.

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

Integrating Genetic and Conventional Risk Factors for Improving Coronary Heart Disease Risk Prediction

by:

Nayla Mohamed Gomaa Nasr

Supervisor:

Dr. Szilvia Fiatal

UNIVERSITY OF DEBRECEN

DOCTORAL SCHOOL OF HEALTH SCIENCES

DEBRECEN, 2023

Integrating Genetic and Conventional Risk Factors for Improving Coronary Heart Disease Risk Prediction

By Nayla Mohamed Gomaa Nasr, Master of Epidemiology (M.Epi)

Supervisor: Dr. Szilvia Fiatal

Doctoral School of Health Sciences - University of Debrecen

Head of the **defence board:** Prof. Dr. György Paragh, PhD, DSc.

Reviewers:

Dr. Melinda Pénzes, Ph.D.

Dr. László Róbert Kolozsvári, Ph.D.

Members of the defence board:

Dr. András Terebessy, Ph.D.

Dr. Rudolf Kolozsvári, Ph.D.

The PhD defence takes place in Lecture Hall of Department of Internal Medicine, Building 'A", Faculty of Medicine, University of Debrecen at 12:00 on March 8th, 2024.

Table of Contents

CHAPTER ONE

Introduction

1.1 Background

Coronary heart disease is one of the leading causes of death and morbidity as well as one of the leading causes of premature disability among adult men and women globally. In 2019, CHD was responsible for 197 million prevalent cases, 182 million DALYs, and 9.14 million fatalities. CHD was expected to affect 244.1 million individuals in 2020, with males (141.0 million) outnumbering females (103.1 million), while mortality rates from CHD were 112.37 per 100,000. These numbers increased steadily by 2021 and reached almost 185 million DALYs and 9.44 million deaths. The WHO estimates that by 2030, the number of annual deaths caused by CVD will reach nearly 23.6 million. Although the incidence and death rate of has decreased in many European countries, CHD remains a major source of morbidity and mortality in central and eastern European nations. CHD is also the leading cause of death in Hungary, with a mortality rate of more than 350 per 100,000 people in 2016, and 32,102 deaths expected in 2020.

CHD also known as ischemic heart disease or coronary artery disease, is caused by atherosclerosis. It is a condition of narrow or blocked-off main arteries supplying the myocardium with oxygen and nutrients, consequently, impairing the heart function due to insufficient blood flow. It is a silent disease that essentially begins with no symptoms, followed by the progression of lesion occurs over years. Manifestations of this disease include angina pectoris, myocardial infarction, and sudden cardiac death. Angina pectoris and myocardial infarction are the most common types of CHD. Angina pectoris is caused by insufficient heart blood flow and presents as chest pain, pressure, and discomfort. A myocardial infarction, often known as a heart attack, occurs when the heart receives insufficient blood flow, resulting in the death of heart muscles due to a lack of oxygen.

CHD development is a result of the complex interaction between modifiable and non-modifiable risk factors. Modifiable risk factors for CHD include dyslipidaemia, high blood pressure, hyperglycaemia, smoking, stress, and a sedentary lifestyle. Non-modifiable risk factors include age, gender, and genetic heritability.

For a better understanding of CHD pathophysiology, and clinical features it is important to demonstrate the cellular concept, structure of blood vessels, and implication of atherosclerosis in the large arteries. Initially, an artery consists of three distinguished layers including: tunica adventitia, tunica media, and tunica intima. In the beginning, endothelium dysfunction which is the fundamental process of atherosclerosis in the main arteries injured, which allows the migration of inflammatory cells that lead to the release of cytokines, and lipids. Consequently, foam cell and plaque formation occur and develop over time in the innermost layer of the main artery, thus decreasing blood flow and narrowing the arteries. The build-up of lipids can be chronic when plaque grows slowly over time or can be acute when plaque forms suddenly.

It is well established that genetic risk plays a pivotal role in etiology and contributes to individual susceptibility. Many risk factors of CHD are influenced by genes. Genome-wide association studies has successfully identified and validated several loci robustly associated with CHD among different populations. Modern technology powered by GWAS helps in identifying the single nucleotides polymorphism that are implicated in the high-risk populations propensity for CHD. This technology is also utilized to establish an appropriate intervention by estimating the alleles frequencies and determining the effect size of genetic risk scores.

Over the past few decades, considerable advances in the clinical diagnosis and curative procedure of CHD have been achieved; however, no significant reduction in the morbidity and mortality from CHD has occurred. The disease burden is increasing steadily, and the DALYs and YLL are also growing substantially. The morbidity and mortality of CHD vary among countries, populations, and specific ethnic minorities, and these variations might be due to large inequalities in socioeconomic status and education, and possibly due to differences in genetic susceptibility.

A wide range of preventive interventions for CHD is available for high-risk individuals through effective medication and comprehensive modification of an aforementioned risk factors, however, a comprehensive and precise assessment of CHD risk is needed to identify those high-risk subgroups. Both types of interventions can be significantly improved by accurate risk assessment. Accurate risk identification enables medical professionals to intervene in risk factor management prior to the onset of disease or critical conditions, thereby improving the patient's quality of life, resulting in a more cost-effective treatment strategy. Risk assessment of CHD requires thoroughness, completeness, and accuracy in obtaining information and

measurements for identifying subgroups with elevated risk and predicting the timing of disease onset.

Early identification of those individuals at high risk, considering both genetic and environmental risk factors for future CHD, is crucial for health promotion and prevention strategy. Such identification can lead to reduced mortality and morbidity and improved cost-effectiveness.

Prediction models play a pivotal role in assessing the risk associated with CHD among population groups, including Framingham, SCORE, QRISK, and ASSIGN models. These prediction models are generally classified as developmental model, validation, or combination of the two. Adoption of such models can assist in identifying high-risk CHD individuals generally as well in Hungary for early intervention.

1.2 Questions for Research

1.2.1 *Questions for Conducting a Systematic Review in the general populations*

1.2.1.1 What is the precise model for predicting the risk of CHD in the general populations?

1.2.1.2 Which biomarkers should be incorporated beside the conventional risk factors for CHD risk predictions?

1.2.1.3 Will risk prediction be significantly improved by genetic information?

1.2.2 *Questions for CHD/AMI Risk Prediction in Hungarian Populations*

1.2.2.1 Dose systematic coronary risk evaluation appropriate risk prediction model (alone or in combination with genetic risk score for CHD/AMI disease in the populations of Hungary?

1.2.2.2 Will the integration of (GRS and CRFs) help in assessing CHD/AMI risk in the Hungarian populations?

1.3 Research Objectives

1.3.1 *Objective for Conducting a Systematic Review of the Literature on CHD Risk Prediction Modelling Studies*

1.3.1.1 To summarize genetic and CRFs modelling studies for predicting CHD risk in the general populations.

1.3.1.2 To explore and identify the “optimal” risk prediction model for CHD risk by assessing its performance.

1.3.1.3 To evaluate the potential improvement in risk prediction by incorporating genetic information into the models.

1.3.2 Obje*ctives Regarding CHD/AMI Risk Assessment and Intervention in the Hungarian Populations*

1.3.2.1 To compare the sociodemographic characteristics, lifestyle factors, and clinical risk factors associated with CHD/AMI risk in Hungarian (Roma and general) populations.

1.3.2.2 To estimate and compare the allele frequencies of GRSs associated with CHD/AMI in the Hungarian Roma and the general populations.

1.3.2.3 To calculate and compare the genetic risk scores and thus genetic load in these Hungarian populations.

1.3.2.4 To assess and compare the role of the modifiable and non-modifiable risk factors in the development of CHD/AMI in the Hungarian populations.

