PLOS ONE

RESEARCH ARTICLE

In silico prediction of blood cholesterol levels from genotype data

Francesco Reggiani[1,2], Marco Carraro[1], Anna Belligoli[3], Marta Sanna[3], Chiara dal Prà[3], Francesca Favaretto[3], Carlo Ferrari[2], Roberto Vettor[3], Silvio C. E. Tosatto[1,4]*

1 Department of Biomedical Sciences, University of Padua, Padua, Italy, **2** Department of Information Engineering, University of Padua, Padua, Italy, **3** Clinica Medica 3, Department of Medicine—DIMED, School of Medicine, University of Padua, Padua, Italy, **4** CNR Institute of Neuroscience, Padua, Italy

* silvio.tosatto@unipd.it

Citation: Reggiani F, Carraro M, Belligoli A, Sanna M, dal Prà C, Favaretto F, et al. (2020) *In silico* prediction of blood cholesterol levels from genotype data. PLoS ONE 15(2): e0227191. https://doi.org/10.1371/journal.pone.0227191

Editor: Alberto G. Passi, University of Insubria, ITALY

Received: June 27, 2019

Accepted: November 14, 2019

Published: February 10, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The project was supported by the Italian Ministry of Health grant GR-2011-02346845. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In this work we present a framework for blood cholesterol levels prediction from genotype data. The predictor is based on an algorithm for cholesterol metabolism simulation available in literature, implemented and optimized by our group in the R language. The main weakness of the former simulation algorithm was the need of experimental data to simulate mutations in genes altering the cholesterol metabolism. This caveat strongly limited the application of the model in the clinical practice. In this work we present how this limitation could be bypassed thanks to an optimization of model parameters based on patient cholesterol levels retrieved from literature. Prediction performance has been assessed taking into consideration several scoring indices currently used for performance evaluation of machine learning methods. Our assessment shows how the optimization phase improved model performance, compared to the original version available in literature.

Introduction

Recent exome-wide association studies [1] started to shed light on the complex genomic architecture behind the regulation of blood cholesterol levels in humans. Reliable tools to predict human cholesterol levels from genotype are not available yet. The huge number of genes involved in the regulation of this trait and the complex interaction with environmental factors as diet, gender and age make modelling cholesterol levels a difficult task. However, particular situations exist where a single mutation is related to significant variations of cholesterol levels. Example are damaging mutations on genes involved in hepatic uptake of Low Density Lipoprotein (LDL), as the Low Density Lipoprotein Receptor (LDLR) gene, causing familial hypercholesterolemia characterized by elevated levels of LDL and total plasma cholesterol but with normal concentrations of triglycerides [2]. Other processes involved in cholesterol metabolism are affected by genetic mutations, with a wide range of phenotypes depending on the gene involved, like marked High Density Lipoprotein (HDL) cholesterol levels deficiency as seen in patients affected by Tangier disease [3]. The aim of this work is to test the reliability of a modelling approach aimed to predict cholesterol levels relying on patient's genotype data only. Different tools have been developed for blood lipid levels prediction, some of them are

regression methods based on a set of variables representing patient genotypes (e.g. presence or absence of SNPs associated to lipid traits)[4] and phenotype (e.g. Body Mass Index, gender, age, etc.). These methods require a huge amount of data for training and test, with predictions having low correlation to lipid profiles [5]. Other research groups have developed tools that are able to predict a familial hypercholesterolemia phenotype from LDLR missense mutations, but not the range of blood lipid values [6]. A different strategy is to develop an *in silico* mathematical model, that represents human cholesterol metabolism, simulate the effect of a mutation and take the response of the model as predicted levels of cholesterol. Effective way to simulate *in silico* metabolism are dynamic models. In this kind of simulations, the development of the system in time is computed through a set of ordinary differential equations, able to simulate the variations of chemical species concentration. Several information are required for the development of these models: interactions between the chemical species involved in the biological process, kinetic parameters associated to chemical reactions occurring in the system and its initial state. The simulation of a biological perturbation could be obtained by modifying model parameters (e.g. decreasing kinetic rates) and observing variations occurring in the system [7]. Several *in silico* models simulating cholesterol metabolism have been proposed so far, both for human and animal models [8]. A recent review [8] has described a set of published mathematical models, based on differential equations, which simulate cholesterol metabolism at different levels. Some of the presented methods were focused on specific reactions, as endocytosis or excretion of lipoproteins by hepatocytes, other attempted to model cholesterol metabolism at a whole body scale. One of these models, published in literature by van de Pas and colleagues in 2012 [9], was developed on the basis of genes and related metabolic reactions that have a relevant role on the control of human cholesterol homeostasis. In this work we decided to adopt an algorithm based on this mathematical model to predict cholesterol levels. This choice was motivated by different factors, from one hand this method has passed a validation process both in the original publication [9] and in a following review by different authors [8]. On the other hand this model is gene based and computes levels of LDL and HDL cholesterol induced by a mutation, making it suitable for the prediction of blood lipid levels from genotype data. This physiologically based kinetic model [9] is based on differential equations, computing the flow of cholesterol in different body organs. The whole process is regulated by a set of rates, each one related to a gene that has a key role in cholesterol metabolism. Simulation of mutations effects depends on reducing rates (f_{mut}) estimated from wet lab experiments. This kind of information is usually not easily accessible, strongly limiting the usability of the model. In this work we implemented and optimized the framework for blood cholesterol levels prediction making it able to perform reliable predictions when only patient's genotype data are available. The model has been improved through a training phase, in which reducing rates (f_{mut}) were estimated from phenotype data of patients affected by mutations on key regulatory genes of cholesterol metabolism. Assessment measures confirmed how the optimized model presents improved performance, reducing the error between experimental and predicted data, compared to the original version available in literature [9].

