SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Genetic Factors Associated with the High Prevalence of Reduced High-density Lipoprotein Cholesterol Level in the Hungarian General and Roma populations: A Comparative Analysis

by Péter Pikó

Supervisor: Róza Ádány, MD, PhD, DSc

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF HEALTH SCIENCES

DEBRECEN, 2021

Genetic Factors Associated with the High Prevalence of Reduced High-density Lipoprotein Cholesterol Level in the Hungarian General and Roma populations: A Comparative Analysis

By Péter Pikó, MSc

Supervisor: Róza Ádány, MD, PhD, DSc

Doctoral School of Health Sciences, University of Debrecen

Head of the **Examination Committee**: Endre Nagy, MD, PhD, DSc
Members of the Examination Committee: Ildikó Seres, PhD
László Márk, MD, PhD

The Examination is held online and starts at 9:30 on 17th February, 2021

Head of the **Defense Committee**: Endre Nagy, MD, PhD, DSc
Reviewers: Zsuzsa Bereczki, MD, PhD
István Kiss, MD, PhD, DSc
Members of the Defense Committee: László Márk, MD, PhD
Ildikó Seres, PhD

The PhD Defense is held online and starts at 11:30 on 17th February, 2021

Publicity is provided online. If you wish to participate, please indicate it in a message sent to the e-mail address egdi@unideb.hu by 2 p.m. on 16 February, 2021 latest. After the deadline, it is no longer possible to connect to the PhD defense due to technical reasons.

INTRODUCTION

Studying ethnicity-related biological, environmental, social, and psychological attributes have always been an essential component of multidisciplinary scientific studies related to the prevention, and the targeted intervention of diseases, with a special emphasis on the ones with a high prevalence among ethnic groups.

Roma ethnicity

The Roma population, which constitutes the largest ethnic minority in Europe, is one of the main topics of ethnicity-based researches. It is estimated that 10-12 million Roma are scattered across the European continent, indicating an accumulation in Central, Eastern and Southern European countries (mainly Bulgaria, Hungary, Slovakia, Romania and North Macedonia). The Roma are often concentrated in severely deprived regions, segregated colonies characterised by most unfavourable environmental conditions. According to the latest census (2011), about 3.2% of the total population is Roma in Hungary; however, their estimated representation is much higher and is reaching some 8.5% of the total population of the country. This discrepancy stems from the fact that due to the widespread prejudices they face, many Roma people do not declare their ethnic origin. Although studies on the Roma population face a number of challenges in terms of both data collection and methodology, the available data strongly suggest that Roma populations suffer from poor health, lower life expectancy and barriers in access to healthcare.

Possible genetic causes in the background of high prevalence of CVD morbidity and mortality among Roma

Comparative studies of the risk profile of Roma adults indicate that the Roma population carries a significantly higher risk burden of cardiovascular diseases (CVDs) than the majority population, regardless of the country where they live. Although some studies have also examined the possible role of genetic factors in relation to individual metabolic traits, the contribution of environmental and genetic factors to the increased risk of CVDs has not been determined in any of those studies.

A low level of plasma HDL-C cholesterol was identified as an independent risk factor for coronary artery disease in the 1980s, and it has been one of the most important risk factors used by clinicians to assess the present or future presence of high cardiovascular risk ever since. HDL-C metabolism has a strong genetic basis since 40-60% of the estimated interindividual variation in serum concentration is regulated at a genetic level.

It was found that the prevalence of metabolic syndrome and its components, especially reduced plasma High-density lipoprotein cholesterol (HDL-C) level (normal values are <1.03 mmol/l in males and <1.29 mmol/l in females) or treated lipid disorders were significantly more frequent in all age groups of the Hungarian Roma population compared to the Hungarian general population.

High-density lipoprotein cholesterol

High-density lipoprotein (HDL) is one of the five major groups of lipoproteins, which are complex particles composed of multiple proteins

which transport lipids (including cholesterol, phospholipids, and triglycerides) in the bloodstream. HDL is mainly secreted by the liver and the small intestines but it is the liver that secretes ~70–80% of the HDL to total HDL-C present in the plasma.

It is a well-known fact that HDL-C levels in the blood show an inverse relationship with the incidence of cardiovascular diseases. In addition to reverse cholesterol transport, HDL is also known to have other health effects (anti-thrombotic, anti-inflammatory, anti-oxidative, and pro-vasodilatory). Therefore, its measurable level in blood plasma serves as an overall indicator of an individual's risk to different non-communicable diseases.

The CETP promotes the transfer of cholesteryl esters from HDL-C (considered as antiatherogenic effect) to apolipoprotein B containing lipoproteins (considered as proatherogenic effect), including very-low-density lipoproteins, intermediate-density lipoproteins (IDL), and low-density lipoproteins (LDL). The reduced level of the CETP is associated with increased HDL-C and decreased LDL-C levels, which is a typically antiatherogenic lipid profile.

Hepatic lipase is a form of lipase with two functions, it serves as triglyceride/phospholipid hydrolase and it has a bridging factor for receptor-mediated lipoprotein uptake. Hepatic lipase is coded by the LIPC gene and is expressed mainly in hepatocytes and endothelial cells. It plays an important role in triglyceride hydrolysis in the blood by maintaining steady levels of IDL, HDL and LDL. Hepatic lipase and ABCL1 protein also help the transfer of free fatty acids from plaques in the arteries to HDL.