1.3.2.5 To develop new models (by integrating GRS and CRFs) for predicting CHD/AMI risk among the Hungarian populations and assess their performance.

1.4 Research Hypothesis

1.4.1 *Hypothesis for Conducting a Systematic Review in the general populations*

1.4.1.1 Adding GRSs to CRFs-based model would improve the ability of these models to predict CHD events in the general populations.

1.4.2 *Hypothesis for CHD/AMI Risk Prediction in Hungarian Populations*

1.4.2.1 Hungarian Roma have more genetic variation and environmental risk factors for CHD than the general populations.

1.4.2.2 Including GRSs into the CRFs-based SCORE model would increase model capacity to predict CHD events in the Hungarian Roma population.

CHAPTER TWO

Literature Review

2.1 CHD Burden among the Hungarian Populations

CHD and stroke are the two leading causes of death in Hungary, accounting for one-third of all fatalities. Despite the fact that the CVD burden has steadily decreased in European countries over the last few decades, the rising prevalence of CHD in Hungary is creating public health concerns. Hungary has the highest CHD risk among European Union members. Half of all deaths in Hungary are attributed to lifestyle risk factors. A large number of risk factors, including poor diet, smoking, obesity, and physical inactivity are more prevalent in Hungary than other European countries. Hungary has a lower life expectancy than the majority of European countries, including gender disparities, and socioeconomic status, which are reflected in educational gaps and living standards. Socioeconomic deprivation is more common among the Hungarian Roma population than the general, and it is associated with health inequalities and life expectancy. It is commonly understood that genetics, physical environment, and access to and usage of the health care system, as well as social environment, all influence an individual's health state. CHD burden increases by social deprivation among countries, groups, and specific ethnic minorities, these disparities could be related to inequalities in socioeconomic status and education, as well as genetic vulnerability. The Hungarian Roma minority is a vulnerable and disadvantaged ethnic group in Hungary, according to the country's ethnic background; the majority of them are severely impoverished, live in inadequate housing facilities, and are below the poverty line. The poor health of Roma population is well documented, Roma population has relatively limited access to healthcare units as a result most Roma CHD patients have a worse cardiometabolic profile at the entry of care, which is characterized by a high risk of premature death. The Roma population is exposed to risk factors for CHD, including smoking, obesity, metabolic syndrome, diabetes mellitus, high triglyceride levels, and low HDL-C concentrations. As there is no data on CHD risk prediction for the Roma population is available, no previous study assessed genetic background of CHD/AMI risk and no model focused on the integration of CRFs and GRS, an accurate estimation of CHD risk is therefore required. After identifying people who live with higher risk we can provide them specific advice for more effective prevention.

2.2 Risk Factors for CHD

In general, the development of the CHD is the result of a complex combination of risk factors. The vast majority of which are linked to lifestyle and human activities. Despite advances in our understanding of the etiology of atherosclerosis and advancements in preventative efforts, CHD remains the major cause of premature adult's mortality and morbidity worldwide. There is still much we do not know about the specific trigger mechanism; however, a wide range of risk factors have been discovered, the majority of which are too general to identify the primary beginning process of atherogenesis. Risk factors such as hypertension, hyperlipidemia, diabetes mellitus and obesity, which are thought to be the main causes of CHD development, are also increasing significantly. Public health practices and clinical practitioners rely on these predictors to assess a patient's risk of developing CHD over the course of a given time period because there is no particular time when CHD emerges, and risk factors are frequently inadequately treated even in high-risk individuals. The common predictors used in CHD risk prediction models were listed below;

2.2.1 Age

Many chronic diseases including CHD, are significantly connected to increasing age. Although age standardized rates for prevalent cases, DALYs, and deaths due to CHD have decreased, the global prevalence of CHD is increasing due to populations growth, aging, as well as the health care system. The incidence of CHD rises dramatically with age in both men and women, and the prevalence of CHD rises with age in men in all age group compared to women until menopause, at the advanced age, women outnumber men, and the absolute number of female patients increases. By the end of the century, it is expected that more than 30% of the populations in Europe would be over the age of 65 years, which could have an impact on the prevalence of disease, the cost of healthcare. It is widely accepted that the risk of dying from CHD increases considerably with age. When integrating with CRFs in multivariable regression models, age is an independent risk factor for developing CHD, therefore, age was included in all models constructed for CHD risk prediction. This predictor, together with other risk factors, is used to assess an individual's risk of future CHD in various risk prediction scores which may be used to indicate the intensity and duration of exposure.

2.2.2 Sex

The most effective predictor for determining who is most likely to develop CHD is sex. This predictor was included in all CHD risk prediction models developed for the general populations. Sex variation exist in CHD pathogenesis, clinical manifestation, responsiveness to treatment, and outcome. These variations may be due to differences in risk factors, comorbidities that influence CHD presentation, and underlying biological differences. CHD are more prevalent in males than in females. In both sexes, CHD increases with age, but males get CHD at a younger age and have a higher predisposition to develop CHD than females. Females present with MI at a later age and have a higher burden of comorbidities than males, and they have a higher mortality rate after MI. It is expected that females born in ESC member countries will live 80.8 years and males 74.8 years. Previous study found that males have a 2 to 5 times higher prevalence of CHD than females among middle-aged persons, and this sex ratio varies between populations. Throughout adulthood, males had a greater CHD death rate than females, but the magnitude of the difference varied by age. The males-to- females CHD death rate ratio was 4-5 throughout middle age (30-64 years) and 2 beyond that (65-89 years).

2.2.3 High Blood Pressure

High blood pressure is the single most useful predictor for identifying people who are at high risk of developing CHD, often known as hypertension, is defined as a rate of raised systolic blood pressure of 140 mm Hg or higher and/or a diastolic pressure of 90 mm Hg or higher that is regarded to be above normal norms. Among adults in ESC member countries was 25%, prevalence of hypertension was lower in females compared to males. Several studies have shown that untreated hypertension is a major contributor to CHD, and the second leading cause of deaths worldwide. The development of atherosclerosis is accelerated by hypertension, especially when it is combined with hyperlipidemia. Many investigators indicated that systolic hypertension is actually more important as a predictor for CHD than DBP. All CHD risk prediction models created for the general populations incorporated (SBP) predictor.

2.2.4 High Blood Cholesterol Level

High blood cholesterol (LDL-C and TG) is a substantial risk factor for CHD death, particularly in young adulthood and middle-aged males. Hypercholesterolemia (>8.0 mmol/L) is frequently inherited, with heterozygous involvement. It is also a significant predictor of premature CHD

morbidity. Low-density lipoprotein cholesterol is the most important risk factor for CHD and the primary focus for treatment. This association is still considered to be etiologically significant, along with age, sex, smoking status, systolic blood pressure, and HDL cholesterol, as a crucial element of cardiovascular risk prediction models that are frequently used in clinical practice to determine a person's risk of CHD and to direct clinical decision-making regarding the start of statin therapy and other lipid-level regulating medications. Low-density lipoprotein cholesterol is the most critical factor in the occurrence of atherosclerotic CHD and is the main target for preventing it.