Materials and methods

In silico kinetic model for cholesterol levels prediction

An available *in silico* kinetic model [9] has been used as basis for predicting plasma cholesterol concentrations in humans. The kinetic model was developed to simulate cholesterol levels for a reference man of 70 kg. The model is composed of 8 pools, representing main sites of cholesterol storage in the human body (Fig 1). These pools can be grouped in 4 main entities corresponding to plasma, intestine, liver and periphery. Each cholesterol pool is modeled by a

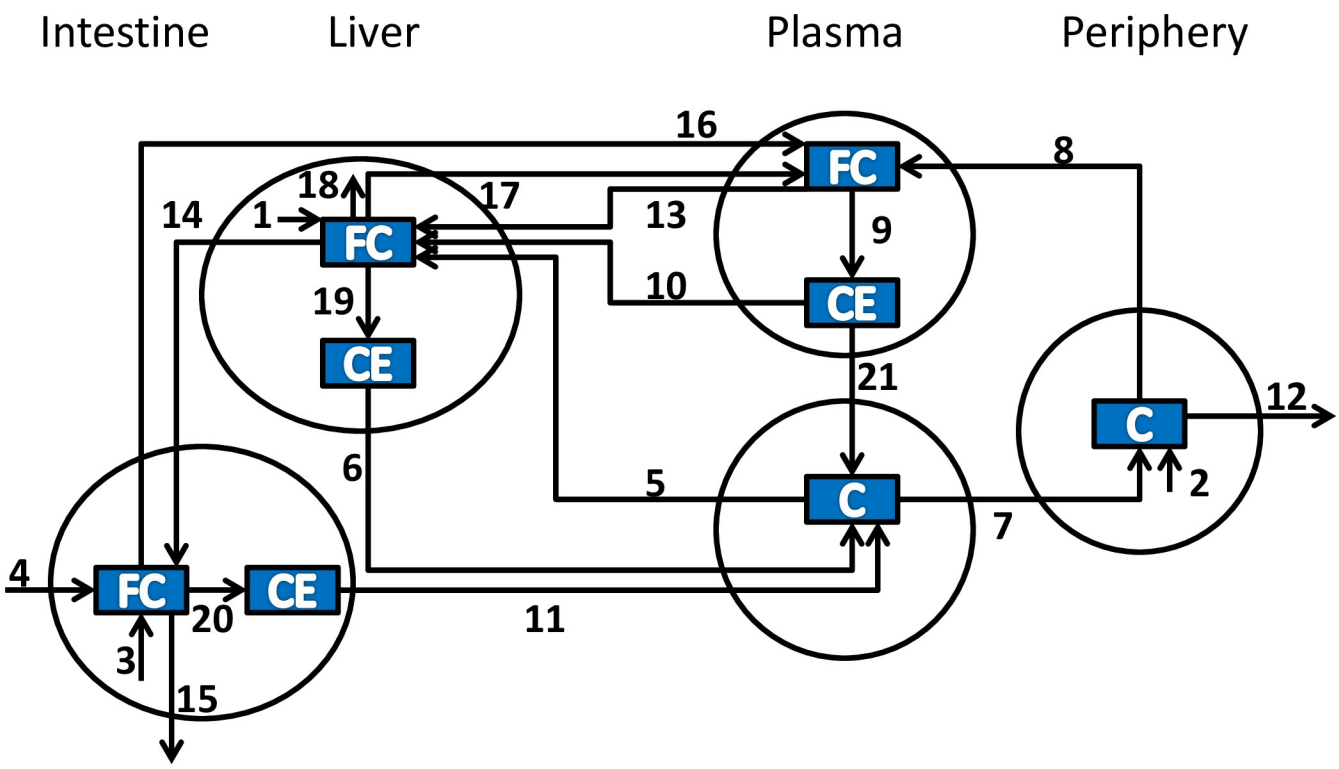

Fig 1. Conceptual model for pathways and genes determining cholesterol plasma levels used van de Pas and colleagues [9], [10]. Process numbers stand for: 1, hepatic cholesterol synthesis (DHCR7); 2, peripheral cholesterol synthesis(DHCR7); 3, intestinal cholesterol synthesis (DHCR7); 4, dietary cholesterol intake (NPC1L1); 5, hepatic uptake of cholesterol from LDL (LDLR,APOB,APOE); 6, VLDL-C secretion (MTTP); 7, peripheral uptake of cholesterol from LDL (LDLR,APOB,APOE); 8, peripheral cholesterol transport to HDL (ABCA1); 9, HDL-associated cholesterol esterification (LCAT); 10, hepatic HDL-CE uptake (SCARB1); 11, intestinal chylomicron cholesterol secretion (MTTP); 12, peripheral cholesterol loss; 13, hepatic HDL-FC uptake (MTTP); 14, biliary cholesterol excretion (ABCG8,NPC1L1); 15, fecal cholesterol excretion; 16, intestinal cholesterol transport to HDL (ABCA1); 17, hepatic cholesterol transport to HDL (ABCA1); 18, hepatic cholesterol catabolism (CYP7A1); 19, hepatic cholesterol esterification (SOAT2); 20, intestinal cholesterol esterification (SOAT2); and 21, CE transfer from HDL to LDL (CETP).

https://doi.org/10.1371/journal.pone.0227191.g001

differential equation, composed by a set of rates moving cholesterol from or to a different one. These pools are connected by 21 kinetic rates, each one representing the main gene responsible of regulating that specific biochemical reaction (Table 1).

Rates depend on kinetic constants, organ volumes, body weight and pool cholesterol concentrations. In the original model, all parameters have been computed from data published in literature [9]. The model was calibrated to immediately reach a steady state, a stable equilibrium in which each compartment has a constant cholesterol concentration in time. To simulate a mutation affecting the activity of a gene, a set of rate reduction parameters (f_{mut}), each one in the interval [0, 1], multiplies the standard rates to represent the effect of the mutated genes. These values were computed on the basis of experimental data available in literature. Example is the value of the f_{mut} related to mutations affecting the gene CYP7A1 involved in byle acid synthesis, where the rate reduction parameter was computed as the ratio of bile acids contents in the stools of patients carring the mutation over controls [9].

These kind of perturbations force a re-tuning of the system, moving from the original steady state to a new one, where blood cholesterol profiles were comparable to the real values detected in patients affected by that particular mutation.

Table 1. Biological process and genes associated to each rate of the model.

rate	Biological process	gene
1	hepatic cholesterol synthesis	DHCR7
2	peripheral cholesterol synthesis	DHCR7
3	intestinal cholesterol synthesis	DHCR7
4	dietary cholesterol intake	NPC1L1
5	hepatic uptake of cholesterol from LDL	LDLR, APOB, APOE
6	VLDL-C secretion	MTTP
7	peripheral uptake of cholesterol from LDL	LDLR, APOB, APOE
8	peripheral cholesterol transport to HDL	ABCA
9	HDL-associated cholesterol esterification	LCAT
10	hepatic HDL-CE uptake	SCARB1
11	intestinal chylomicron cholesterol secretion	MTTP
12	peripheral cholesterol loss	
13	hepatic HDL-FC uptake	MTTP
14	biliary cholesterol excretion	ABCG8, NPC1L1
15	fecal cholesterol excretion	
16	intestinal cholesterol transport to HDL	ABCA1
17	hepatic cholesterol transport to HDL	ABCA1
18	hepatic cholesterol catabolism	CYP7A1
19	hepatic cholesterol esterification	SOAT2
20	intestinal cholesterol esterification	SOAT2
21	CE transfer from HDL to LDL	CETP

Reaction rates present in the model and the associated biological process they represent, also the main genes involved in the process are reported [9], [10].

https://doi.org/10.1371/journal.pone.0227191.t001

Model implementation

The algorithm of the available physiologically based kinetic model [8], was implemented in R language [11]. The *deSolve* package [12] was used for solving differential equations.