Genetic risk score calculations

Genetic variations describe the differences occurring among individuals. When a specific nucleotide at a particular position in the genome differs from the reference nucleotide, it is called a single nucleotide polymorphism (SNP). SNPs are the most common genetic variations that exist in the human genome, and we all carry about 4-5 millions of these variations some of which have a real effect on different traits (lipid metabolism, carbohydrate metabolism, pathways involved in blood pressure regulation, etc.) while others do not. The individual risk alleles associated with these traits have only a small effect on them, with most genotype relative risks being in the range of 1.1–2.0; consequently, the predictive value of a single SNP is very low, in general.

Calculating the polygenic/genetic risk score (GRS) is a very effective way of estimating the cumulative effect of selected SNPs on a biological outcome. However, in order to do these calculations we need independently published genome-wide association studies (GWAS) which estimate the full impact of SNPs. In addition to the many benefits of GRS, one of the critical points against it is that the impact of SNPs has been described mainly on populations of European descent and may not be applicable to ethnic groups with non-European origin such as the Roma.

Applicability of SNPs described in European populations to a population with a non-European origin

The effect of SNPs associated with lipid metabolism is lesser known in non-European populations. Most genetic studies dealing with ethnic minorities underperform as a result of relatively small sample sizes, what is more, most of the genotyping platforms are designed to optimally cover the genetic variants previously identified in European populations. Therefore, it cannot be excluded that the directions and/or magnitudes of the effect of SNPs on HDL-C concentration differ significantly in non-European populations when compared to those obtained in populations of European origin.

According to the NHGRI-EBI Catalogue of published genome-wide association studies, the majority of GWASs for HDL-C levels were performed in populations of European descent, some of them in other populations, and none of them in the Roma.

A comprehensive research on the effect of SNPs discovered in different populations is needed that would explain either the generalizability or the lack of them across populations before applying them in GRS calculations for the Roma population. The received generalization data will show whether the SNPs identified by GWASs are simply tagSNPs or they are more likely to be true functional SNPs across the different populations.

Haplotype analysis

Haplotype analysis is a tool for arranging alleles/SNPs that are inherited together on the same genes/chromosomes. Haplotyping involves an

identification process where the recombination events are located between different markers. In order to avoid multicollinearity, it is enough to examine one of the SNPs with a linkage disequilibrium (LD) above 0.8 during the GRS calculation. Nevertheless, haplotype analysis provides an opportunity to examine the combined effect of these SNPs on the observed trait and what is more, it can also be considered an extension of the GRS calculation.

Aims

The aim of our study was to define whether genetic susceptibility contributes to the higher prevalence of reduced HDL-C levels among the Roma in addition to the effects of unfavourable environmental factors (e.g. unhealthy nutrition, smoking, alcohol consumption) being more common among them.

In our study we aim to

1. conduct a structured literature search to identify polymorphisms that are thought to contribute to the occurrence/development of low HDL-C levels and to select the most relevant from the identified SNPs for genotyping in the Hungarian general and Roma populations
2. determine whether the effect of genotyped SNPs described in European-descent population-based studies can be transferred to the Roma population (with Asian origin)
3. calculate unweighted and weighted genetic risk scores for both study populations if the applicability of SNPs in Roma populations is demonstrated. To investigate by using biostatistical methods whether the calculated GRS correlates with HDL-C levels in both

populations or not. If the significant association between GRSs and HDL-C levels is confirmed, it will be possible to compare the genetic burdens of study populations.

4. perform a haplotype analysis to characterize the combined effect of SNPs in the same gene on HDL-C levels and to identify the unique co-occurrence of SNPs in the Hungarian general and Roma populations.

MATERIALS AND METHODS

Study design

The studies described in the present thesis involved 757 randomly selected Hungarian Roma individuals living in segregated colonies in North-East Hungary and 1783 individuals selected also at random from the Hungarian general population.

Sample of Roma living in segregated colonies

Roma participants were enrolled from two Northeast Hungarian counties (Hajdú-Bihar and Szabolcs-Szatmár-Bereg) where the majority of their segregations are found by using stratified multistage sampling. Briefly, segregated colonies of the Roma were identified by field workers with Roma origins within the framework of a project of the Hungarian Ministry of Environmental Protection and the University of Debrecen. The ethnic affiliation of individuals was assessed by self-declaration.

As a part of this health examination survey, medical histories and socio-demographic characteristics (age, sex, marital status, education, financial situation, number of people living in a household and employment) were

recorded and physical examinations were carried out (weight, height, waist circumference, blood pressure measurements) for each participant. Medical history information was collected by GPs. This included data collection on the occurrence of known hypertension, known carbohydrate metabolism disorder, known lipid metabolism disorder, known obesity, and known metabolic syndrome, as well as co-morbidities based on hospital or specialist diagnosis for stroke, peripheral vascular disease, acute myocardial infarction, ischemic heart disease, heart failure, diabetic nephropathy, asthma, AV block, sick sinus syndrome, and kidney disease. Blood pressure measurements were performed twice on the same day with a standard mercury sphygmomanometer at intervals of at least 3 minutes. The average of the two measurements was used in the analysis.

Blood samples were taken for laboratory tests (among others serum triglyceride, HDL-cholesterol, glucose levels were determined) and genotype investigations. A total of 757 DNA samples were eligible from Roma individuals with full records for genotyping.

In 757 of the 925 planned sample collections, informed consent was given, the questionnaire was completed and the physical examinations were performed.