2.2.5 Smoking

Smoking is a primary cause of morbidity and premature mortality in avoidable diseases such as CHD, accounting for 25% of CHD deaths under 65 years of age and causing sudden deaths in males under 50 years of age. There is no such thing as a safe level of smoking; light smokers who consume one cigarette per day have approximately a 50% increased chance of developing CHD. Passive smokers are also at risk, and significant secondhand smoke exposure is associated with a similar relative risk of CVD as low active exposure. A cross ESC member nations, more than 20 % of adult's smoke on a daily basis, with the prevalence ranging between males (28.3%) and females (14.8%). Smoking contributes significantly to premature CHD and promotes atherosclerosis by increasing oxidation of low-density lipoprotein and causing coronary endothelial vasodilation damage. Smoking interacts synergistically with other risk factors such as hypertension and increased blood cholesterol. Smokers who stop can reduce their risk of CHD by 39% in 5 years, but the effect of quitting in lowering CHD risk takes at least five to ten years, and possibly 25 years, after quitting.

2.2.6 Diet

It's widely accepted that dietary factors contribute significantly to CHD risk, dietary fat, sugar, a higher intake of dietary sodium, and a low intake of fruit and vegetables are all major risk factors for CHD. Diet control can dramatically reduce the number of deaths caused by CHD.

2.2.7 Physical Inactivity

Physical activity has been demonstrated to reduce death and CHD risk in middle age by lowering blood pressure, losing weight, increasing insulin sensitivity, and lowering cholesterol.

2.2.8 Obesity

Obesity affects more than one in every three females and one in every four males in European member countries, with similar frequency in high- and low-income countries. Over the last 35 years, the prevalence of overweight and obesity has more than doubled, and it continues to climb in both industrialized and developing countries. Obesity with a body mass of 25 km/m^2 is a risk factor for CHD, with the lowest all-cause mortality reported at BMI, however, BMI of 20 km/m^2 increased all-cause mortality. BMI is a decent predictor of CVD risk, especially at higher levels, although there is strong evidence that visceral adiposity and liver fat are major risk factors at all levels of BMI. This helps to explain why the CHD risk profile in the overweight differs based on the location of adipose deposition. Some argue that, in addition to BMI reduction, waist circumference reduction as a proxy for visceral fat reduction should be a more essential objective for preventing CHD.

2.2.9 Diabetes Mellitus

Diabetes mellitus (type 1, and 2), defined as a chronic hyperglycaemic state characterized by a lack in the synthesis or action of the insulin hormone, which regulates glucose, lipid, and amino acid metabolism, is a key risk factor for CHD. A fasting blood glucose level of 70 to 99 mg/dL (3.9 to 6 mmol/L) is considered low or normal. Higher values could suggest pre-diabetes or diabetes, with prediabetes ranging from 100 to 125 mg/dL (5.6 to 6.9 mmol/L) and diabetes ranging from >125mg/dL or higher. People with DM have 2 to 4-fold higher risk of developing CHD compared with non-diabetic people in both sexes for all age groups. Incidence of CHD was found to be higher among diabetic patients, and about 75% of deaths in people with diabetes. Diabetes was discovered to work synergistically with other variables such as obesity, smoking, hyperlipidaemia, and hypertension in hastening the atherosclerosis process.

2.1.10 Genetic Risk Predictors for CHD

Despite the value of CRFs and the utility of risk estimate models, many high- and low-risk individuals are misclassified as low and high risk by CRF algorithms, resulting in overuse or underuse of preventative methods for predicting CHD in the general populations. CHD risk classification for primary prevention based solely on CRFs appears inefficient. The GRS computed from recently discovered genetic variants could provide a potential solution in cardiovascular primary prevention. Studies shown that genotyping the population with a

microarray containing these genetic risk variants, and genetic risk stratification based on the GRS is superior to that of conventional risk factors in detecting those at high risk and who would benefit most from statin therapy. The identification of genetic risk variants might lead the development of a novel therapy like the discovery of PCSK9 has led to the development of a novel treatment for high plasma LDL-C. Risk variants in the genome does not change over an individual's lifetime, it does not vary with time, and it can be easily detected by a single blood test, and a single variant can exert multiple influences. It is now universally accepted that CHD risk is known to be modified by the interaction of both multiple genetics and environmental components. GWAS has so far identified a hundred loci associated with many CVD risk and traits. Out of these, more than 97 single-nucleotide polymorphisms have been associated with CHD risk and myocardial infarction. SNPs on chromosome 9p21 have been consistency associated with CHD, variant at 3q22, and 6p24, 6q23, 6q26 and 12q24, also found to be a risk factors for developing CHD. Individual risk of CHD prognosis is determined by hereditary and lifestyle factors. GRS-based risk classification, followed by lifestyle adjustments or statin treatment, was found to be associated with a significant 40% to 50% reduction in cardiac events in the high-genetic-risk group. Previous study indicated that, incorporation of GRSs based stratification for primary and secondary preventions have several advantages over CRFs because it is independent of age and can be determined at birth or anytime thereafter. GRS is constructed from a list of common genetic variations associated with CAD on a risk-weighted basis. Each risk variant's weight is multiplied by the number of variants at that site (0 for absent risk, 1 for heterozygous (moderate risk), or 2 for homozygous (high risk), and the final score is simply the sum of the weighted dosage for each risk variant included in the GRS. Formerly, the association between some SNPs, and coronary microvascular function independently from CHD were defined; specifically, the role of adenosine triphosphate-sensitive potassium channels, which are the end effectors of several regulatory mechanisms for coronary flow reserves.

2.2.11 Other Predictors

Several studies have indicated that CRFs are insufficient for identifying patients at high risk of CHD, a novel biomarker was then added to models besides the CRFs in order to contribute to CHD improvement in models performance which includes beside the genetic marker for quantifying the added value of genetic biomarkers and family history of premature CHD,

coronary artery calcium, C-reactive protein, fibrinogen, and homocysteine, lipoprotein, cystatin c, and apolipoproteins.

2.3 Concept of Prognostic Models for CHD Risk

Prognostic models are used to estimate the probability of developing a particular outcome in the future with the aim of assisting clinicians in disease prediction and enhancing informed decision-making with the patients. These models in general, use two types of performance measures: discrimination and calibration. Prognostic models are more likely to be reliable and useful in practice when they are developed using a large, high-quality data set, based on a study protocol with a sound statistical analysis plan, evaluated a long-term outcome, and externally validated by using independent data sets. Despite the importance of predicting future CHD among initially healthy adults, the predictive accuracy of the models often seemed disappointing because most individuals who eventually suffered a CHD event were previously at average risk rather than high risk. In observational studies, data from the cohort, nested case-control, or case-cohort studies are recommended for prognostic modelling studies.
Several prognostic models, such the Framingham, SCORE, QRISK, QRISK2, and ASSIGN models, have been created in the recent decades to assess the risk of developing the CVD outlined below;

2.3.1 Framingham

The Framingham heart study was the first and most widely used risk prediction model for CHD, created by Framingham, a town in Massachusetts, USA in 1968. This model was built for predicting 10 years' risk of CHD, using three generations of residents includes 1968-1971, 1971-1975, and 1984-1987. Members of the generations were age (30-75) years, the second generation was an offspring cohort made up of children of the original cohort and their spouses. The original Framingham model included age, sex, LDL-C, HDL cholesterol levels, blood pressure level, hypertension medication, smoking, and diabetes mellitus. This study discovered important risk variables that predispose to the development of CHD, which may help in classify patients and prescribe statin therapy for those with high risk to develop CHD. Framingham risk score functions have overestimated the CHD risk in some populations leading to a concern that it may not be appropriate for other populations.

