EDITORIAL COMMENT

Advancing Prediction of Pathogenicity of Familial Hypercholesterolemia LDL Receptor Commonest Variants With Machine Learning Models*

Raul D. Santos, MD, PhD, MSc

Heterozygous familial hypercholesterolemia (FH) is a frequent autosomal dominant disease characterized by elevated low-density lipoprotein (LDL) cholesterol levels, family history of hypercholesterolemia, and early onset of atherosclerotic cardiovascular disease (ASCVD) (1,2). The main pathogenic mechanism behind FH is reduction in the clearance from plasma of pro-atherogenic LDL that leads to accumulation and consequent early atherosclerosis development. Adequate management of FH involves the diagnosis and treatment of index patients with potent lipid-lowering therapy as well as the identification of affected relatives by cascade screening.

In most situations, FH is diagnosed using clinical scores like the Dutch Lipid Clinics Network or Simon-Broome criteria. However, the advent of robust and less expensive targeted next-generation sequencing has increased the availability of molecular testing in clinical practice (2). In most situations (60%-80%), the FH phenotype is caused by loss of function variants in the LDL receptor gene (*LDLR*), followed by those in the apolipoprotein B gene (*APOB*) in 5% to 10% and by gain-of-function variants in the proprotein convertase subtilisin kexin type 9 (*PCSK9*) gene in <1% (1). Rarely, the phenotype may be caused by variants in the apolipoprotein E gene. In 20% to 40% of cases, no variants are encountered in the aforementioned genes, and some individuals may present aggregation of single-nucleotide variation (SNV, formerly SNP) in various genes that raise LDL cholesterol (polygenic hypercholesterolemia).

Genetic diagnosis is important because the presence of monogenic defects is implicated in higher ASCVD risk because of exposure to high LDL cholesterol since birth, which is different from the impact of raised LDL cholesterol caused by polygenes, which usually occurs later in life (2). Also, the presence of monogenic defects facilitates a definitive diagnosis of FH, and this may increase adherence to lipid-lowering therapies and improve the effectiveness of cascade screening (2).

The problem with the incrementing of genetic testing is that many previously not described variants are being encountered, mainly in the *LDLR* gene, and there is uncertainty as to whether they are responsible or not for the FH phenotype (3,4). To reduce uncertainty, the Clinical Genome FH genetic curation variant expert panel encourages the submission of FH-related variants to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), a U.S. National Center for Biotechnology Information–sponsored genomic database (3). ClinVar follows the recommendations of the American College of Medical Genetics, and FH-related variants are classified as pathogenic, likely pathogenic, variants of unknown significance, likely benign, and benign (4). The first 2 are considered as a positive molecular diagnosis of FH.

When a variant is encountered in an individual suspected of having FH, it can be checked for its pathogenicity on available databases like ClinVar. However, when not previously described, further steps are necessary for its validation, like 1) in silico testing using dedicated different software, such as Polymorphism Phenotyping v2 (PolyPhen2), Mutation Taster, Sorting Intolerant Form Tolerant (SIFT), SPIF-MutID (a specific software for missense *LDLR* variants), and Combined Annotation-Dependent Depletion (CADD), to evaluate possible damage to the protein; 2) cosegregation studies; and 3) ex vivo or in vitro functional tests (4).

Depending on the genetic defect, *LDLR* variants may be classified as nonsense, missense, synonymous, variants in the promoter and splice sequences, frameshift (small insertions or deletions) and large DNA rearrangements (2). Nonsense and frameshift alterations and large rearrangements have a strong impact on the LDL receptor protein (4). Thus, these are classified as pathogenic variants, as characterized by in silico computerized models. Missense *LDLR* variants (that may change a single amino acid) are the most common (40%-50% of those described), may not be pathogenic (1), and therefore need further testing. The problem with additional testing is that for cosegregation analysis, a large number of individuals from the same family may need to be tested for variant presence and LDL cholesterol levels, and this might not be feasible. Functional testing is done by a small number of laboratories worldwide and involves extensive laboratory work. Therefore, the task of validating pathogenic FH-causing variants is not simple, and more practical and less laborious ways are indeed necessary.