Sample of Hungarian general population

The Hungarian General Practitioners' Morbidity Sentinel Stations Programme (GPMSSP) was established in 1998 by the School of Public Health, University of Debrecen and the National Public Health and Medical Officer Service. The main purpose of the GPMSSP is to monitor the incidence and the prevalence of non-communicable chronic diseases (e.g.:

lipid disorders, diabetes mellitus, ischaemic heart disease, hypertension, stroke, acute myocardial infarction, liver cirrhosis and malignancies of the colon and rectum, breast, prostate, cervix, and the respiratory tract) of great public health importance.

At the start of the program, samples were collected from four counties (Hajdú-Bihar, Győr-Moson-Sopron, Szabolcs-Szatmár-Bereg and Zala counties), which were later supplemented by four more counties (Komárom-Esztergom and Bács-Kiskun, Baranya, and Heves counties). The source population of the sample collection was selected from the Hungarian citizens aged 20-69 in 59 participating GP's practices. Samples were randomly selected from the files of residents and 2006 individuals were selected at random, proportional to the size of the practice, and stratified for age and gender. The resulting sample population can be considered representative of the Hungarian general population on the basis of age, gender and geographical distribution.

The study followed a methodology similar to that of the Roma one: in health examination surveys, medical histories and socio-demographic characteristics were recorded and physical examinations were carried out (weight, height, waist circumference, blood pressure measurements) for each participant. Blood samples were taken for laboratory tests (among them serum triglyceride, HDL-C, fasting glucose levels were measured) and genotype investigations.

DNA isolation

DNA was isolated by using a MagNA Pure LC system (Roche Diagnostics, Basel, Switzerland) with a MagNA Pure LC DNA Isolation Kit–Large

Volume according to the manufacturer's instructions. Extracted DNA was eluted in 200 µl MagNA Pure LC DNA Isolation Kit-Large Volume elution buffer.

Selection of SNPs

A systematic literature review on the PubMed, HuGE Navigator and Ensembl databases was conducted to identify SNPs most strongly associated with HDL synthesis and cholesterol transport using different combinations of the following keywords and terms: high-density lipoprotein cholesterol, cholesterol transport and synthesis, single nucleotide polymorphism, candidate gene, and meta-analysis. Those studies were used to identify the SNPs for our study which were original researches conducted on human samples and which applied either the candidate gene or the GWAS approach since both of them result in susceptible/protective alleles. The references of the selected articles and reviews were also examined to identify additional related studies. SNPs were selected from the above studies only if they were found to be consistently associated with plasma HDL-C levels in samples with a biostatistically acceptable size.

The literature search resulted in the selection of 130 SNPs influencing HDL-C levels. During the assay design, a pool of 33 SNPs was created for genotyping by the service provider (Mutation Analysis Core Facility of the Karolinska University Hospital, Sweden). Genotyping was successful for 21 SNPs which then became the subjects of our further studies. To avoid multicollinearity, only one SNP is selected from each identified LD block.

Genotyping

Genotyping was performed on a MassARRAY platform (Sequenom Inc., San Diego, CA, USA) with iPLEX Gold chemistry by a service provider (Mutation Analysis Core Facility of the Karolinska University Hospital, Stockholm, Sweden). Validation, concordance analysis and quality control were conducted by the service provider according to their protocols.

Comparison of the individual effects of SNPs on HDL-C levels in the Hungarian general and Roma populations

The individual effects of SNPs (size and direction) on HDL-C levels were estimated and compared between the HG and HR populations; furthermore, the data obtained from the available literature on the estimated effect of SNPs in independent European populations were compared to our results. To test for the effects of genotypes on the quantitative trait (HDL-C levels), linear regression analyses were performed and all results were adjusted for relevant covariates (age, sex, and BMI). The transferability of the results obtained was defined as having no significant differences in the effect of SNPs on HDL-C levels between the two study groups. This definition is based on the assumption that the effect in the Hungarian general population is consistent with the results previously obtained from populations with European ancestry.

Calculation of GRS and wGRS values

Unweighted and weighted genetic risk scores (GRS and wGRS) were calculated to define the combined effect of several HDL-C related SNPs. Individuals without full genotypic data were excluded from the studies.

In the GRS, each person was assigned a score based on the number of risk alleles carried. Thus, risk allele homozygotes were coded as genotype “2”, heterozygotes as genotype “1”, while “0” indicated the absence of the risk allele. When the effect allele was reported to be protective, the coding was “0” for effect allele homozygotes and “2” for reverse allele homozygotes. By using these codes, a simple count score (unweighted) was calculated as it is described by equation (1) in which Gi is the number of the risk alleles for the ith SNP. This model sums up all risk alleles of all loci as a summary score assuming that all alleles have the same effect.

$$GRS = \sum_{i=1}^{I} G_i \qquad (1)$$

In a weighted approach, rather than giving equal weight to all SNPs, we used their effects described in the literature selected as weight scores, and those with a greater impact on the outcome contribute more to the wGRS. Weights were derived from the risk coefficient for each loci based on the reported beta values of previous HDL-C associated studies. The estimated effect on HDL-C levels was available for 18 SNPs in 7 genes (ABCA1 gene: rs4149268; CETP gene: rs1532624, rs5882, rs708272, rs7499892, and rs9989419; GALNT2 gene: rs2144300 and rs4846914; KCTD10 gene: rs2338104; LIPC gene: rs10468017, rs1077834, rs1532085, rs1800588, rs2070895, and rs4775041; LIPG gene: rs2000813 and rs4939883; LPL gene: rs328). The calculation of the weighted genetic risk score is described by equation (2). In this weighted genetic risk score, weights (wβ_i) were derived from the risk effect for each SNP based on relative effect size

determined (in beta value) by previous studies. These weights (wβ_i) were multiplied by 0, 1 or 2 according to the number of risk alleles carried by each person (Xi).