2.3.2 Systematic Coronary Risk Evaluation (SCORE), SCORE2, and SCORE-OP

The SCORE, SCORE2, and SCORE2-OP models are a risk models developed by the ESC for use in clinical cardiovascular risk management in European clinical practice, these models recommended for high and low risk regions of Europe including Hungary, SCORE model is based on data from 117,098 males and 88,080 females who participated in 12 European cohort studies between the ages of 40 and 65, and it estimates the 10-year risk of overall CVD death at baseline in (1972-1991). This risk assessment estimates fatal CVD events over a ten-year period based on integrated CRFs such as sex, age, TC, or TC/HDL-C ratio, SBP, and smoking status. SCORE provides calibrated risk estimation for total CVD events for low, moderate, and high-risk populations. The validity of this risk functions was analyzed with the area under the ROC curve and the Hosmer-Lemeshow test respectively, SCORE is also overestimated the CHD risk in some populations. SCORE2 risk prediction algorithm is a revised version of SCORE, developed to estimate the 10-year risk of first-onset CVD in the European population, in individuals without prior CVD or diabetes mellitus in the age range (40-69) years, using data from 45 cohorts in 13 countries. This model included age, smoking status, SBP, total- and HDL-cholesterol predictors, and it is recommended for apparently healthy people <70 years of age without a history of CVD, DM, CKD, genetic, lipid, or blood pressure disorders to estimate the 10-year fatal or nonfatal CVD, however patients with stablished CVD and or DM, CKD are to be considered at high or very high CVD risk. The competing risk adjusted SCORE2-Older Persons risk model is recommended in apparently healthy people aged 70 years or older. It was developed to estimate the 5- and 10-year risk of CVD in older adults in four geographical risk regions, with the models including age, smoking status, diabetes, systolic blood pressure, and total- and high-density lipoprotein cholesterol. In this model, the 10 years CVD risk were classified as; low to moderate (<2.5%), high (2.5-7.5%), and very high (≥7.5%) based on age categories (<50, 50-69, and ≥70). CVD risk is higher (≥7.5%) in apparently healthy people aged <50 years and having SBP (140 to 130 mmHG), but very high (≥10%) if people aged (50-69), with SBP (<140 to 130 mmHG), and LDL-C (<2.6 mmol/L, or <100mg/dL) level, and greater (≥15%) in people aged ≥70 years old.

2.3.3 QRISK 1, and QRISK2

QRISK is a CVD risk prediction model developed using data collected from general practice databases in the United Kingdom between 1993 and 2008. This risk stratification method

included 1.28 million participants in QRISK1, and 2.29 million in QRISK2, aged 30 to 74, calculated a 10-year risk of CVD. The CRFs included in QRISK1 were gender, age, TC to HDL-C ratio, SBP, smoking status, and diabetes, as well as indices of social deprivation, family history of CVD, BMI, ethnicity, chronic conditions, and antihypertensive medication, however, QRISK2 includes age, gender, ethnicity, deprivation, SBP, BMI, TC to HDL-C ratio, smoking, family history of CHD, antihypertensive medication, and some of the medical condition variables such as DM, CKD, atrial fibrillation, and rheumatoid arthritis.

2.3.4 ASSIGN

ASSIGN is a CVD risk prediction developed using the Scottish Heart Health Extended Cohort Study from the general populations in Scotland between 1984 and 1987. It included 6,540 males and 6,757 females, ages 30 to 75 years, it estimates the 10-year risk of overall CVD event. ASSIGN uses conventional risk factors including sex, age, TC, HDL-C, SBP, smoking, and DM, measures of area based index for social deprivation, and family history of CVD predictors.

2.3.5 PROCAM

PROCAM is a CVD risk prediction developed using health employee's databases (1978-1995). It is based on data from 18,460 males and 8,515 females, ages 20 to 75 years who participated in two separate scores calculate 10-year risks of major coronary events and cerebral ischemic events. The CRFs includes age, sex, LDL-C, DM, smoking and SBP. This model developed using Weibull methods, which allow extension of risk estimation to females and broader age range. PROCAM functions has miscalibrated for some European populations.

2.3.6 Pooled Cohort Studies Equations

Pooled Cohort Studies Equations is a CVD risk prediction model based on data from four Pooled prospective studies: ARIC, CHS, CARDIA, and Framingham, with populations baselines from 1987-1989 (ARIC), 1990 and 1992-1993 (CHS), 1985-1986 (CARDIA), 1968-1971, 1971-1975, and 1984-1987 (Framingham). It comprised 11,240 white females, 9,098 white males, 2,641 African-American females, and 1,647 African-American males between the ages of 20 and 79 to estimate a 10-year risk of a CVD incident. Predictors of lifetime risk

Age, gender, race, total cholesterol, HDL-C, SBP, antihypertensive medication, diabetes, and smoking are all CRFs.

2.3.7 CUORE

CUORE is a CVD risk prediction model that was developed between the 1980s and 1990s, and it includes 7,520 males and 13,127 females aged 35-69. It calculates the 10-year risk of developing an event based on CRF factors such as age, gender, SBP, total cholesterol, HDL-C, antihypertensive treatment, and smoking habit.

2.3.8 Globorisk

Globorisk CVD risk prediction model based on 8 pooled prospective studies includes Atherosclerosis Risk in Communities, Cardiovascular Health Study, Framingham Heart Study original cohort and offspring cohort, Honolulu Program, Multiple Risk Factor Intervention Trial, Puerto Rico Heart Health Program, and Women's Health Initiative Clinical Trial, included populations from 8 prospective studies from North America (1948-1993), included 33,323 males and 16,806 females, aged 40-80 years.

Most of the existing models are based on the Framingham model. Different markers were then added to this model as a response to deficiencies in improving performance, such as coronary artery calcification scores, C-reactive protein, fibrinogen, homocysteine, and apolipoprotein. Previous studies found that all three models based on the Framingham score; the Framingham Adult Treatment Panel III model, the Framingham Wilson model, and pooled cohort equations provide an incomplete prognosis of CHD events. However, two problems remain: first, there is no consensus about the most suitable and optimal model for predicting CHD in the general populations, and second, it is not clear which biomarkers or genetic markers associated with events should be incorporated into the risk model in addition to conventional factors.

For any novel CHD risk factor to be useful in a clinical setting, it must significantly enhance event prognosis based on easily measurable CRFs such as age, cholesterol level, blood pressure, or body mass index; thus, any such factor(s) must have a major impact on risk.

2.4 Estimating A Model Performance

There are several ways to evaluate prediction model performance, including R^2 statistic and Brier score to indicate over all model performance, discrimination ability using sensitivity,

specificity, and the AUC curve, or concordance (c) statistic and calibration measures via Hosmer-Lemeshow "goodness of fit" as described below;

2.4.1 Discrimination

Discrimination assesses the model's ability to distinguish between people who develop CHD events and those who do not. It is determined by the distribution of patient characteristics in the populations where the model is applied, which includes variables such as age, gender, clinical, and genetic data. Several measurements can be used to assess its including; 1. The concordance (c) statistic for survival is a rank order statistic related to the D statistic that reflects generalized linear regression models' discriminative capabilities, 2. The AUC or ROC curve for a binary outcome, which plots the sensitivity (true positive rate) against 1- (false positive rate) for consecutive cutoffs for the probability of CHD outcome. If the model predicted a higher probability for patients with CHD than those who do not, the c statistic or (ROC) value is 1.0; however, if the c statistic or (ROC) value is equal to 0.50, the model cannot discriminate between high risk and low risk individuals. Even though physicians should not utilize a model that fails to differentiate between high risk and low risk people, discrimination alone is insufficient to assess a model's prediction capability.