New f_{mut} values have been obtained thanks to a training procedure exploiting a dataset composed of cholesterol levels and genotypes of mutated patients (S1 Table). This operation required the usage of the Levenberg-Marquardt algorithm as implemented in the *Minpack.lm* package [13].

The R scripts are publicly available from the GitHub repository at URL: https://github.com/BioComputingUP/Cholesterol-model

Training phase

To improve performance in predicting genetic mutations' effect on cholesterol levels, f_{mut} parameters, each one related to a particular gene mutation and rates of the model, have been trained on phenotype data of a dataset of patients, retrieved from literature. The Levenberg-Marquardt minimization method has been used to estimate the f_{mut} parameters able to minimize the difference between predicted and experimentally measured levels of HDL and LDL, divided by the control, intended as level of cholesterol of the model when no mutation is present (Eqs 1 and 2). Exceptions are patients affected by mutations on the DHCR7 genes where only total cholesterol (TC) levels were found in literature. In this case the difference between

real and predicted total cholesterol rate was taken in account (Eq 3).

$$\Delta HDL = \frac{HDL_{experimental}}{HDL_{control}} - \frac{HDL_{predicted}}{HDL_{control}} \quad (1)$$

$$\Delta LDL = \frac{LDL_{experimental}}{LDL_{control}} - \frac{LDL_{predicted}}{LDL_{control}} \quad (2)$$

$$\Delta TC = \frac{TC_{experimental}}{TC_{control}} - \frac{TC_{predicted}}{TC_{control}} \quad (3)$$

The optimized f_{mut} parameters are reported in Table 2. The values of the first column are the result of a training procedure, which is based on a dataset of patients and regulated by the sensitivity of the rates involved. This is not true for the f_{mut} based on experimentally determined variables, since they were computed on the basis of specific molecule concentrations [9]. The consequence of these two different strategies of parameter estimation is reflected by the difference between the f_{mut} associated to the same gene in the two columns. Example is the CYP7A1 gene: the f_{mut} estimated by van de Pas and coauthors is equal to 0.05, as the value of bile acids in the stools of patients compared to controls. On the contrary, the corresponding value is higher after an optimization procedure, which has been influenced by the sensitivity of that rate and the cholesterol levels of the training set elements.

Training set

The training set is represented by a custom dataset of patients affected by single mutations (either in homozygous or heterozygous form), in one of the key genes regulating cholesterol metabolism (Fig 2). For each patient the levels of HDL, LDL or total cholesterol and the causative mutation were extracted from literature (S1 Table). Each gene is covered by a different number of individuals due to the relative abundance of works in literature (Table 3). Special cases are the CETP gene, where only information regarding the mean levels of blood cholesterol were found in literature and the DHCR7 gene, where only the levels of blood total cholesterol were found. The training set has been divided in two sections (Table 3). The first group is represented by hypercholesterolemic patients with mutations affecting a set of genes involved in the development of Autosomal Dominant Hypercholesterolemia [14]: LDLR, APOB and

Table 2. Optimized f_{mut} parameters and related genes.

Gene	Reggiani et al f_{mut}	van de Pas et al 2012 f_{mut}
LDLR	0.58	0.38
APOB	0.9	0.31
APOB (hom.)	0.55	0.32
ABCA1	0.53	0.41
APOE	0.72	0.45
CETP	0.43	0.65
LCAT	0.48	0.62
LCAT (hom.)	1	0
DHCR7	0	0
CYP7A1	0.81	0.05

Genes represented in the training, test set and related f_{mut} as computed by the optimization procedure or by using experimental variables, as reported by van de Pas and colleagues[9]