$$wGRS = \sum_{i=1}^{I} w_{\beta\_i} X_i \qquad (2)$$

Two-sided t tests were applied to compare the distribution of GRSs between the study populations. To reveal if the association between genetic risk and ethnicity depends on the influence of demographic factors (age and sex) multivariate linear regression analyses were used in which GRSs were the dependent variable, while the type of population (HR and HG), sex and age were considered as independent variables. Quintiles were generated based on wGRS, and the association of quintiles with changes in HDL-C levels and the prevalence of reduced HDL-C levels was examined by trend analysis. The association of GRSs with reduced plasma HDL-C levels (as binary outcome) and plasma HDL-C levels (as continuous outcome) were further evaluated by using unadjusted regression models (Model I) and using regression models adjusted for age and sex (Model II) separately in HR and HG subjects. The possibility of ethnic differences in association patterns was further confirmed by analysing study populations (HR and HG) together in one multiple logistic regression model by adding ethnicity, age and sex as covariates. The applied thresholds for reduced serum HDL cholesterol levels were <1.03 mmol/l for men and <1.29 mmol/l for women. Based on the consensus definition of the International Diabetes Federation for the metabolic syndrome, those individuals were considered to have reduced

HDL-C levels whose parameters were below the above threshold or who were treated with lipid abnormality.

Haplotype block analyses

To avoid effects resulting from ethnicity related factors (e.g. environment and cultural effects), the study populations were examined together in a combined population, and then ethnicity was used as a covariate in the statistical models. All models were adjusted by relevant covariates (ethnicity, sex, age, BMI, systolic and diastolic blood pressure, fasting glucose levels, antihypertensive, antidiabetic and lipid-lowering treatment).

Bonferroni corrected p-values were used for haplotype analyses. The number of independent SNPs for each gene (CETP and LIPC) was defined by using the SNPsnap web-based tool, which was 4 in case of the CETP gene and 2 for the LIPC. For haplotype analysis, the result was considered significant if the p-value was less than 0.0125 for haplotypes in CETP and less than 0.025 for haplotypes in the LIPC gene according to the Bonferroni correction method.

Statistical analysis

Statistical tests were conducted with Stata (version 13) and PLINK (version 1.07) softwares. Mann-Whitney U test was used to compare the age, BMI and HDL-C levels of the populations. Prevalence data were compared by $\chi 2$ test. The Shapiro-Wilk test was applied for testing the normality of data. When necessary, dependent quantitative variables were transformed by using a two-step approach suggested by Templeton to reduce the effect of

non-normality. The Bonferroni adjustment was applied when genetic risk calculation was being performed ($p<0.0023$).

The existence of Hardy-Weinberg equilibrium (HWE) and the differences of allele frequencies for all SNP variants between the Hungarian general and Hungarian Roma populations were evaluated with $\chi 2$ test. Linkage disequilibrium between polymorphisms was analysed by using Haploview software (version 4.2). Generally, the conventional p threshold of 0.05 was applied.

RESULTS

Characteristics of the study populations

Only those subjects were considered whose genotype data were available in addition to clinical and demographic data. In total, 616 Hungarian Roma (HR) and 1497 Hungarian general (HG) samples were included in the baseline comparison. The proportion of males in Roma samples was significantly lower (HG: 47.3% vs. HR: 38.3%, $p<0.001$). The age distribution of Roma subjects was shifted towards the younger age groups and deviated from the HG one (HG: 44.16 vs. HR: 40.30, $p<0.001$). There were significant differences in mean plasma HDL-C levels (HG: 1.43 mmol/L vs. HR: 1.21 mmol/L, $p<0.001$), but the distribution of BMI values kg/m2did not differ significantly (HG: 27.43 kg/m2 vs. HR: 27.497 kg/m2; $p=0.898$). The proportion of subjects with reduced plasma HDL-C levels was significantly higher among the Roma compared to the Hungarian general population (HG: 28.20% vs. HR: 53.00%, $p<0.001$). Lipid-lowering (HG: 13.90% vs. HR: 10.80%, $p=0.050$) and antihypertensive treatment (HG:

29.90% vs. HR: 25.40%, p=0.042) was more common in the Hungarian general population than in the Roma. There was no difference in the prevalence of antidiabetic treatment between the study populations.

Results of linkage analysis

The SNPs of the CETP gene (rs708272 - rs1532624), the SNPs of the LIPC gene (rs4775041-rs104688017 and rs1077843-rs1800588-rs2070895) and rs2144300-rs4846914 SNPs of the GALNT2 gene were in complete linkage disequilibrium. Therefore, from among these genes, only rs708272 of the CETP, rs4775041 and rs1077843 of the LIPC gene and rs2144300 of the GALNT2 gene were taken into consideration in the calculation of the weighted genetic risk score, resulting in a wGRS composed of 13 SNPs.

Comparison of allele frequencies

Twenty-one SNPs were tested to determine whether the observed genotype frequencies were consistent with the Hardy-Weinberg equilibrium in both study populations. No significant deviation from HWE was observed. Differences between the HR and HG populations remained significant for 12 SNPs after the multiple test correction.

Six protective SNPs with a HDL-C levels increasing effect (rs1532085, rs1800588, rs2070895, rs2000813, rs4939883, and rs328) were more frequent in the Roma population, and three alleles (rs4149268, rs2144300, and rs38466629) were more prevalent in the HG one. Out of 7 susceptibility SNPs, four (rs7499892, rs4846914, rs2338104, and rs2548861) were more frequent in the Roma population and two of them (rs5882 and rs9989419) were more common in the Hungarian general one.