2.4.2 Calibration

Calibration (goodness or fit) refers to how well the model forecast matches the overall observed event rates, or how well the observed outcomes and predictions accord. It is recommended to report calibration performance. A good calibration should be present in a useful model. It may be effective in some patients but not in others. A poorly calibrated model will either underestimate or overestimate the disease outcome. When the average predicted risk is higher than the overall event rate, the algorithm overestimates risk in general, and overestimates risk in individuals at high risk by more than 20%. Underestimation happens when the observed event rate exceeds the average anticipated risk. Even if they overstated risk by more than 20% in high-risk people, this model would still be clinically valuable. The calibration intercept and calibration slope can be used to detect poor calibration. The Hosmer-Lemeshow goodness-of-fit test is the most commonly used calibration test, and it was extended for survival data by Grnnesby and Borgan test and Nam and D'Agostino test.

2.4.3 Other Performance Metrics

The likelihood function is a statistical method used to evaluate how well a model fits the data; it indicates how much the likelihood increases by the novel marker, and it measures using the likelihood ratio test, the Wald test in nested models, or log likelihood for binary outcomes in nonparametric models and machine learning. Other likelihood-based measures, such as the Akaike Information Criterion or the Bayes Information Criterion can be used. These are particularly valuable when non-nested models are used. The easiest technique to compare a model with or without a new marker is to validate it in completely independent, external data. Reclassification is used to show how many subjects are reclassified by adding a marker to a model. The net reclassification improvement computes the proportion moving up or down in risk strata in cases and non-cases separately. The overall NRI is a sum of improvement in each set. Integrated discrimination improvement is defined as the difference between the discrimination slopes of two models, one with and one without the new variable.

2.5 Prevention of CHD

Strong prediction for incidents of CHD by identifying high-risk subgroups. The American Heart Association developed seven criteria for a healthier life including exercise, maintain a healthy weight, educate yourself on cholesterol, abstain from smoking or using smokeless tobacco, eat a heart-healthy diet, maintain a healthy blood pressure, and educate yourself on blood sugar and diabetes. Exercise is commonly acknowledged to have a good impact on the majority of health outcomes, including CVD. Even at highly intense levels of exercise, the risk of death and morbidity is negligible, and in the vast majority of cases, the benefits outweigh the risks. The single most cost-effective CVD prevention strategy is quitting smoking, and some benefits can be noticed as soon as a few months have passed.

CHAPTER THREE

Methodology

3.1 Methodology for Conducting a Systematic Review for Predicting CHD/AMI Risk in the General Populations

3.1.1 Definition and Objectives

Recognizing the need for larger studies, we first performed a systematic literature review. It's a process of identifying and collecting available studies related to CHD risk factors. Our objectives were to provide an overview of multivariable prognostic models developed to predict the risk of CHD in the general populations, explore and identify the optimal models by evaluating how well they performed in estimating CHD risk, and provide researchers with prognostic models by outlining the optimal combination of predictors, including conventional and genetic risk scores, and biomarkers covariates.

3.1.2 Protocol and Registration

The protocol was registered in PROSPERO (ID: CRD42021234224). We conducted a systematic review based on the PRISMA guidelines by following the recently published Cochrane Prognosis Methods Group guidelines, by using the Checklist for Critical Appraisal and Data Extraction for the Systematic Review of Prediction Modelling statement for assessing the quality of the prognostic modelling studies. The Genetic Risk Prediction Studies. Statement was used to assess genetic prognostic modelling studies. We performed the search of using items of the PICO framework including P(opulation) for subjects (people) free of coronary heart disease, I(ntervention) for developmental prediction models, C(omparator) for validation prediction models, and O(utcome) for the incidence of CHD within a specified time interval. The developmental prediction model seeks to derive a prognostic model by selecting relevant predictors and statistically combining them into a multivariable model, and then quantifying the model's predictive power, validation modelling studies aim to assess and compare the predictive performance of an existing prediction model using new participant data that weren't utilized to develop a prediction model.

3.1.3 Information Sources, and Search

An intensive systematic search was conducted utilizing five databases including Embase, PubMed, Cochrane, Web of Science, and Scopus. We applied a human filter on 30 November 2019 to identify original articles of the developmental and/or validation of prognostic models describing the combination of conventional and genomic risk factors for incident CHD. We searched the databases using the following key search terms: *("validation" OR "prediction" OR "predict" OR "risk" OR "prognosis") AND ("ROC" OR "area under the curve" OR "c-statistic" OR "c statistic" OR "discrimination" OR "discriminate") AND ("coronary heart disease" OR "CHD" OR "coronary disease").*

3.1.4 Eligibility Criteria

We included all original articles describing the estimation of risk associated with CHD morbidity or mortality in individuals in developmental and/or validation modelling studies where the models' performance for predicting CHD in the general populations were available. Two study designs were included: nested case-control and cohort studies. Articles describing clinical models with intervention and studies describing the prediction models of CHD in individuals with certain health conditions, such as HIV, HBV, congenital heart disease, kidney failure, diabetes mellitus, hypertension and cancer were excluded.

3.1.5 Study Selection

Initially, two reviewers independently screened the titles and abstracts of all studies identified according to the keywords and inclusion criteria, and then duplicates were removed. After consensus, full-text articles were then obtained and examined for quality. If there was any disagreement regarding the article's inclusion, a third person evaluation was performed to reach a consensus. We compared the work of the reviewers using the Epi Info7 program developed by the Centre for Disease Control and Prevention to minimize bias. The Preferred Reporting Items for Systematic Review and Meta-Analysis flow chart summarizes the selection process. According to our objectives, we selected all articles describing the prognostic modelling studies for CHD in the general populations, in cohort design (n= 66) and nested case-control studies (n=6), these prospective studies allow the optimal documentation of the predictors and outcomes. Case-control studies are not suitable due to recommendations; therefore, they were

excluded from this study, leaving both logistic regression and proportional hazard analyses to describe the effect of the categorical predictors.

3.1.6 Data Collection Process

The two reviewers extracted the articles independently. The list of extracted items was based on the CHARMS and GRIPS statements for reviewing the prognostic modelling of conventional and genetic studies. We categorized the eligible full-text articles into three groups: (1) *Developmental studies* are such models commonly aim to identify important predictors by selection, combining them into a multivariable model, and then developing a final model and quantifying the predictive performance and validating the model internally using forms such as bootstrapping or cross-validation; (2) *Validation studies* with or without updating a model, aim at assessing and comparing the predictive performance of an existing prognostic model using new participants' data that were not used to develop the prognostic model; and possibly adjust or update the model in case of poor performance based on the validation data; and (3) *Developmental studies with external validation* in independent data.

3.1.7 Risk of Bias Across Studies

The two reviewers separately assessed the quality of the included articles. Based on our objectives, the strengthening of the report of observational studies in epidemiology guideline was utilized to assess the epidemiological quality of the included studies.

3.2 CHD/AMI Risk Assessment and Intervention in the Hungarian Populations

3.2.1 Study Design and Setting

Second, an observational cross-sectional study was conducted to predict CHD/AMI risk in Hungarian Roma and general populations. We compared the prevalence of CHD/AMI in both populations based on the collected data to determine which group is more likely to develop CHD/AMI, by relating CHD/AMI, associated with an independent variable of interest. We used preliminary data obtained in a complex health survey conducted by the MTA-DE Public Health Research Group that began in 2018, with the recruitment of 1000 participants from the Hungarian general (500), and Roma (500) randomly selected, aged 18 years or older. The study included 558 participants who provided complete genetic and phenotypic data. After completing the questionnaire, all participants were invited to a medical evaluation and blood

sample collection by their primary health care provider. Notably, Northeast Hungary is the geographical region where most of the Roma Hungarians live in segregated colonies.