In this issue of *JACC: Basic and Translational Science*, Larrea-Sebal et al (5) describe a machine learning-based model and software developed to test the pathogenicity of missense *LDLR* variants called MLb-LDLr. For their study, 80 benign and 664 pathogenic variants from the ClinVar database were considered. In the model, specific changes in amino acid distribution and complex impact on protein conformation and function that would predict a dysfunctional LDL receptor were considered. The authors were careful in the developing/training (499 pathogenic and 54 benign variants) and validation subsets (166 pathogenic and 26 benign variants). To prove their case, 13 *LDLR* variants of unknown significance were selected to test the accuracy of the MLb-LDLr software by functional testing. In elegant

state-of-the-art studies, CHO-*ldl*Δ7 cells (that express only residual LDL receptor activity) (4) were transfected with plasmids carrying the genetic variants to be tested and were incubated with LDL to test the cycling and function of induced LDL receptors (binding, internalization and recycling to cell surface) as well as labeled LDL uptake. According to the authors, the best predictive machine learning algorithm provides a 92.5% specificity and a 91.6% sensitivity for pathogenicity.

The new algorithm was compared with PolyPhen2, Mutation Taster, SIFT, and CADD in bootstrapped and nonbootstrapped analyses. In summary, approximately 60% of the ClinVar variants were correctly predicted by any of the software programs, except for one specific variant p.(Ala299Thr) that could be correctly identified by a combination of them. Overall discrimination was excellent and did not differ much among the different software programs, all of which presented values for area under the receiver-operating characteristic curves of >0.90 (0.932 for MLb-LDLr, with the highest being 0.959 for CADD).

Regarding the results of functional studies, MLb-LDLr showed the second-best accuracy for detecting both pathogenic (72%) and benign variants (50%), being, according to authors, the most balanced software, with a 69% accuracy versus SPIF-MutID, which showed the highest accuracy (77%) but had 0% capacity to predict the 2 benign variants.

The authors are to be commended by their very elegant work—the difficulty in developing generalizable tools for confirmation of the pathogenicity of *LDLR* variants precludes a greater accuracy of molecular diagnosis of FH. The presence of a real pathogenic-causing FH variant has many clinical implications, not only for the patient but also for the family (2), and therefore an accurate diagnosis is of extreme importance. The study clearly shows, however, that none of the prediction models is perfect, but they can complement each other. The study limitations are well considered by the authors, such as the fact that they focused only on missense variants (although these are the commonest and hardest to validate) and the need to test the algorithm in other genetic FH data sets. The study advances the field; it shows that none of the available software programs is perfect but strongly suggests that a combination of them will improve capabilities to make genetic diagnoses. FH remains an underdiagnosed and under-treated disease; where ASCVD can be prevented by

early diagnosis and therapy [1], studies like this will help improve the scenario.

ADDRESS FOR CORRESPONDENCE: Dr Raul D. Santos, Heart Institute (InCor) University of Sao Paulo, Avenida Dr Eneas C. Aguiar 44, 05403-000 São Paulo, Brazil. E-mail: rauldsf@gmail.com. Twitter: @rauldsf_santos.

REFERENCES

1. Defesche JC, Gidding SS, Harada-Shiba M, Hegele RA, Santos RD, Wierzbicki AS. Familial hypercholesterolaemia. *Nat Rev Dis Primers*. 2017;3:17093:1–20.

2. Sturm AC, Knowles JW, Gidding SS, et al. Clinical genetic testing for familial hypercholesterolemia: JACC scientific expert panel. *J Am Coll Cardiol*. 2018;72:662–680.

3. Iacocca MA, Chora JR, Carrie A, et al. ClinVar database of global familial hypercholesterolemia-associated DNA variants. *Hum Mutat*. 2018;39: 1631–1640.

4. Bourbon M, Alves AC, Sijbrands EJ. Low-density lipoprotein receptor mutational analysis in diagnosis of familial hypercholesterolemia. *Curr Opin Lipidol*. 2017;28:120–129.

5. Larrea-Sebal A, Benito-Vicente A, Fernandez-Higuero J, et al. MLb-LDLr: a machine learning model for predicting the pathogenicity of LDL receptor missense variants. *J Am Coll Cardiol Basic Trans Science*. 2021;6(11):815–827.

KEY WORDS familial hypercholesterolemia, genetic variants, LDL receptor, machine learning