Modified Classification Trees applied to Pediatric Familial Hypercholesterolemia: a comparative study including Simon Broome Criteria

João Albuquerque, MSc[1,2]; Ana Catarina Alves, PhD[3,4]; Ana Margarida Medeiros, Msc[3,4] Mafalda Bourbon, PhD[3,4]; Marília Antunes, PhD[1]

1 - Centro de Estatística e Aplicações, FCUL, Portugal.

2 - Departamento de Bioquímica, FMUP, Portugal.

3 - Cardiovascular Research Group, INSA, Lisboa, Portugal.

4 - Instituto de Biossistemas e Ciências Integrativas – BioISI, FCUL, Portugal.

Corresponding author: João Albuquerque, joaodavid.alb@gmail.com

Keywords: Familial Hypercholesterolemia; Decision Tree models; Simon Broome Criteria.



**Introduction:** Familial Hypercholesterolemia (FH) is an autosomal dominant disorder of lipid metabolism, characterized by elevated plasmatic cholesterol (TC) concentrations, in particular low density lipoprotein cholesterol (LDLc). The resulting severe dyslipidemia leads to early development of atherosclerosis, which represents a major risk factor for cardiovascular disease (CVD). Early diagnosis of FH has been associated with a significant reduction in CVD risk, supporting the introduction of precocious and/or more aggressive therapeutic measures. Simon Broome (SB) criteria for the diagnostic of FH are among the most frequently used in clinical setting [1]. When compared to genetic test results however, clinical diagnosis criteria present a high false positive rate, which poses a serious problem, since the correct identification of dyslipidemia etiology is crucial for the assessment of CVD risk and therapeutic approach.

**Purpose:** The main purpose of this work was to develop a classification model for FH based on a modified version of the classic decision tree (DT), using several biochemical markers as predictor variables. Two different DT models were compared with SB clinical criteria in terms of accuracy and efficiency.

**Methods:** The sample used in this study was constituted by 260 participants of the Portuguese FH Study of both sexes, at pediatric age (2 to 17 years). Patients met the clinical criteria for dyslipidemia, and were not under hypolipidemic medication at referral. All participants had an informed consent form signed by the legal guardian, and information was approved by the National Data Protection Commission. Plasmatic concentrations of TC, LDLc, high density lipoprotein cholesterol (HDLc), triglycerides (TG), apolipoproteins AI (apoAI) and B (apoB), and lipoprotein(a) (Lp(a)), were determined, in mg/ dL. Genomic DNA was extracted and molecular diagnosis was performed through the study of LDLR, APOB and PCSK9 genes. Entropy and information gain measures were calculated for all biochemical variables. A modified version of the DT method [2], consisting in the sequential exclusion of predictor variables as they are used in each tree node, was implemented. Two different trees, using all 7 (DT7) and only 4 (DT4) biomarkers (TC, LDLc, HDLc and TG) were built, and pruned using *rpart* algorithm to avoid overfitting. The reason to build DT4 was the fact that it uses only the more common and accessible biomarkers, and is therefore cost-effective in case of similar performance between both models. The modified DT models were compared with the biochemical values used in SB criteria for FH diagnosis in pediatric subjects (LDLc>155.0 mg/dL and TC>260.0 mg/dL). The bootstrap resampling method was used to assess the models' performance, with a total of 100 bootstrap samples generated from the original dataset. In all bootstrap samples,

a confusion matrix was obtained for each classification method, by comparison with molecular study results. Median and mean values of correspondent operating characteristics were used for performance comparison: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

**Results:** DT-based models showed increased accuracy, specificity and PPV compared to SB criteria. This means that DT models correctly classify the patient more often than SB criteria on one hand, and also that they have better ability to exclude negative cases. On the other side, SB criteria presented better sensitivity and NPV, suggesting higher ability to retain positive cases (Table 1.). Given the conservative cut-off values in SB criteria, this may be accomplished at the expense of retaining a high number of false positive cases, which can prove to be costly and inefficient in clinical practice. Between both DT methods, it was observed that classification rules are similar at the top nodes of the tree. In both cases, LDLc was the most informative variable, with a cutpoint of 167 mg/dL, which makes sense since this disease primary affects LDLc metabolism. Also in both cases, TG was suggested as the following reference variable for patients with an LDLc≥167 mg/dL, with a cutpoint of 70 mg/dL. In this case, low TG values are related to the presence of FH, which can be understood from the stand point that high TG concentrations are probably related to dyslipidemia triggered essentially by environmental factors. At the bottom nodes the trees began to differ. In DT7, the variables ApoAI, ApoB and HDLc are pointed out, and in DT4, both TC and HDLc are used, by hierarchical order (Figure 1.). This may suggest that apolipoprotein values may be more informative than cholesterol fractions alone, and should therefore be used for diagnostic purposes if available.

