our explanation: "Strategy 0 represented a "no model" approach in the sense that the additional predictor variables were ignored" is sufficiently clear not to mislead the reader.

Page 5: HBP: Why is blood pressure taken as binary variable, rather than as a continuous?

RESPONSE:
This was a simplification along the lines of the BMI & WHR "action levels" used previously by other researchers. In a very comprehensive model, all continuous variables would be analyzed as such.

Page 5: estimated at Y >5.0
1) Does this mean 'estimated' from the model? Than readers would be interested in \( \beta \)-coefficients rather than \( b_0 \)-k. I have the impression that 'measured' is meant here. Consider avoiding use of 'estimated' then.

RESPONSE:
Yes, estimated, not measured. Also, please see our response to comment 7 of Reviewer (2).

2) Why it is important to mention here that \( Y > 5.0 \) is considered as dyslipidaemia. If I get it right, continuous levels are used here.

RESPONSE:
Yes, but even for the regression models we are assuming that a dichotomous (binary) outcome is what needs to be predicted.

\( P > 0.50 \) was classified as dyslipidaemic: This is unclear to this reviewer. I assume that authors KNOW cholesterol levels. I start doubting whether the authors have MEASURED cholesterol levels. Indeed, it is not described how (and whether) cholesterol levels are measured. If cholesterol levels are indeed not measured but solely based on prediction formulas, more clarity is definitely needed on producing cholesterol levels. Table 1 presents cholesterol levels. How are these values produced?

RESPONSE:
All cholesterol was measured (see reference [16]). Of course, one develops the prediction models based on the observed measurements. The key point here is how the predictions are evaluated. If one relies solely on the same data used to develop the model to also evaluate it, serious bias results. That is why we relied mainly on the external validations (develop model in one sample, evaluate on the other, and vice versa) for that purpose (please see also our response to comment 6 of Reviewer 2).

This model assumes the relationship.....is linear:

Why stating that the model is linear here and not making a statement on linearity regarding strategy 2. Consider consistency. Either make the statement on both or no strategies.

RESPONSE:
There was/still is such a statement for Strategy 2: "This model assumes the relationship between \( \log(p/(1-p)) \) and the predictor variables is linear." But this is not the same type of linearity as assumed in Strategy 1.

Page 6, first paragraph.
This paragraph is unclear without looking at the trees.
This reviewer feels it is important to mention how cut-off points in the tree structures have been defined.

RESPONSE:
We intended for the reader to be looking at the figures (and the figure legends) while reading this. We don't see how to be more clear than that. The process by which the cut-offs were defined is described in the subsections on the CART Strategies 3 & 4 in the Methods.

Page 7: Switzerland has.....:
As I know the MONICA study, samples are not necessarily representative of the countries. Are the studied samples representative for Switzerland?

RESPONSE:
No, but they are representative of the two Swiss regions (e.g., one is French-speaking, the other is Italian-speaking, as mentioned in the text).

Page 7: bivariate relationship patterns ALSO were similar. It is unclear where 'also' refers to

RESPONSE:
It refers to the observation that, just as the univariate distributions were similar between the two regions (Table 1, previous paragraph), the correlations ALSO were similar between the two regions (Table 2, and as further explained in the same paragraph).

The first sentence of the conclusion paragraph is disappointingly weak: We wonder less now......more effective.

RESPONSE:
The penultimate paragraph of the Conclusions now reads:
"At least for predicting dyslipidemia from WHR and BMI in the context of the small set of other predictor variables examined, we conclude that the simple additive logistic models obtained in previous studies were about as effective as the more comprehensive statistical models examined 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."

Also, the Abstract Conclusions now read:
"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. 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 larger sets of predictor variables."

What next?: Reject because scientifically unsound
Level of interest: Too insignificant to warrant publication in any journal
Quality of written English: Acceptable
Statistical review: Yes
Declaration of competing interests: none