Comparison of individual effects (size and direction) of SNPs on HDL-C levels between the Hungarian general and Roma populations

Eleven SNPs were associated significantly with HDL-C levels in the HG population, while only 4 SNPs showed a significant effect on HDL-C levels among the Roma. Three of these 4 SNPs (rs1532624, rs708272 and rs7499892) are identical to those we identified in the CETP gene in the HG population, while rs2000813 in the LIPC gene showed a nominally significant association with HDL-C levels only in the Roma population.

As opposed to this, 8 of the 11 SNPs were found to be associated with HDL-C levels only in the Hungarian general population. Six of these genetic variants showed only a nominally significant association with HDL-C levels. In general, significant associations were more likely to be observed among the Hungarian general participants (with European origin) than among the Roma participants (with South Asian origin). Power analyses in the Roma group demonstrated decreased statistical power to detect the reported effects, which were largely attributable to its lower sample size (NHG=1497 vs. NHR=616).

Comparison of our results with data obtained in other studies on European populations

The effect direction and size, expressed as β-values, of the SNPs obtained from our regression analyses were compared to the effect size measures from literature data on European populations. For the SNPs we examined, there was no measurable difference in the effect size and direction when compared with the literature data in any of the study populations.

Comparison of genetic risk scores between study populations

The unweighted GRS calculated on the basis of 21 SNPs ranged from 11 to 31 in the Hungarian general population and from 13 to 31 among the Roma. GRS data showed normal distribution in both study populations. The mean value of the score was 21.5±3.3 in the HG population and 22.2±3.2 in the Roma group. The distribution of GRS in case of the Roma population shows a strong shift to the right (higher risk) compared to the Hungarian general one and a strong, significant difference ($p<0.001$) was measured between the two groups using two-sided t-test. The results showed a similar pattern when examined by gender, where individuals belonging to the Roma population in both sexes have, on average, a higher genetic burden than those belonging to the Hungarian general population (men: $GRS_{HR}=22.1\pm3.3$ vs. $GRS_{HG}=21.5\pm3.1$, $p<0.050$; women: $GRS_{HR}=22.3\pm3.4$ vs. $GRS_{HG}=21.6\pm3.2$, $p<0.001$).

The wGRS data were normally distributed in both study populations. The average wGRS value was 0.53±0.1 in the Hungarian general population, while this value was significantly higher in case of the Roma (0.57±0.1, $p<0.001$)

Both study populations were classified into quintiles based on wGRS. The first quintile included the genetically least vulnerable (wGRS: 0.15 - ≤0.30), while the 5th quintile included the most vulnerable individuals (wGRS: 0.75 – 0.88). The genetic risk of individuals in the third quintile is considered to be average. After this, it was examined how the distribution of individuals evolved between the quintiles in both populations.

Only half a percent of the Roma population belonged to the least vulnerable group compared to 1.83% in the Hungarian general population (p=0.025). Nineteen percent of the Hungarian general population and only 11% of the Roma one belonged to the lower-than-average genetic risk group. The 5th highest risk quintile included 2.61% of the HG population while 5.14% of the HR one (p=0.004).

Association of genetic risk scores with plasma HDL-C levels

Based on descriptive statistics, there is a clear trend for both the average level of HDL-C and the prevalence of reduced HDL-C by wGRS-based quintiles in both populations. Nevertheless, regardless of the quintiles, lower average HDL-C levels as well as higher representations of reduced HDL-C levels were more common in the Roma population.

The unweighted GRS showed a significant correlation with HDL-C levels in the unadjusted models in case of the study populations, where a one-unit rise in the GRS increased the risk of reduced plasma HDL-C levels (ORHG=1.065, p=0.001; ORHR=1.080, p=0.004). These results remained unchanged even after a correction for age and sex only (ORHG=1.066, p<0.001; ORHR=1.076, p=0.007), and for all relevant covariates (ORHG=1.075, p<0.001; ORHR=1.075, p=0.012) in linear regression analyses.

The weighted GRS showed a significant association with the outcome in both populations (ORHG=3.411, p=0.029; ORHR=10.385, p=0.005), and this correlation remained even after a correction for age and sex in logistic regression models (ORHG=3.665, p=0.021; ORHR=8.866, p=0.009). After correction for all relevant covariates, a similarly strong association between

wGRS and reduced HDL-C cholesterol levels was measured in both populations (ORHG=4.331, p=0.015; ORHR=9.303, p=0.012).

In case of unweighted GRS, the estimated effect per allele was similar in both populations (about 6-7%), while in case of weighted GRS, a one-unit change showed a greater effect in the Roma population than in the Hungarian general one regardless of the number of covariates (ORHG=3.411 vs. ORHR=10.382; ORHG=3.665 vs. ORHR=8.866; ORHG=4.331 vs. ORHR=9.303).

Similar associations of unweighted and weighted GRS with plasma HDL-C levels (as a continuous variable) were observed in both study populations. Similar to the results described above, in this case, the change of one unit in the Roma population had a greater effect on the development of HDL-C levels in both GRS and wGRS.

The effect of GRS and wGRS on reduced plasma HDL-C levels remained significant when the two study populations were analysed together, i.e. ethnicity was used as a covariate in the regression models (GRS: OR=1.076, 95% CI: 1.042 - 1.111, $p<0.001$; wGRS: OR=6.087, 95% CI: 2.306 - 16.071, $p<0.001$).

In addition to GRSs, we were able to identify a number of factors that influence the development of reduced HDL-C levels in both populations. The risk-increasing factors were Roma ethnicity, female gender, increased BMI, and increased glucose levels. Older age has a protective effect.