https://doi.org/10.1371/journal.pone.0227191.t002

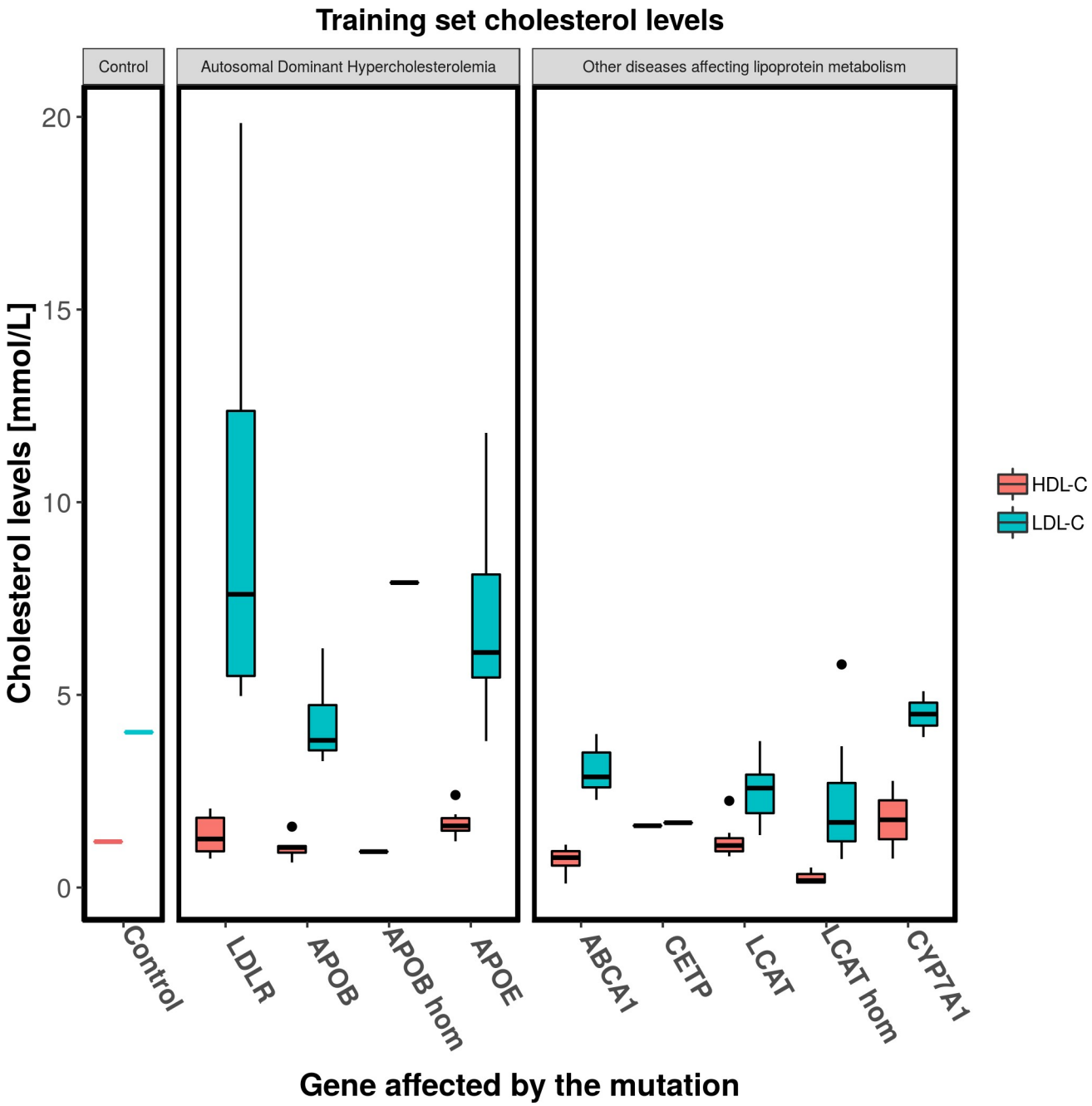

Fig 2. Training set patients cholesterol levels. Boxplot of HDL and LDL cholesterol levels of the patients composing the training set. From left to right: cholesterol levels of the model at the steady state, patients affected by Autosomal Dominant Hypercholesterolemia (with high levels of LDL and low HDL) and patients affected by other disease altering lipoprotein metabolism.

https://doi.org/10.1371/journal.pone.0227191.g002

APOE genes, represented by reaction 5 and 7 of the model (Table 1). The second part of the dataset is composed of patients with damaging mutations on 5 different genes: ABCA1, CETP, LCAT, DHCR7 and CYP7A1 (affected rates are shown in Table 1). Patients of the Autosomal Dominant Hypercholesterolemia dataset are characterized by high levels of LDL cholesterol, while the second part of the dataset is composed by different ranges of HDL and LDL, depending on the gene affected by the mutation (Fig 2).

Table 3. Training set composition.

Dataset	Gene	Patients	Mutations	type	rate
Autosomal Dominant Hypercholesterolemia	*LDLR*	13	9	heterozygous	5, 7
	APOB	7	1	heterozygous	5, 7
	APOB (hom)	1	1	homozygous	5, 7
	APOE	12	2	heterozygous	5, 7
Other disease altering lipoprotein metabolism	*ABCA1*	7	3	6 heterozygous, 1 compound heterozygous	8, 16, 17
	CETP	1	1	heterozygous	21
	LCAT	17	2	heterozygous	9
	LCAT (hom)	7	4	homozygous	9
	CYP7A1	2	1	heterozygous	18

Disease, gene, number of patients with a mutation in that gene, number of different mutations, type of mutation (heterozygous, homozygous or compound heterozygous), rates representing that gene in the model

https://doi.org/10.1371/journal.pone.0227191.t003

Test phase

Prediction performance was tested on a dataset retrieved from literature [9]. The dataset is the same one used to test performance of the former version of the model. This test set has been used in order to highlight performance comparison between the versions of the algorithm. The effect of a genetic mutation was simulated for each individual of the dataset until a steady state was reached (fixed threshold: 1000 days). Predicted HDL, LDL and total cholesterol were than compared to experimental data.

Test set

Test set is composed by patients affected by 10 mutations. All mutations affect genes present in the training set of this work. The first group of mutations maps on the LDLR, APOB and APOE genes, involved in hepatic cholesterol uptake. Patients affected by this kind of mutations have high levels of LDL and total cholesterol. Genetic mutations affecting the other genes of the dataset have different effects on lipid profiles. Mutations on the ABCA1 gene can cause marked HDL cholesterol levels deficiency as reported for different diseases like hypoalphalipoproteinemia or Tangier disease [3]. CETP is a protein involved in the transport of cholesterol esters from HDL to LDL, deficiency of this protein can cause a marked increase of HDL levels [3]. LCAT is a gene involved in cholesterol esterification in HDL particles, mutation on this gene can cause LCAT deficiency, characterized by low levels of HDL and LDL cholesterol [3]. Patients with mutations in heterozygous or homozygous form has been included in the training set. DHCR7 gene is responsible for the last step of the cholesterol biosynthesis pathway. Reduced enzyme activity cause low levels of blood cholesterol, as reported in patients affected by the Smith-Lemli-Opitz syndrome[15]. CYP7A1 gene is involved in cholesterol catabolism and bile acids synthesis, mutations affecting this gene cause an increase of total, hepatic cholesterol and a decrease in bile acids secretion [16].

Results and discussion

Performance assessment

The assessment approach used in this work was influenced by the methods used for the evaluation of tools predicting the effect of variants on continuous phenotypes [17]. Model performance has been evaluated in terms of distance and correlation, measuring the deviation from

experimental values while assessing model capability to predict a decrease or increase of cholesterol levels. The analysis has been conducted at two levels. In the first part of the assessment, predictions were evaluated at the level of the single gene to understand if prediction error was homogeneous or significantly different for some of the mutations. The second part of the assessment focused on the overall performance of the predictor. In the first phase, the analysis was focused on assessing model performance in terms of prediction error computed on each element (i) of the test set: the deviation was evaluated by computing the difference between predicted and experimental data, in terms of rate of cholesterol levels (CL), for TC, HDL or LDL, in case and control (Eq 4).

$$Error(i) = \frac{CL_{predicted}(i)}{CL_{control}} - \frac{CL_{experimental}(i)}{CL_{experimental\ control}(i)} \tag{4}$$

To evaluate the magnitude of the error, compared to real values, this measure (Eq 4) was divided by the corresponding experimental value and multiplied by 100 (Eq 5).

$$Error(i)\% = \frac{Error(i)}{\frac{CL_{experimental}(i)}{CL_{experimental\ control}(i)}} \cdot 100 \tag{5}$$

This analysis was aimed to highlight mutation effects that where under or over-predicted.