**Conclusion:** Overall, DT classification methods seem to be a viable alternative to traditional clinical criteria for FH diagnosis. The modified version of DT models allows simplifying its interpretation for clinical practice, since each predictor variable is only used once.

**Acknowledgements:** Research supported by the programme Norte2020 (operação NORTE-08-5369-FSE-000018) and by national FCT funds under the projects UID/MAT/00006/2019 and PTDC/SAU-SER/29180/2017.

**References:** [1] Scientific Steering Committee on behalf of the Simon Broome Register. BMJ. 1991;303:893-6; [2] Breiman L. Classification and regression trees. Routledge; 2017;

Table 1. Median, mean and standard deviation of operating characteristics for the SB, DT4 and DT7 criteria, over 100 bootstrap samples.

| | DT7 method [c] | | | DT4 method [c] | | | SB method [d] | | |
|---|---|---|---|---|---|---|---|---|---|
| | Median | Mean | sd [e] | Median | Mean | sd [e] | Median | Mean | sd [e] |
| Accuracy | 0.911 | 0.912 | 0.02 | 0.901 | 0.902 | 0.02 | 0.728 | 0.726 | 0.03 |
| Sensitivity | 0.865 | 0.867 | 0.03 | 0.797 | 0.797 | 0.04 | 0.976 | 0.975 | 0.02 |
| Specificity | 0.933 | 0.933 | 0.02 | 0.954 | 0.953 | 0.01 | 0.609 | 0.606 | 0.04 |
| VPP [a] | 0.861 | 0.862 | 0.03 | 0.894 | 0.891 | 0.03 | 0.547 | 0.545 | 0.04 |
| VPN [b] | 0.937 | 0.936 | 0.02 | 0.910 | 0.907 | 0.02 | 0.981 | 0.981 | 0.01 |

a) Positive Predictive Values; b) Negative Predictive Value; c) Decision Tree; d) Simon Broome; e) Standard Deviation.

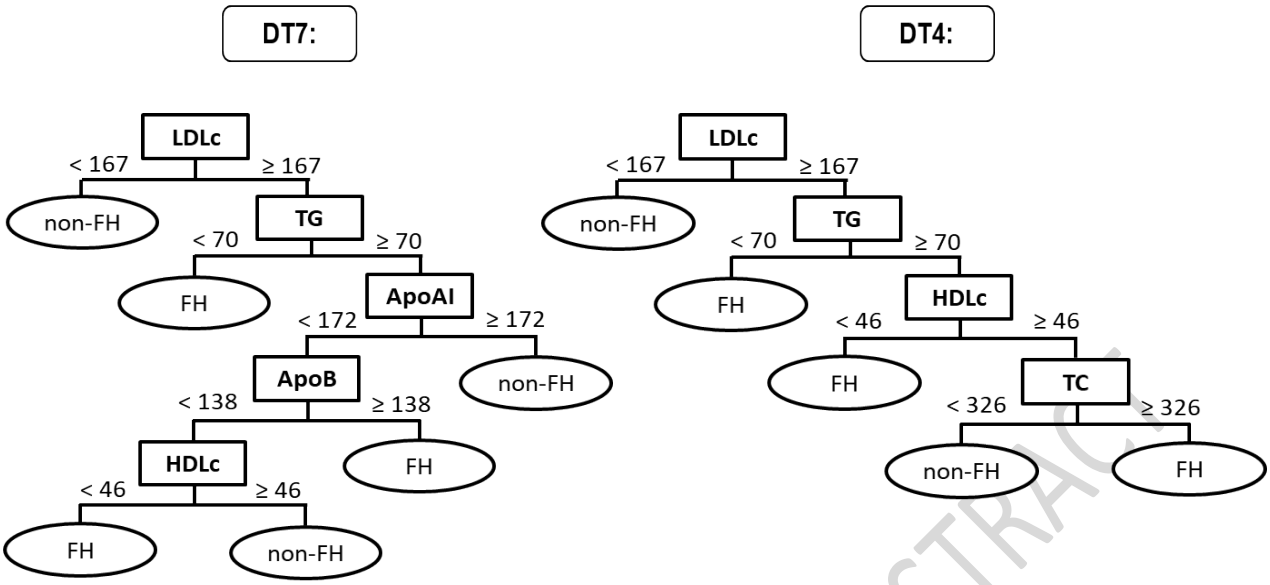


Figure 1. Representation of DT7 (left) and DT4 (right) decision tree models.