Haplotypes in the CETP and LIPC genes and their frequencies in the Hungarian general and Roma populations

Haplotype analysis involved different combinations of the 5 SNPs in the CETP (rs1532624, rs5882, rs708272, rs7499892 and rs9989419) and the 6 SNPs in the LIPC gene (rs10468017, rs1077834, rs1532085, rs1800588, rs2070895, rs4775041).

We identified 10 haplotype blocks in the CETP and six in the LIPC genes, the prevalence of which had been higher than 1% in the combined population (HR and HG together). A total 8 out of 10 in the haplotypes blocks in case of CETP (H1–H5 and H8–H10) and 4 out of 6 in LIPC (H1 and H4–H6) showed a significant difference in prevalence between the study populations. The H8CETP occurs almost exclusively in the Roma population (HR: 7.28% vs. HG: 0.14%; $p< 0.001$).

Association of haplotypes in CETP and LIPC genes with HDL-C levels in the combined study population

The most prevalent haplotype of the investigated two genes in the combined population (H1CETP: AGACG and H1LIPC: CTGCGG) was used as a reference for the comparative analysis on the relationship of haplotypes with HDL-C levels.

$H3_{CETP}$ ($\beta=-0.05$, $p=0.016$ and $OR=1.34$, $p=0.040$) and $H8_{CETP}$ ($\beta=-0.14$ $p=0.001$ and $OR=2.60$, $p=0.002$) have at least a nominally significant lipid-lowering effect on the outcome. The prevalence of $H3_{CETP}$ in the Hungarian general population (HG: 15.05% vs. HR: 11.45%, $p=0.015$), and the prevalence of $H8_{CETP}$ in the Roma population (HG: 0.14% vs. HR: 7.28%,

p<0.001) were found to be significantly higher in comparison with each other.

The $H2_{LIPC}$ (β=0.05, p=0.003 and OR=0.74, p=0.006) and $H3_{LIPC}$ (β=0.07, p=0.001) have a significant effect on HDL-C levels as continuous and binary outcome, and their prevalence did not differ significantly between the study groups. $H5_{LIPC}$ (β=0.09, p=0.004; OR=0.50, p=0.005) was significantly associated with HDL-C levels, and its prevalence was significantly higher in the HG population (HG: 5.48% vs. HR: 2.81%, p=0.006).

DISCUSSION

Low HDL-C levels is one of the most important predictors of cardiometabolic diseases, especially atherosclerosis. It has long been a well-known epidemiological fact that an individual's HDL-C levels is highly dependent on his or her age, sex, and lifestyle factors (characteristics of diet, levels and type of physical activity, smoking, and alcohol consumption). Furthermore, a low socio-economic situation has also been shown to predict adverse changes in a number of cardiometabolic factors. In addition to all these life-style related factors, the degree of genetic sustainability, as evidenced by numerous family and twin studies, must also be taken into consideration.

Significantly lower serum HDL-C levels were reported in the Roma population (independently from sex) compared to the general one, and reduced HDL-C levels were also found to be a lot more frequent among Roma children and adolescents as well.

These facts, together with the findings of our previous study which showed that the average HDL-C levels was significantly lower and the prevalence of reduced HDL cholesterol levels was higher in all age groups of the Hungarian Roma population than in the corresponding age groups of the Hungarian general one, strongly suggest that in the Roma there are genetic factors behind this phenomenon.

The present study is the first one that examined in detail the existence of genetic causes underlying the high prevalence of low HDL-C levels among the Roma. The SNPs in the current study are located in genes encoding proteins involved in HDL-C metabolism including lipid transfer. Twenty-one polymorphisms associated with HDL-C levels were genotyped, and differences in the prevalence of 15 SNPs proved to be significant when the two study groups were compared with each other.

The effect of SNPs, as estimated in independent studies, was validated and utilized to model and compare the genetic risk related to reduced plasma HDL-C levels in the Hungarian Roma and general populations. The individual effect of all twenty-one SNPs in 10 genes closely related to plasma HDL-C levels was evaluated in both study populations. Out of the 21 SNPs, 12 showed a significant association with HDL-C levels in at least one of the study populations - eight SNPs in the Hungarian general population and one among the Roma, and three in both populations. All the three common SNPs (rs1532624, rs708272, and rs7499892) are located in the CETP gene. There was no significant difference between the literature data and our results in any of the study populations. Thus, the studied SNPs are suitable for calculating genetic risk scores for both populations.

The genetic risk score was calculated in both populations in unweighted and weighted forms. For the unweighted GRS, all 21 SNPs were used, while only 13 were used to calculate the weighted GRS. In case of both unweighted (HR: 22.2 vs. HG: 21.5, $p<0.001$) and weighted (HR: 0.57 vs. HG: 0.53, $p<0.001$) GRS, the Roma population carried a significantly higher genetic burden compared to the Hungarian general one. Furthermore, it was demonstrated that both GRSs showed a significant association with HDL-C levels regardless of the effect of confounding factors.

Most of the SNPs with a significant effect on HDL-C levels are found in 2 genes, CETP and LIPC; thus, haplotype analysis was performed using SNPs in CETP and LIPC genes in the study. Three haplotypes in the CETP gene (H3, H5, and H8) showed at least a nominally significant association with decreased HDL-C levels and 3 in the LIPC gene (H2, H3, and H5) with elevated HDL-C levels. Haplotype 8 in the CETP gene, which has the greatest reducing effect on HDL-C levels ($\beta=-0.14$, $p<0.001$), was found almost exclusively in the Roma population (HR: 7.28% vs. HG: 0.14%). The 0.14% representation of this haplotype in the HG population is so low that the possibility that the two persons showed to have the haplotype 8 might belong to the Roma ethnicity (in the HG population selected at random the estimated number of Roma is about 140).