In the second part of the assessment, model performance has been evaluated in terms of correlation and error measures on the whole dataset. Correlation measures used for the assessment were Pearson (r or PCC) and Kendall's tau (τ or KCC) correlation coefficients (Eqs 6 and 7).

$$r = \frac{n\sum_{i=1}^{n} y_i \bar{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \bar{y}_i)}{\sqrt{[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2][n\sum_{i=1}^{n} \bar{y}_i^2 - (\sum_{i=1}^{n} \bar{y}_i)^2]}} \tag{6}$$

$$\tau = \frac{2}{n(n-1)} n\sum_{i<j} sgn(x_i - x_j) sgn\left(y_i - y_j\right) \tag{7}$$

The PCC has been used to evaluate the correlation between real and predicted data as continuous measures, while KCC estimated the conservation of the order of magnitude of the experimental cholesterol levels in predicted ones.

To better understand the amount of variability described by the model compared to the variability inside the data, the R^2 index was used (Eq 8).

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{8}$$

RMSE (Root Mean Squared Error) has been used to evaluate if the method predicted cholesterol levels with huge deviation from real ones (Eq 9).

$$RMSE = \sqrt{\frac{(\sum_{i=1}^{N} Error(i))^2}{N}} \tag{9}$$

The MAE (Mean Absolute Error) has been computed as the mean absolute error between model predictions and experimental values (Eq 10).

$$MAE = \frac{\sum_{i=1}^{N} |Error(i)|}{N} \tag{10}$$

A bootstrap procedure has been used to evaluate the robustness of the performance measures presented in this work: the probability of obtaining the same or better scores with a random shuffle of model predictions, as seen in [17], [18]. In particular each index has been computed 10000 times, considering either the rate of HDL, LDL or total cholesterol of all the elements of the test set each time, on the vector of model predictions in a random order and the corresponding vector of experimental values. From the resulting distribution of scores, the probability (p-value) of obtaining a score greater or equal to the real one was computed. The only exception was the RMSE index, in this case the probability of obtaining a value lower or equal to the score computed in the original assessment was calculated. All indices with a p-value lower than 0.05 were considered as statistically significant.

A sensitivity analysis was performed on a set of rates, corresponding to genes represented in the test set. The aim of this analysis was to understand the effect of a perturbation of specific model parameters on the output [7]. In this case, we decreased rates associated to genes represented in the test set, using a reducing factor [0.1, 1] and measured model cholesterol levels when a steady state was reached.

Performance assessment on single genes mutations

The first part of the assessment was aimed to understand how the model performs on the single mutations represented in the test set. This type of analysis highlighted cases where the model overestimated or underestimated cholesterol levels, respectively called positive or negative errors. The error represents the increase or decrease of cholesterol in case relative to controls (Eqs 1, 2 and 3), which is not observed in experimental data. The errors were divided by real data and converted to percentages as reported in Table 4. The standard deviation of model predictions, computed as the standard deviation of the predicted cholesterol levels for the elements of the training set given a mutated gene, has been reported in Table 5. As already introduced, the datasets of patients have been divided in two sections. The first group of elements of the test set is composed by patients affected by damaging mutations on genes that have a role in the onset of the Autosomal Dominant Hypercholesterolemia: LDLR, APOB and APOE. The main effect of simulating these mutations is an increase of blood cholesterol levels of LDL and decrease of HDL (Table 5), as observed in real cases [19]. The algorithm predicted cholesterol levels caused by mutations in LDLR and APOB with a reduced error intervals: [-35.3%, 11.5%] for HDL, [-26.2%, -12.9%] for LDL and [-20.5%, -13.7%] for total cholesterol, respect to the former version of the model. The original model in fact, has shown to drive predictions towards an overestimation of the mutation effect, as shown by the prediction errors of HDL [-52%, -30.8%], LDL [35%, 139.7%], and total cholesterol [29.7%, 115.9%]. A particular case is the one regarding mutations in APOE, where the algorithm strongly underestimated the effect of damaging mutations on total cholesterol levels. In this case, a higher error has been registered for our optimized model (-53.1%) compared with the former version (-27.4%). This situation is mainly related to the fact that the average levels of total cholesterol of patients in the training set was lower than the one of the test set.

The effect of damaging mutations on the other genes of the test set have been simulated by reducing different set of rates of the model. The model predicted the effect of ABCA1 mutations as a decrease in HDL levels, but also produced overestimated decrease in LDL levels (Table 4), which is not usually observed in patients affected by related disease like Hypoalphalipoproteinemia [3]. CETP is a protein involved in the transport of cholesterol esters from HDL to LDL, deficiency of this protein can cause a marked increase of HDL levels [3]. In this case, the model correctly predicted an increase in HDL cholesterol levels, with a bigger error when optimized f_{mut} was used (Table 4). The LCAT gene is involved in cholesterol esterification in

Table 4. Models predictions percentage error on elements of the test set.

Mutation	Gene	Predicted van de Pas et al			Predicted Reggiani et al		
		HDL	LDL	TC	HDL	LDL	TC
1	LDLR	-30.83	35.05	29.65	-13.52	-14.82	-13.69
2	APOB	-36.7	139.71	115.91	11.53	-26.23	-20.45
3	APOB *(hom)*	-51.96	62.66	49.05	-35.34	-12.89	-15.14
4	ABCA1	147.99	-57.44	-44.77	200.05	-51.25	-35.99
5	APOE	NA	NA	-27.4	NA	NA	-53.15
6	CETP	12.7	-4.82	-0.72	34.04	-11.53	-0.46
7	LCAT	34.3	-0.66	21.7	39.18	-3.08	20.55
8	LCAT *(hom)*	679.37	-19.37	10.11	426.32	21.95	29.87
9	DHCR7	NA	NA	171.51	NA	NA	171.51
10	CYP7A1	-4.42	-42.09	-34.15	1.37	-50.07	-40.81

Mutation numeric ID, gene, HDL, LDL and total cholesterol error (as percentage of experimental value), of predictions based on f_{mut} as reported by van de Pas and colleagues[9], or trained f_{mut}.

https://doi.org/10.1371/journal.pone.0227191.t004

HDL particles, patients with mutations on this gene generally have low levels of HDL and LDL cholesterol [3]. In this case the model was not able to accurately simulate HDL and LDL levels in all cases (Table 3). Explanation could be that it was not possible to train the parameter for patients with a homozygous mutation on the LCAT gene (f_{mut} has been assumed to be equal to 1). DHCR7 gene is involved in cholesterol biosynthesis pathway, mutations reducing related enzymatic activity cause low levels of blood cholesterol [15]. In all cases the model predicted a bigger decrease in total cholesterol levels with an error of -171.5%. CYP7A1 gene is involved in cholesterol catabolism and bile acids synthesis, mutations affecting this gene can cause an increase of total and hepatic cholesterol [16]. In this case the model generally predicted an increase of LDL and total cholesterol levels and a decrease in HDL cholesterol. Nevertheless, CYP7A1 simulations showed an underestimation of LDL and total cholesterol levels (Table 4).