3.2.2 Participants and procedure

Overall, 558 study subjects were enrolled from two Hungarian populations: the general population (n=279) and Roma individuals (n=279). They were selected randomly in the framework of a complex comparative health survey by MTA-DE Public Health Research Group in 2018, this group are focused mainly on the investigation of the health of Roma population. Initially, 92 segregated colonies were identified in a complex health survey; of these, 25 colonies were chosen at random using certified household lists from general practitioners. Following that, 20 families were chosen from each colony, and one individual aged 18 years and older from each household was interviewed face-to-face at the respondent's home by Roma University students who were supervised by public health coordinators. Participants who were interviewed were also invited for a physical examination and blood sample collection. The respondents' self-declaration indicated their Roma ethnicity. Hungarian general subjects were drown from the same source population. Data was collected through paper questionnaires, physical examinations, and laboratory testing. Blood samples for DNA extraction were also drawn from all the subjects. The details of the sample and data collection have been previously described. Subjects with complete genotype and phenotype data were included in this study. The details of the sample and data collection have been previously described. The Hungarian Scientific Council on Health Research committee approved the protocol (61327-2017/EKU). All participants provided written consent before their participation.

3.2.3 Variables, Data Sources, and Measurements

In general, demographic information, physical examination, and blood sample collection for genetic analysis and laboratory investigation data were collected.

CHD and AMI was confirmed when participants answered “yes” to one of the following questions: “Did you have CHD or AMI during the last 12 months?”, “Have you been diagnosed with CHD or AMI by a medical doctor?”, or “Have you received hospital treatment for CHD or AMI?”.

Blood pressure was measured during the physical investigation. Hypertension identified when participants answered "yes" to one of the following questions: "Did you have HTN during the last 12 months?", "Have you been diagnosed with HTN by a medical doctor?" or "Have you received a hospital treatment for HTN?". In addition to that, elevated hypertension was defined if the average SBP ≥140 mmHg or a DBP ≥ 90 mmHg based on the Fifth Joint National Committee Guideline. Thus, in case of measured raised blood pressure or hypertension in the history we considered that the subject has hypertension.

The lipid profiles data were determined in the laboratory (mmol/l), however, they were not employed in models' construction. HDL-C, TC and LDL-C were all found to be protective factors. The protective characteristics in the lipid profiles could be due to the low prevalence of CHD/AMI, or small enough sample size, or possibly due to the confounding effect. In addition to that an elevated TC was confirmed when participants answered "yes" to one of the following questions: "Did you have a high cholesterol level during the last 12 months?", "Have you been diagnosed with a high cholesterol level by a doctor?" or "Have you received a hospital treatment for a high cholesterol level?"

Diabetes mellitus was measured by laboratory examination. In addition to that DM was confirmed when participants answered "yes" to one of the following questions: "Did you have a DM during the last 12 months?", "Have you been diagnosed with DM by the medical doctor?", or "Have you received a hospital treatment for DM?".

Stroke was confirmed when participants answered "yes" to one of the following questions "Did you have a stroke during the last 10 months?", "Have you been diagnosed with a stroke by a medical doctor?", or "Have you received a hospital treatment for stroke?".

Chronic kidney disease was confirmed when participants answered "yes" to one of the following questions: "Did you have CKD during the last 12 months?", "Have you been diagnosed with CKD by a medical doctor?", or "Have you received hospital treatment for CKD?"

Smoking status (yes/no) as lifestyle risk factors was confirmed when participants answered "yes" to the following question: "Are you a current smoker?". Other environmental and lifestyles risk factors including air pollution, noise and neighbourhood characteristics,

physical activity, alcohol consumption were not a part of this study, however, it was investigated by the MTA-DE Public Health Research group.

3.2.4 SNPs Selection Procedure and Genotyping

Genetic variants (30 SNPs) were selected based on on their robust association with CHD/AMI risk in GWAS and had been extensively investigated in a systematic literature search of previous investigations. In general, twenty-five SNPs that were significantly associated with CHD at a genome-wide level in prior analyses were selected. Furthermore, an additional three variants were added based on a paper by Tikkanen et al., and two SNPs were selected from Schenkert et al., and Teslovich et al. These publications that had several overlapping SNPs were the most cited GWASs on CHD. All SNPs were successfully genotyped by the Mutation Analysis Facility, Clinical Research Centre, Karolinska University Hospital (Sweden).

3.2.5 Weighted and Unweighted GRS Constructions

SNPs were coded based on the number of risk alleles as follows: zero was assigned in the absence of the risk allele, subjects homozygous for the risk allele were coded as 2, and the single risk allele was coded as 1. The unweighted GRS was calculated by simply counting the number of risk alleles present for each SNP for every study subject, while the weighted GRS was computed by multiplying the risk allele score (0, 1, 2) carried for each SNP by the published effect size measure.

3.2.6 Statistical Analysis and Software Were Utilized

First, DNA samples, clinical data and the questionnaire responses were matched in order to link the data and avoid duplicate subjects. Individuals with any missing genotype and phenotype values or individuals who did not specify their gender were excluded before the statistical analyses to minimize the possibility of systematic errors. Data quality control guidelines were applied based on a previous publication.

Based on the normality distribution, the difference between the baseline demographic and clinical data were compared using Pearson's chi-square and two-sided students t-tests. The difference in means of the genetic risk scores between the two populations was examined using two-sided students t-tests. Additional analyses were performed with the goal of estimating and comparing the allele frequencies of 30 selected SNPs linked to CHD/AMI risk, as well as to

calculate and compare the genetic load. The allele frequencies of 30 SNPs associated with CHD in the Hungarian general and Hungarian Roma populations were analysed using Plink 1.07 software. The force-specific allele technique was used to ensure that the affected risk alleles were assigned to be first before running the commands in the allele frequency comparison analyses. The HWE test in Plink 1.07 was used to exclude any SNPs that failed to be in HWE. LD calculation was performed by using Haploview 4.2 software to examine whether there was any correlation between these SNPs. Bonferroni correction was performed.

Multivariable logistic regression analyses were conducted by using Stata 13 software to assess the possible interaction between the CHD/AMI when integrated with CRFs includes age, sex, HTC-Med, HTN-Med, smoking, DM, ethnicity, and smoking as independent variables. We created different models using the CRFs variables suggested by the SCORE model. In addition, we updated the model by adding some potential explanatory predictors, such as genetic risk scores and diabetes mellitus. HDL-C, TG, and LDL-C were not included in the final derived model.

Ethnicity were also included in the statistical models. To evaluate the potential value of the integration of CRFs and genetic risk score in risk prediction, the model's performance was assessed. First, the area under the receiving operating characteristic curve of models with and without the GRSs was computed. Second, the calibration by the Hosmer- Lemeshow goodness of fit test was measured. Marginal plot analysis was also used to predict the interaction between the genetic risk score, age, and sex for CHD/AMI risk prediction.

CHAPTER FOUR

Results

4.1 Results of the Systematic Review for CHD/AMI Risk Prediction in General Population

4.1.1 Study Selection, and Characteristic

The search strategy of our systematic review identified 7187 potential articles; 2328 duplicates were removed automatically by Endnote X7 software, and 838 articles were removed after exporting the Endnote file to the CSV file. A total of 2658 articles were excluded based on title and abstract not being related to conventional or/and genetic risk modelling of CHD, or prognostic modelling studies with subjects having comorbidities. In total, 477 full texts were included after the exclusion of 405 other studies, such as Suspected patients of CHD or CVD or stroke (n=280), case-control studies (n=73), cross-sectional studies (n=21), case report (n=2), treatment (n=1), poor-quality studies with no follow up (n=6) or blinded comparison (n=1), pooled analyses (n=9), and no study design information (n=12). Finally, 72 eligible articles were included in this review.