The results show that the majority of the susceptibility alleles were accumulated in the Roma population, and they override the slight predominance of protective alleles among them. The comparison of unweighted and weighted genetic risk score distributions showed that the Hungarian Roma carry a greater load of risk alleles compared to the Hungarian general population. The haplotype analysis confirmed that lower

HDL-C levels among the Roma can be partly explained by the special coexistence of polymorphisms in the CETP gene.

In conclusion, we demonstrated that the vast majority of the effects of targeted SNPs identified in European populations could be replicated in Hungarian general and Roma populations. Consequently, effect size measures obtained from the literature can be used for risk estimation not only in case of the Hungarian general population (European origin) but also in case of the Roma (South Asian origin). It was strongly suggested that the Roma population has a genetic susceptibility for reduced HDL-C levels. Further genetic susceptibility studies on HDL-C levels can utilize the effect size of SNPs identified in European populations to estimate the weight of genetic factors on this trait among Roma populations.

This is the first study carried out to investigate the possible genetic background of the high prevalence of reduced HDL-C levels among the Roma. The SNPs investigated include those that are mainly involved in HDL-C metabolism. By using DNA samples from study groups that are representative of the Hungarian general and the Roma populations, it was possible to estimate and compare the genetic risk. Although the genetic variants in our study could explain differences in plasma HDL-C levels and are stronger predictors of reduced plasma HDL-C levels in the Roma population compared to the Hungarian general population, it is still important to emphasise that further genetic researches on minority populations are required. At present, there is no sufficient support for the clinical application of gene-based prediction models in lipid disturbances, yet, there are promising signs pointing towards its future applicability. We must accept

that careful clinical trial programmes are needed in order to determine which HDL-raising therapeutic interventions may indeed exert protective effect.

MAIN STAEMENTS AND RESULTS

The aim of our study was to define whether genetic susceptibility contributes to the higher prevalence of reduced HDL-C levels among Roma in addition to the effects of unfavourable environmental factors (e.g. unhealthy nutrition, smoking, alcohol consumption) being more common among them.

Our aims were to

1. conduct a structured literature search to identify polymorphisms that are thought to contribute to the occurrence/development of low HDL-C levels, and to select the most relevant SNPs from the among the identified ones for genotyping in the Hungarian general and Roma populations
   - we were able to identify 21 SNPs that are related to the development of HDL-C levels based on literature data
2. determine whether the effect of genotyped SNPs described in European-descent population-based studies can be transferred to the Roma population (with Asian origin)
   - we demonstrated that the vast majority of the effects of targeted SNPs identified in European populations could be replicated in Hungarian general and Roma populations as well. Consequently, effect size measures obtained from the literature can be used for risk estimation not only in case of the Hungarian general population (European origin) but also in case of the Roma one (South Asian origin)
3. calculate unweighted and weighted genetic risk scores for both study populations if the applicability of SNPs in Roma populations is

demonstrated. To investigate by using biostatistical methods whether the calculated GRS correlates with HDL-C levels in both populations or not. If the significant association between GRSs and HDL-C levels is confirmed, it will be possible to compare the genetic burdens of study populations.

- the genetic risk estimated on the basis of both the unweighted and the weighted GRS calculation is significantly higher in the Roma population compared to the Hungarian general population. Both unweighted and weighted GRS showed a significant association with HDL-C levels and the risk of decreased HDL levels.

4. perform a haplotype analysis to characterize the combined effect of SNPs in the same gene on HDL-C levels and to identify the unique co-occurrence of SNPs in the Hungarian general and Roma populations.

- we were able to identify several haplotype blocks that had a significant effect on the development of HDL-C levels in both populations

Based on the results, genetic specificity can be clearly demonstrated in the background of the differences in the HDL-C levels of the Hungarian general and Roma population. For both populations, interventions to increase HDL-C levels should take into account the genetic background of individuals in addition to lifestyle and environmental risk factors.

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/6/2021.PL
Subject: PhD Publication List

Candidate: Péter Pikó
Doctoral School: Doctoral School of Health Sciences
MTMT ID: 10061786

List of publications related to the dissertation

1. **Pikó, P.**, Fiatal, S., Werissa, N. A., Begashaw, B., Rácz, G., Kósa, Z., Sándor, J., Ádány, R.: The Effect of Haplotypes in the CETP and LIPC Genes on the Triglycerides to HDL-C Ratio and Its Components in the Roma and Hungarian General Populations.
Genes. 11 (56), 1-13, 2020.
DOI: http://dx.doi.org/10.3390/genes11010056
IF: 3.759 (2019)

2. **Pikó, P.**, Fiatal, S., Kósa, Z., Sándor, J., Ádány, R.: Generalizability and applicability of results obtained from populations of European descent regarding the effect direction and size of HDL-C level-associated genetic variants to the Hungarian general and Roma populations.
Gene. 686, 187-193, 2019.
DOI: http://dx.doi.org/10.1016/j.gene.2018.11.067
IF: 2.984

3. **Pikó, P.**, Fiatal, S., Kósa, Z., Sándor, J., Ádány, R.: Genetic factors exist behind the high prevalence of reduced high-density lipoprotein cholesterol levels in the Roma population.
Atherosclerosis. 263, 119-126, 2017.
IF: 4.467