Performance assessment on the overall dataset

The overall assessment highlighted that the training phase increased model performance (Table 6). Both Pearson and Kendall correlation coefficients show that the use of trained f_{mut}

Table 5. Experimental and predicted cholesterol levels of the test set.

Mutation	Gene	Experimental value			Predicted van de Pas et al			Predicted Reggiani et al		
		HDL	LDL	TC	HDL	LDL	TC	HDL	LDL	TC
1	LDLR	0.86	2.17	1.85	0.59	2.93	2.4	0.74±0.17	1.85±1.31	1.6±0.97
2	APOB	0.85	1.52	1.36	0.54	3.64	2.94	0.95±0.07	1.12±0.21	1.08±0.14
3	APOB *(hom)*	1.12	2.24	1.97	0.54	3.64	2.94	0.72	1.95	1.67
4	ABCA1	0.22	1.42	1.07	0.55	0.6	0.59	0.66±0.19	0.69±0.15	0.68±0.16
5	APOE	NA	NA	2.8	0.65	2.44	2.03	0.84±0.13	1.45±0.57	1.31±0.41
6	CETP	1.1	0.98	1.01	1.24	0.93	1	1.47	0.87	1.01
7	LCAT	0.79	0.97	0.81	1.06	0.96	0.99	1.1±0.16	0.94±0.11	0.98±0.05
8	LCAT *(hom)*	0.19	0.82	0.77	1.48	0.66	0.85	1	1	1
9	DHCR7	NA	NA	0.2	1.13	0.37	0.54	1.13±0.01	0.37±0.04	0.54±0.03
10	CYP7A1	0.97	2.09	1.74	0.93	1.21	1.15	0.98±0.02	1.04±0.06	1.03±0.04

Mutation numeric ID, gene, HDL, LDL and total cholesterol, from wet lab experiments[9], from predictions based on f_{mut} as reported by van de Pas and colleagues[9], or trained f_{mut} with standard deviation.

https://doi.org/10.1371/journal.pone.0227191.t005

Table 6. Models performances on the whole test set.

Prediction	PCC	KCC	MAE	RMSD	R^2
van de Pas et al predicted HDL ratio	-0.22	-0.18	0.4	0.54	0.05
Reggiani et al predicted HDL ratio	0.32	0	0.32	0.4	0.11
van de Pas et al predicted LDL ratio	**0.65**	**0.55**	**0.77**	**1.03**	0.43
Reggiani et al predicted LDL ratio	**0.74**	0.5	**0.39**	**0.5**	**0.55**
van de Pas et al predicted TC ratio	**0.66**	**0.63**	**0.55**	**0.71**	**0.43**
Reggiani et al predicted TC ratio	**0.75**	**0.69**	**0.42**	**0.57**	**0.56**

Cholesterol level and predictor: Pearson Correlation Coefficient, Kendall rank Correlation Coefficient, Root Mean Squared Error, Mean Absolute Error and R-squared index computed on the test set. Values in bold have a p-value lower than 0.05, computed as the probability of obtaining an index better than the original one in a distribution of 10000 random scores, generated by a bootstrap procedure.

https://doi.org/10.1371/journal.pone.0227191.t006

increased algorithm capability to predict variations on cholesterol levels caused by gene mutations. In particular, HDL levels predicted by the former version of the model have shown negative correlation with experimental values. The MAE and RMSE index computed on HDL and total cholesterol levels have been decreased thanks to the training procedure, and the second index on predicted LDL is one half of the one obtained with the original version of the model. R^2 indices on blood cholesterol levels show an increase in the amount of variability explained by the model when a training phase is added. The bootstrap procedure has shown that for all indices computed on HDL levels predictions, model performance was not better than random.

Conclusions

In this work we improved and assessed the performance of an *in silico* prediction method for blood cholesterol levels. The addition of a training phase has generally improved model performance, as shown in Table 6. Our training phase overcomes the problem of model usability when no experimental data is available for f_{mut} parameters estimation. The reducing parameters presented by van de Pas and colleagues were computed from variables obtained in wet lab experiments[9]. This procedure, in contrast with the training methodology we applied in this work, did not take in account that decreasing different rates by the same factor can lead to modification of cholesterol profiles with different magnitude. To better understand model responses to different simulations, we performed a sensitivity analysis on the rates involved in the test set (Fig 3). This analysis showed that reduction of rate 5 and 7 produced a consistent decrease of model predicted HDL, while increasing LDL and total cholesterol. The training procedure has computed f_{mut} on the basis of the difference between experimental levels and model response to the reduction of selected rates, as previously explained. This procedure avoid an overestimation of the effect of mutations on the LDLR and APOB genes, as observed when f_{mut} based on experimental variables were used (Table 4). The use of trained parameters has decreased prediction error when model was not able to correctly simulate the effect of a mutated gene on cholesterol levels. In particular rate 9 regulates the flow of cholesterol from free to esterified form in HDL particles, LCAT gene product activity. The effect of a mutation on this gene is predicted by the model as an increase of HDL cholesterol while the opposite is observed in real data (Table 5). In this case the training procedures had hampered in part model inability to correctly predict HDL and LDL deviations caused by mutations on this gene by fixing the f_{mut} to 1 in the homozygous case, since the reduction of this parameter was not able to reduce the difference between experimental and predicted values. A similar behavior has been observed for the estimation of the reducing CYP7A1 f_{mut}: in this case the trained parameter (0.81) was greater than the value computed by van de Pas and coauthors (0,