We identified (n = 48) articles concerning the developmental CHD risk prognostic models; 14 articles described the external validation of the models, and 10 articles described the combinations of developmental and external validation.

4.1.2 Developmental Risk Prognostic Models for CHD

4.1.2.1 Frequency of Models, Study Designs, and Study Populations

In general, 58 articles described more than 157 different models. Most of the prognostic models (*n* = 68, 94%) were developed using data from cohort studies. Most of the models originated in the United States and Canada (*n* = 32, 45%) or Europe (*n* = 26, 36%); few studies originated from Asia (*n* = 14, 19%), and no developmental modelling studies originated from African countries. Framingham risk models developed for the US population were used multiple times to derive a novel model for different populations and countries. Framingham models which developed by Wilson and/or D'Agostino was used in 30 articles, and Framingham Adult Treatment Panel III (2001 and 2002) was used in 9 articles. The SCORE risk-estimation model

developed for the European population was used seven times, PROCAM was used five times and QRISK2 was used two times.

In general, there was variation between the study populations regarding the age groups: seven models (10%) were developed for people with ages ranging from 30 to 74 years, and eight (11%) models were developed for the subjects with ages between 45 to 64 years, while the majority (*n* = 57, 79%) of the models used several different age groups.

Most of the models (*n* =53, 74%) targeted the general population (males and females), few models (*n* =17, 23%) were developed for males, and only two (3%) models were available for females. Regarding the inclusion and exclusion criteria in most studies (*n* =47, 71%), the researchers stated that participants with a history of coronary heart diseases, other diseases such as cancer (*n* =6, 9%), diabetes mellitus (*n* =10, 15%), or chronic medical conditions were excluded. participants who were taking lipid-lowering medication or aspirin were also excluded from several studies. Additionally, few studies (*n*=7, 5%) excluded participants because of race/ethnicity status, and one model had no information. In the modelling studies with genetic parameters, the investigators explicitly stated that they excluded study participants with no genotypic data (*n* = 11, 15%).

4.1.2.2 CHD Definition and Outcomes

We observed a wide variety of fundamental definitions of CHD disease as well as CHD outcomes among the general populations. There were almost 20 distinct classifications for the definition of CHD, however, the international classification definitions codes of CHD were reported in (n=28, 39.9%) models, but it showed heterogeneity, which described as ninth and tenth revision (codes 410-414) (n=18, 25%), hospitalization or death with any of the following primary diagnoses: acute MI and unstable angina and surgical codes(n= 2, 2.8%), 10th revision codes 121 (n=5, 7%), 9th edition (ICD-9) codes (410-414) or ICD-10 codes (I20-I25) (n= 2, 2.8%), and ICD-8 (n=2, 2.8%). The CHD outcomes as endpoints were also showed considerable heterogeneity; the majority (*n* = 42, 58%) of the prognostic models characterized CHD disease as an occurrence of CHD with no categorization, whereas some models (*n* = 27, 38%) specified the endpoints of CHD as (fatal or nonfatal) myocardial infarction, stable or unstable angina, percutaneous coronary revascularization or bypass grafting, or death due to CHD. There were more than eight classifications for the outcomes of CHD. Other outcomes

were identified, such as fatal/nonfatal CVD events (*n* = 1, 1%), and three models (n = 3, 4%) with no information.

4.1.2.3 Time Span, Follow Up and Duration in the Prognostic Modelling Studies

Prognostic models follow participants over a period of time and record whether a specific outcome occurs after the prognostic time origins, this period start at time zero and goes through the end of follow up that specified in the study. Duration involved in prediction models should be described accurately in the study. Between the zero-time and the end of the follow-up researcher should do at least one more data collection for accurate result. Follow-up time period in the reviewed prognostic models ranged between 3-30 years, 4 models (6%) predicted the incidence of CHD for less than 5 years, 33 models (56%) predicted CHD outcomes for 5-10 years, a longer (>10-15 years) follow-up was described in 29 models (40%), and the length of follow-up was longer than 15 years in a few models (n = 6, 8%).

4.1.2.4 The Candidate Predictors

The candidate predictors in general categorized into different factors subgroups such as demographic and anthropometric, genetic, biomarker, comorbidities, reproductive, behavioral, metabolic syndromes, and psychological factors. The major categories of the predictors used were conventional risk factors, genetic risk variables, and biomarker variables. Age and smoking as predictors for CHD were used in all the studies, total cholesterol level was reported in 67 (93%) models, HDL cholesterol level was used in 62 (86 %) models, diabetes mellitus and systolic blood pressure were used in 63 (87%) models, sex was included as a predictor in 57 (79%) models. Most of the models (*n* = 46, 63%) included a set of similar predictors, such as age, sex, smoking, total cholesterol, blood pressure, BMI, blood cholesterol/HDL cholesterol level, and diabetes mellitus. Other prognostic models included several different variables, such as hypertension (*n* = 25, 35%), family history of CHD and LDL-C (*n* = 27, 38%), triglycerides (*n* = 29, 40%), genetic risk score (*n* = 17, 23 %), C-reactive protein (*n* = 12, 16%), apolipoprotein B (*n* = 8, 11%), and coronary artery calcification (*n* = 6, 8 %). Treatment as a predictor for CHD was included in a few studies (*n* = 6, 8%), described as the use of antihypertensive/antidiabetic and lipid-lowering medications.

4.1.2.5 Sample Size and Number of Outcomes

It is beneficial to utilize a large sample size when creating a risk prediction model in order to prevent overfitting. According to the evaluated articles, the number of participants used to develop and validate the prognostic modelling studies were ranged between 112 and 268,315 (median 4,651); almost half of the models (*n* = 50, 69%) recruited their participants from multiple centres, while (*n* = 16, 22 %) recruited subjects from one centre, (*n* = 2, 3%) recruited from two centres, and (n=4, 6%) models did not describe the recruitment method. Regarding the study setting (primary or secondary health care centres), (n=45, 62%) models described that they selected the participants from primary health care centres, (n=4, 6%) models selected the participants from secondary health care centres, and (n=23, 32%) models described that they selected the participants from the communities by using a random selective sampling procedure. The number of outcomes that occurred during the follow-up period (CHD events) ranged between 56 and 203,666 (median 467), and (*n* =2, 3%) models did not report the number of outcomes.

4.1.2.6 Missing Data

Missing data for developed and validated models should be kept to a minimum, if its present in the model, researcher should avoid leaving out individuals with incomplete data, if they do, they should explain the strategy used to deal with the missing data. As a result, we identified the number of participants with any missing value as well as whether the subjects were censored, or whether migration happened to the participants during the follow-up were described in (n=27, 37%) models. Methods for handling the missing data were described for several models. Seventeen models explained that they excluded the participants with missing data before starting the analysis, whereas four models reported that they used the imputation resampling technique, and only one reported that they repeated the measurement. Genetic modelling studies reported that they excluded the participants if they missed information related to genetic data, ECG, and C-reactive protein, blood pressure, total serum cholesterol, fasting serum glucose, smoking status, and body mass index.

4.1.2.7 Modelling Method

In relation to the evaluation of the model type used to create a prognostic model; (n=47, 65%) models used Cox proportional hazards to develop the prognostic model for CHD; logistic regression was used in (n=17, 24%) studies. Furthermore, conditional logistic regression and lifetime survival analysis were described in some models (*n* = 6, 8%).