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

List of other publications

4. Llanaj, E., **Pikó, P.**, Nagy, K., Rácz, G., Sándor, J., Kósa, Z., Fiatal, S., Ádány, R.: Applicability of Obesity-Related SNPs and their Effect Size Measures Defined on Populations with European Ancestry for Genetic Risk Estimation among Roma.
Genes. 11, 1-13, 2020.
DOI: http://dx.doi.org/10.3390/genes11050516
IF: 3.759 (2019)

5. Fiatal, S., **Pikó, P.**, Ádány, R.: Application of Single-Nucleotide Polymorphism-Related Risk Estimates in Identification of Increased Genetic Susceptibility to Non-communicable Diseases.
In: Personalised Health Care. Fostering Precision Medicine Advancements for Gaining Population Health Impact. / Stefania Boccia, Róza Ádány, Paolo Villari, Martina C. Cornel, Corrado De Vito, Roberta Pastorino, Springer Nature Switzerland AG, Cham, Switzerland, 22-26, 2020.

6. **Pikó, P.**, Wenissa, N. A., Fiatal, S., Sándor, J., Ádány, R.: Impact of Genetic Factors on the Age of Onset for Type 2 Diabetes Mellitus in Addition to the Conventional Risk Factors.
JPM. 11 (1), 6-, 2020.
DOI: http://dx.doi.org/10.3390/jpm11010006
IF: 4.433 (2019)

7. Ádány, R., **Pikó, P.**, Fiatal, S., Kósa, Z., Sándor, J., Bíró, É., Kósa, K., Paragh, G., Bácsné Bába, É., Veres-Balajti, I., Biró, K., Varga, O., Balázs, M.: Prevalence of Insulin Resistance in the Hungarian General and Roma Populations as Defined by Using Data Generated in a Complex Health (Interview and Examination) Survey.
Int. J. Environ. Res. Public Health. 17 (4833), 1-22, 2020.
DOI: http://dx.doi.org/10.3390/ijerph17134833
IF: 2.849 (2019)

8. Diószegi, J., **Pikó, P.**, Kósa, Z., Sándor, J., Llanaj, E., Ádány, R.: Taste and Food Preferences of the Hungarian Roma Population.
Front. Public Health. 8, 1-11, 2020.
DOI: https://doi.org/10.3389/fpubh.2020.00359
IF: 2.483 (2019)

9. Soltész, B., **Pikó, P.**, Sándor, J., Kósa, Z., Ádány, R., Fiatal, S.: The genetic risk for hypertension is lower among the Hungarian Roma population compared to the general population.
PLoS One. 15 (6), 1-17, 2020.
DOI: http://dx.doi.org/10.1371/journal.pone.0234547
IF: 2.74 (2019)

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

10. Fiatal, S., **Pikó, P.**, Kósa, Z., Sándor, J., Ádány, R.: Genetic profiling revealed an increased risk of venous thrombosis in the Hungarian Roma population.
Thromb. Res. 179, 37-44, 2019.
DOI: http://dx.doi.org/10.1016/j.thromres.2019.04.031
IF: 2.869

11. Werissa, N. A., **Pikó, P.**, Fiatal, S., Kósa, Z., Sándor, J., Ádány, R.: SNP-Based Genetic Risk Score Modeling Suggests No Increased Genetic Susceptibility of the Roma Population to Type 2 Diabetes Mellitus.
Genes. 10 (11), 1-16, 2019.
DOI: http://dx.doi.org/10.3390/genes10110942
IF: 3.759

12. **Pikó, P.**, Fiatal, S., Kósa, Z., Sándor, J., Ádány, R.: A csökkent HDL-koleszterin szint a roma lakosság körében tapasztalt magas prevalenciájában szerepet játszó genetikai tényezők.
Népegészségügy. 96 (1), 39-50, 2018.

13. van El, C. G., Baccolini, V., **Pikó, P.**, Cornel, M. C.: Stakeholder Views on Active Cascade Screening for Familial Hypercholesterolemia.
Healthcare. 6 (3), 1-11, 2018.
DOI: http://dx.doi.org/10.3390/healthcare6030108

14. **Pikó, P.**, Fiatal, S., Kósa, Z., Sándor, J., Ádány, R.: Data to genetic risk assessment on high-density cholesterol level associated polymorphisms in Hungarian general and Roma populations.
Data Brief. 14, 354-359, 2017.
DOI: http://dx.doi.org/10.1016/j.dib.2017.07.053

Total IF of journals (all publications): 34,102
Total IF of journals (publications related to the dissertation): 11,21

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

07 January, 2021

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my PhD supervisor, Prof. Róza Ádány for her help and guidance during my PhD work.

I owe many thanks to Szilvia Fiatal for her help and support. I would also like to thank all of my colleagues involved in the publications on which this thesis is based. I am grateful to Zsuzsa Tóth and Beáta Soltész for their tireless work in the laboratories. My thanks also go out to my colleagues in the MTA-DE Public Health Research Group and Department of Public Health and Epidemiology, Faculty of Medicine from whom I received a lot of support.

I would like to thank my family and friends for their support and patience.

This work was supported by the TÁMOP-4.2.2.AA-11/1/KONV-2012-0031 [IGEN HUNGARIAN), TÁMOP 4.2.1. B-09/1/KONV-2010-0007 and GINOP-2.3.2-15-2016-00005 projects which are co-financed by the European Union and the European Social Fund, and by the Hungarian Academy of Sciences [MTA 11003, 2006TKI227). This research was also supported by ÚNKP-19-3 New National Excellence Program of the Ministry for Innovation and Technology.