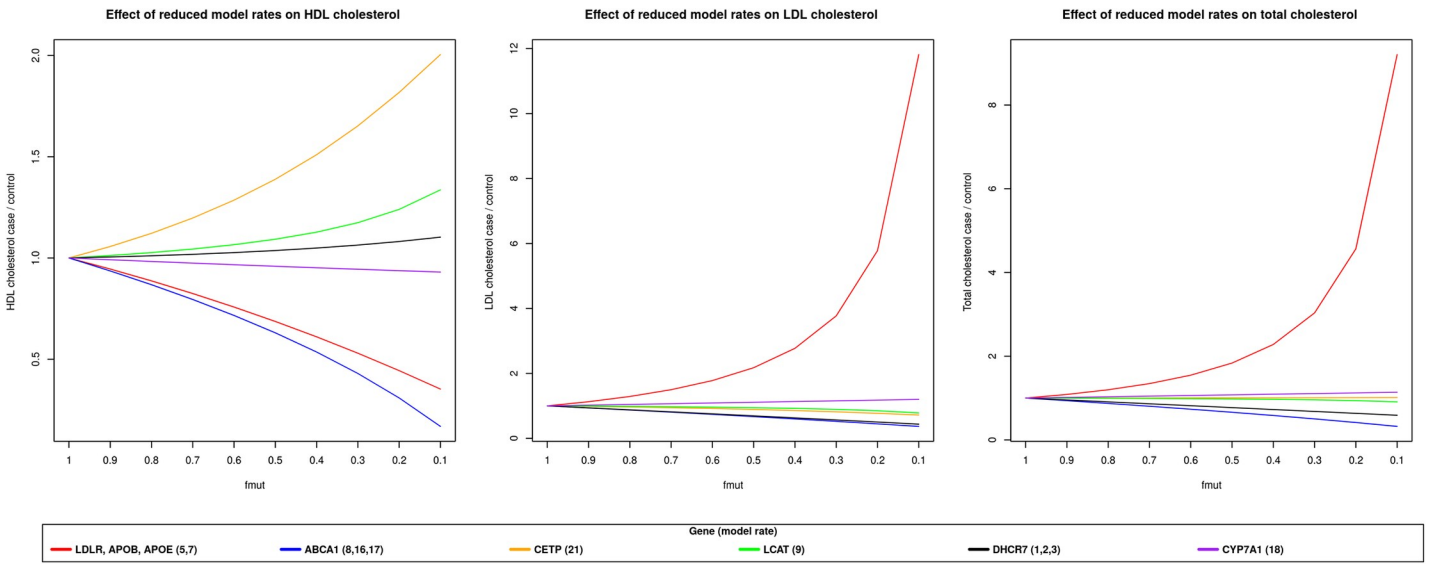

Fig 3. Model response in terms of HDL, LDL and total blood cholesterol at different values of f_{mut}. The effect of reducing model rates, involved in the test procedure, on HDL, LDL or total cholesterol levels.

https://doi.org/10.1371/journal.pone.0227191.g003

Table 2), while the difference in predicted cholesterol levels was relatively small (Table 5). This difference is related both to the low sensitivity of the rate (Fig 3) and the inability to produce a consistent increase of LDL and Total cholesterol levels, while producing a limited decrease of HDL as experimentally observed (Table 5).

This model can be considered a valid tool for the study of cholesterol metabolism *in silico*, considering the other models currently available [8] and the predictions error: the average relative deviations between model predictions and experimental data were 49% for HDL-C, 43% for LDL-C and 36% for total cholesterol [9]. Mathematical models are a simplified representation of the original system, this from one hand results in a relatively simple tool for making inference and simulate different experimental conditions *in silico*. From the other hand, they don't represent the selected system completely, hence deviation from real data are expected. Prediction error in principle could be decreased by increasing the number of parameters, however this process will increase model complexity and present problems related to parameter identifiability and fitting to experimental error [20]. Prediction of *in silico* cholesterol levels is a complex procedure, the physiologically based *in silico* cholesterol model optimized in this review has proven its ability to predict cholesterol levels behavior with reduced error when only genotype data is available. Given the huge number of genomic loci controlling cholesterol homeostasis, much of that still unknown, gender-related effects and environmental factors that affects blood cholesterol levels, the possibility of developing a software able to accurately predict cholesterol levels seems far from true. Despite these critical points, in this work we considered patients affected by monogenic dominant diseases only, so we expect that ethnicity and other factors will have a relatively low contribution on the onset of the phenotype. Nowadays genetic assays are increasingly used to support the diagnosis of monogenic diseases affecting blood lipid levels, as Familial Hypercholesterolemia (FH) [19]. Studies have shown that coronary artery disease risk is higher in carrier of FH mutations compared to those without, this is likely the consequence of a higher life-long exposure to LDL. In this context our work could be considered a further step in the process of using genetic tests for the detection and treatment of patients affected by FH and other genetic disorders affecting blood cholesterol

levels. Under this perspective we think that our work could be useful to simulate and study the effect of genetic variants on human cholesterol metabolism, in particular for variants affecting genes involved in hepatic cholesterol uptake (model rate 5, 7) and FH, as LDLR and APOB, where the trained model has predicted blood cholesterol levels with little error (Table 4). Furthermore, given the newly developed therapies against molecular targets (such as the anti PCSK9 monoclonal antibodies) [21] the model could be useful to identify the patients that are best candidates for treatment. Simulation of drug actions could be another possible application of the proposed model. It is well known that different genetic backgrounds have a strong effect on drug activity and *in silico* prediction methods can have poor performance with patients that don't have the same ethnicity of individuals used during the training procedure, as seen in CAGI Warfarin dosing challenge [22].

In light of these considerations, the physiologically based *in silico* model of human cholesterol metabolism, optimized in this work, can be a useful tool for studying the effect of damaging mutation on genes involved in cholesterol homeostasis.

Supporting information

S1 Table. *In silico* cholesterol model training set. Phenotype, genotype information and reference paper of each element of the data set used to train the model.
(XLSX)

S1 File. Data set and algorithms used in the paper.
(ZIP)

Acknowledgments

The project was supported by the Italian Ministry of Health grant GR-2011-02346845. The authors are grateful to Francesco Tabaro for his useful suggestions. Ggplot2 has been used to produce Fig 2[23]. Boot package has been used to implement the bootstrap procedure [24], [25].

Author Contributions

Conceptualization: Francesco Reggiani, Silvio C. E. Tosatto.

Data curation: Francesco Reggiani, Anna Belligoli, Marta Sanna, Chiara dal Prà, Francesca Favaretto.

Formal analysis: Francesco Reggiani, Marco Carraro, Silvio C. E. Tosatto.

Funding acquisition: Roberto Vettor, Silvio C. E. Tosatto.

Investigation: Marco Carraro, Anna Belligoli, Marta Sanna, Chiara dal Prà, Francesca Favaretto, Carlo Ferrari, Roberto Vettor.

Methodology: Francesco Reggiani.

Project administration: Roberto Vettor, Silvio C. E. Tosatto.

Software: Francesco Reggiani, Marco Carraro.

Supervision: Marco Carraro, Carlo Ferrari, Roberto Vettor, Silvio C. E. Tosatto.

Validation: Silvio C. E. Tosatto.

Visualization: Carlo Ferrari, Silvio C. E. Tosatto.

Writing – original draft: Francesco Reggiani, Marco Carraro, Anna Belligoli, Marta Sanna, Chiara dal Prà, Francesca Favaretto, Roberto Vettor, Silvio C. E. Tosatto.

