Journal of Statistics on Health Decision

EXTENDED ABSTRACT

# A3 Validation and Extension of a Logistic Regression Model Developed from a Cohort of Portuguese Children to Classify Familial Hypercholesterolemia in a Brazilian Cohort

João Albuquerque[1,2], Ana M. Medeiros[3,4], Ana C. Alves[3,4], Cinthia E. Jannes[5], Alexandre da Costa Pereira[5], Jose E. Krieger[5], Mafalda Bourbon[3,4], Marília Antunes[1,6]

[1]Centro de Estatística e Aplicações (CEAUL), Faculdade de Ciências da Universidade de Lisboa, Campo Grande 016, 1749-016, Lisboa, Portugal
[2]Departamento de Bioquímica, Faculdade de Medicina da Universidade do Porto, Alameda Prof. Hernâni Monteiro, 4200-319, Porto, Portugal
[3]Grupo de Investigação Cardiovascular, Instituto Nacional de Saúde Doutor Ricardo Jorge, Av. Padre Cruz, 1600-609, Lisboa, Portugal
[4]Instituto de Biossistemas e Ciências Integrativas (BioISI), Faculdade de Ciências da Universidade de Lisboa, Campo Grande 016, 1749-016, Lisboa, Portugal
[5]Laboratório de Genética e Cardiologia Molecular, Instituto do Coração (InCor), São Paulo, Brasil
[6]Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Campo Grande 016, 1749-016, Lisboa, Portugal

## Introduction

Familial Hypercholesterolemia (FH) is an inherited disorder of lipid metabolism, characterized by increased low density lipoprotein cholesterol (LDLc) levels from birth. The early diagnosis of FH has been associated with a significant reduction in cardiovascular disease (CVD) risk [1]. Traditional clinical criteria for FH diagnosis can however present important limitations: they are often very conservative, retaining a high percentage of false positive cases, and generally rely on variables that are either absent, such as records of family history, or very rare to find on individuals at pediatric age, such as physical signs or presence of CVD. For these reasons, alternative classification algorithms, such as logistic regression (LR) models, have been proposed [2]. In order for such models to be widely applicable, it is important that they are previously validated across different populations. The main purpose of this work was therefore to validate a LR model built from a sample of Portuguese children to classify FH in a pediatric sample from Brazil.

## Methods

The LR model was built from a sample of 364 subjects at pediatric age, from the Portuguese FH study. Candidate predictor variables for the LR model were Age, Gender, body mass index z-scores (zBMI), calculated according to the World Health Organization (WHO) guidelines [3], and a basic lipid panel, constituted by total cholesterol (TC), low density lipoprotein cholesterol (LDLc), high density lipoprotein cholesterol (HDLc), and triglycerides (TG), assessed in mg/ dL. FH molecular diagnosis was performed by next generation sequencing (NGS) [4]. The validation data was constituted by a sample of 157 children from Brazil, participants in HipercolBrasil study. Inclusion criteria consisted of not being under lipid lowering therapy (LLT) at time of assessment, and having LDLc values >120 mg/ dL, for the Portuguese sample, or >160 mg/dL, for the Brazilian cohort. The area under the ROC curve (AUROC) was used to assess the quality of the model in the original and validation datasets. Additionally, an extension of the LR model was built, by merging the Portuguese (PT) and Brazilian (BR) datasets, and re-estimating the models' parameters, also incorporating a binary candidate variable referring to "country". Statistical analysis was performed using R and R Studio software.

## Results

Compared to the Portuguese sample, the Brazilian FH patients presented a more severe phenotype, namely significantly higher TC (p=0.04), LDLc (p=0.02) and TG (p<0.01) values, and lower HDLc values (p=0.04). The equation for the LR model built from the Portuguese sample was

$$FH\ diagnosis = -4.38 + 0.04 \times LDLc - 0.06 \times HDLc - 0.01 \times TG - 0.24 \times zBMI + 0.78 \times Male$$

**Keywords:**
Familial Hypercholesterolemia; Logistic Regression; Validation

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

**Conflict of interest:**
The authors declare no conflict of interests

First published: 22JUN2021

The AUROC for this model was 0.88 for the Portuguese sample, and 0.80 when used on the Brazilian sample. When merging both datasets, significant predictor variables in the LR model were the same, plus the binary categorical variable referring to "country", and a significant interaction, between LDLc and zBMI. The equation for this new LR model was

$$FH\ diagnosis = -4.46 + 0.03 \times LDLc - 0.05 \times HDLc - 0.01 \times TG - 1.74 \times zBMI + 0.75 \times Male + 0.88 \times PT + 0.01 \times LDLc \times zBMI$$

The global AUROC of this model was 0.86. When applying the new equation solely to Portuguese patients, the AUROC was 0.88, and 0.81 when applied to the Brazilian sample. This value was not far from the maximum AUROC value obtained when training a LR model on the Brazilian sample, which was 0.82.

## Discussion

The LR model trained only on the Portuguese sample yielded an AUROC of 0.88 in the training sample, and an AUROC of 0.80 in the Brazilian testing sample. An extended model, trained with both samples merged together, yielded the same AUROC of 0.88 for the Portuguese sample, and an improved AUROC of 0.81 for the Brazilian sample, close to the maximum AUROC of 0.82, obtained by training a LR model solely on the Brazilian sample. In spite of the phenotypical differences among the two samples, it seems to be possible to validate and extend a LR model with good discriminatory ability for both cohorts. The use of a binary categorical variable to adjust for the two countries should be faced with caution, and further work is necessary to ensure that such differences can be attributed to the actual profile of the Portuguese and Brazilian populations, and not to sample characteristics.

**Ethics committee and informed consent:**

The current research was approved by an independent ethics committee and subjects gave their informed consent before they were enrolled in the study.

**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. Martin AC, Gidding SS, Wiegman A, Watts GF. Knowns and unknowns in the care of pediatric familial hypercholesterolemia. J. Lipid Res. 2017; 58(9):1765-76. https://doi.org/10.1194/jlr.S074039
2. Benlian P, Turquet A, Carrat F, Amsellem S, Sanchez L, Briffaut D, Girardet JP. Diagnosis scoring for clinical identification of children with heterozygous familial hypercholesterolemia. J. Pediatr. Gastroenterol. Nutr. 2009; 48(4):456-63. https://doi.org/10.1097/MPG.0b013e3181810b30
3. Onis MD, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. Bulletin of the World Health Organization. 2007; 85:660-7. https://doi.org/10.2471/BLT.07.043497
4. Benito-Vicente A, Alves AC, Etxebarria A, Medeiros AM, Martin C, Bourbon M. The importance of an integrated analysis of clinical, molecular, and functional data for the genetic diagnosis of familial hypercholesterolemia. Genet. Med. 2015; 17(12):980-8. https://doi.org/10.1038/gim.2015.14