Writing – review & editing: Francesco Reggiani, Marco Carraro, Carlo Ferrari.

References

1. Liu D. J. et al., "Exome-wide association study of plasma lipids in >300,000 individuals," *Nat. Genet.*, vol. 49, no. 12, pp. 1758–1766, Oct. 2017. https://doi.org/10.1038/ng.3977 PMID: 29083408
2. Golan D. E. and Tashjian A. H., Eds., *Principles of pharmacology*: *the pathophysiologic basis of drug therapy*, 3. ed., internat. ed., not Authorised for sale in United States, Canada, Australia and New Zealand. Philadelphia: Wolters Kluwer Health, Lippincott Williams & Wilkins, 2012.
3. Shapiro M. D., "Rare Genetic Disorders Altering Lipoproteins," in Endotext, De Groot L. J., Chrousos G., Dungan K., Feingold K. R., Grossman A., Hershman J. M., Koch C., Korbonits M., McLachlan R., New M., Purnell J., Rebar R., Singer F., and Vinik A., Eds. South Dartmouth (MA): MDText.com, Inc., 2000.
4. Spiliopoulou A. et al., "Genomic prediction of complex human traits: relatedness, trait architecture and predictive meta-models," *Hum. Mol. Genet.*, vol. 24, no. 14, pp. 4167–4182, Jul. 2015. https://doi.org/10.1093/hmg/ddv145 PMID: 25918167
5. Ramos-Lopez O., Riezu-Boj J. I., Milagro F. I., Cuervo M., Goni L., and Martinez J. A., "Prediction of Blood Lipid Phenotypes Using Obesity-Related Genetic Polymorphisms and Lifestyle Data in Subjects with Excessive Body Weight," *Int. J. Genomics*, vol. 2018, p. 4283078, 2018. https://doi.org/10.1155/2018/4283078 PMID: 30581838
6. Guo J. et al., "Systematic prediction of familial hypercholesterolemia caused by low-density lipoprotein receptor missense mutations," *Atherosclerosis*, vol. 281, pp. 1–8, Feb. 2019. https://doi.org/10.1016/j.atherosclerosis.2018.12.003 PMID: 30583242
7. Cazzaniga P. et al., "Computational strategies for a system-level understanding of metabolism," *Metabolites*, vol. 4, no. 4, pp. 1034–1087, Nov. 2014. https://doi.org/10.3390/metabo4041034 PMID: 25427076
8. Paalvast Y., Kuivenhoven J. A., and Groen A. K., "Evaluating computational models of cholesterol metabolism," *Biochim. Biophys. Acta BBA*—Mol. Cell Biol. Lipids, vol. 1851, no. 10, pp. 1360–1376, Oct. 2015.
9. van de Pas N. C. A., Woutersen R. A., van Ommen B., Rietjens I. M. C. M., and de Graaf A. A., "A physiologically based in silico kinetic model predicting plasma cholesterol concentrations in humans," *J. Lipid Res.*, vol. 53, no. 12, pp. 2734–2746, Dec. 2012. https://doi.org/10.1194/jlr.M031930 PMID: 23024287
10. van de Pas N. C. A. et al., "Systematic construction of a conceptual minimal model of plasma cholesterol levels based on knockout mouse phenotypes," Biochim. Biophys. Acta BBA—Mol. Cell Biol. Lipids, vol. 1801, no. 6, pp. 646–654, Jun. 2010.
11. R Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.," 2015.
12. Soetaert K., Petzoldt T., and Setzer R. W., "Solving Differential Equations in *R*: Package deSolve," *J. Stat. Softw.*, vol. 33, no. 9, 2010.
13. Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess and Ben Bolker, minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. 2016.
14. Marduel M. et al., "Description of a Large Family with Autosomal Dominant Hypercholesterolemia Associated with the *APOE* p.Leu167del Mutation," *Hum. Mutat.*, vol. 34, no. 1, pp. 83–87, Jan. 2013. https://doi.org/10.1002/humu.22215 PMID: 22949395
15. Yu H. et al., "Spectrum of Delta(7)-dehydrocholesterol reductase mutations in patients with the Smith-Lemli-Opitz (RSH) syndrome," *Hum. Mol. Genet.*, vol. 9, no. 9, pp. 1385–1391, May 2000. https://doi.org/10.1093/hmg/9.9.1385 PMID: 10814720
16. Pullinger C. R. et al., "Human cholesterol 7alpha-hydroxylase (CYP7A1) deficiency has a hypercholesterolemic phenotype," *J. Clin. Invest.*, vol. 110, no. 1, pp. 109–117, Jul. 2002. https://doi.org/10.1172/JCI15387 PMID: 12093894
17. Carraro M. et al., "Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI: CARRARO et al.," *Hum. Mutat.*, vol. 38, no. 9, pp. 1042–1050, Sep. 2017. https://doi.org/10.1002/humu.23235 PMID: 28440912
18. Monzon A. M. et al., "Performance of computational methods for the evaluation of Pericentriolar Material 1 missense variants in CAGI-5," *Hum. Mutat.*, Jul. 2019.

19. Gidding S. S. et al., "The Agenda for Familial Hypercholesterolemia: A Scientific Statement From the American Heart Association," *Circulation*, vol. 132, no. 22, pp. 2167–2192, Dec. 2015. https://doi.org/10.1161/CIR.0000000000000297 PMID: 26510694

20. White A., Tolman M., Thames H. D., Withers H. R., Mason K. A., and Transtrum M. K., "The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems," *PLOS Comput. Biol.*, vol. 12, no. 12, p. e1005227, Dec. 2016. https://doi.org/10.1371/journal.pcbi.1005227 PMID: 27923060

21. Verbeek R., Stoekenbroek R. M., and Hovingh G. K., "PCSK9 inhibitors: Novel therapeutic agents for the treatment of hypercholesterolemia," *Eur. J. Pharmacol.*, vol. 763, pp. 38–47, Sep. 2015. https://doi.org/10.1016/j.ejphar.2015.03.099 PMID: 25989132

22. Daneshjou R. et al., "Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges," *Hum. Mutat.*, vol. 38, no. 9, pp. 1182–1192, Sep. 2017. https://doi.org/10.1002/humu.23280 PMID: 28634997

23. Wickham H., *Ggplot2*: *elegant graphics for data analysis*. New York: Springer, 2009.

24. Angelo Canty and Brian Ripley, "boot: Bootstrap R (S-Plus) Functions," R package version 1.3–20, 2017.

25. Davison A. C. and Hinkley D. V., *Bootstrap methods and their application*. Cambridge; New York, NY, USA: Cambridge University Press, 1997.