4.1.2.8 Models' Assumptions and Normality Distribution

Model assumptions underpin statistical approaches. One of the most important factors to consider when evaluating prediction accuracy utilizing discrimination and calibration measurements, both of which should be unbiasedly confirmed using bootstrapping or cross validation. As results, the distributional assumptions about the residuals as well as whether the researchers selected the right predictors in their models, (n=18,) models reported how they checked the assumption of the normality distribution using linear regression, seven models used Schoenfeld residuals to verify the proportional hazard assumption, and two models reported that they fit the models by Grambsch and Therneau.

The methods used for selection of the best predictors during multivariable modelling were a backward approach in (n=3, 4%) models, forward selection in 3 (4%) different models, Bayes information criterion (BIC) in (n=9, 13%) models, Akaike information criterion in (n=5, 7%) models, likelihood ratio test (LR) in (n=12, 17%) models, and Shrinkage or penalized estimation in (n=3, 4%) models.

4.1.2.9 Assessing Models Performance of the Reviewed Studies to Explore the Optimal Risk Prediction Model

Models' performance can be assessed usually by using different methods which included discrimination, calibration, and reclassification measures. When evaluating the models performance, discrimination and calibration are both crucial factors. Discrimination measured by AUC ROC curve, or concordance (or c) statistic; calibration measured by calibration slope, survival analysis, Grønnesby and Borgan test. Reclassification is measured by net reclassification improvement (NRI). The net reclassification index is a measure for evaluating the improvement in prediction performance gained by adding a marker to a set of baselines predictors.

The study of a statistical prognostic model's performance revealed significant heterogeneity, and discrimination measures of predictive performance were given for (n=72, 100%) models. The concordance index (Harrell's C statistic) or ROC curve was utilized by the majority of the models (n = 57, 79%); The D statistics were employed in seven (10%) of the models; (n=6, 8%) models reported lifetime CHD risks, but only one model characterized the log rank. Calibration measurements were provided in 29 (45%) of the models, Calibration slope and intercept were assessed in (n=3, 4%) models, and calibration plots were presented in (n=2, 3%)

models, Hosmer-Lemeshow was employed to examine the variation between observed and predicted rates in (n=20, 28%) models, and Grønnesby and Borgan was used to test the goodness of fit of 5 (7%) models. Classification measurements were employed in (n=54, 75%) models to evaluate risk predictions, (n=24, 33%) models applied sensitivity and specificity measurements, net reclassification improvement (NRI) was reported in (n=28, 39%) investigation, in (n=16, 22%) models, integrated discrimination improvement (IDI) was utilized to quantify how close prognostics were to the actual outcome, and clinical NRI was calculated in (n=3, 4%) studies to assess the improvement between the basic and extended models. In addition, Kaplan-Meier survival curves were employed in (n=16, 22%) models. In total, 45 (63%) of the 72 models were developmental, with most models (n=18, 25%) employing a random split of the dataset, bootstrapping (n=21, 29.2%), or cross-validation (n=4, 6%), and only a few models (n=6, 8%) using multiple imputation to impute the missing values on all predictors.

4.1.3 Validation Risk Prognostic Models for CHD

Model validations were performed on only 24 models: (n=10, 16%) studies were subjected to developmental validation, whereas 14 were subjected to external validation investigations. However, the ten models had been developed and validated in the same study using different populations. Three of the 10 models stated genetic risk scores, while the remaining 14 models were internally validated. Framingham models were the most commonly used by researchers: Framingham which Wilson and D'Agostino developed in 1998 was reported in 33 models, Adult Treatment Panel III Framingham 2002/2001 has been described in 9 models, five different investigations reported on the Framingham models suggested by Anderson in 1991 and Kannel in 1979. Other models used to predict CHD development that were identified in our analysis were SCORE (2003), which was reported in seven models. PROCAM was mentioned in three separate papers. QRISK2 (2008) was utilized in two models, whereas the reference models used in 20 other experiments were not specified.

The qualifying criteria utilized for the subjects engaged in the external validation modeling studies varied significantly. The models were developed and validated using different age groups. The majority of the models assessed discrimination ability solely using C-statistics, while just a few models described the calibration measure (n = 5, 7%).

4.1.4 Genetic Risk Prognostic Models for CHD

The genetic risk modelling studies identified in this review were used to determine whether the including genetic factors in addition to conventional risk factors based on the Framingham score, improved the CHD risk prognosis. The majority of the models (*n* =16, 22.22%), were developed in healthy populations in longitudinal cohort studies while one study employed a nested case-control strategy.

The majority of the genetic modelling research originated in Europe (*n* = 11, 15.28%) and the United States (*n* = 5, 6.94%), with only one model was from Asia. The European modeling studies included four studies from the United Kingdom, two models from Sweden and one model each from Norway, Spain, Switzerland, the Netherlands, and Scotland. there was no model for predicting CHD risk in Africa. Most of the models (*n* = 11, 15.28%) were developed with Caucasian populations.

The genetic models were recruited using healthy participants from multiple centres (*n* = 12, 16.67%). The recruitment periods were ranged from 1987 to 2007; seven models recruited participants between 1987 and 2001, five models recruited people between 2003 to 2007, and two models provided no information. In terms of age, the majority of the models (n=11, 15.28%) were developed for adults aged 45-75 years, three models for people aged 25-64 years, and two models had no information.

The majority of the investigations used healthy Caucasian populations of both sexes with complete genetic data. The authors excluded all subjects with missing genotype data, those with prevalent CHD or stroke, and those with no follow-up data. The number of participants ranged from 840 to 51,954, with a total of 183 to 3,217 events.

The models examined incorporated genetic risk scores with varying numbers of SNPs. There were a total of 230 SNPs reported in the articles. It ranged from 1 to 156 SNPs per article: four models used a small number of SNPs (1-19), while thirteen models used a large number of GRSs ranging from 38 to 156 SNPs. rs17114036 (gene *PLPP3*) has been reported in 12 articles; rs1122608 (gene *SMARCA4*) and rs3184504 (gene *SH2B3*) in 7 articles; rs9818870 (gene *MRAS*), rs67258870 (gene *DHRSX*), and rs501120 (unknown gene) in 6 articles; and rs7692387 (gene *GUCY1A1*), rs12413409 (gene *CNNM2*), rs9515203 (gene *COL4A2*), rs11556924 (gene *ZC3HC1*), rs11206510 (unknown gene), rs273909 (gene *SLC22A4*, and gene *MIR3936HG*), rs12190287 (gene *TCF21*), rs2048327 (gene *SLC22A3*), rs12526453 (gene *PHACTR1*), rs4252120 (gene *PLG*), rs2505083 (gene *JCAD*), rs974819 (unknown gene),

and rs9982601 (unknown gene) in 5 articles. Other SNPs were utilized less frequently. These SNPs were discovered to be associated with well-known phenotypic traits or biomarkers such as systolic blood pressure, total cholesterol, LDL-C, HDL-C, apolipoprotein-B, lipoprotein (a), plasma C-reactive protein, health behavioral factor (smoking). Family history of premature CHD was also investigated because plasma cholesterol concentration and hypertension are both heritable risk factors for CHD. The majority of the SNPs chosen (*n*=9) were discovered in genome-wide association studies and the CARDIoGRAMplusC4D (*n*= 4) investigation. Based on a literature analysis, one model stated that SNPs were incorporated.

Heterogeneity was found in the genotyping methods used for the analysis, which included TaqMan technology, and Illumina MetaboChip both of which reported in three different models, Affymetrix GeneChip (*n*= 2), custom-designed Affymetrix Axiom arrays and genome-wide arrays (*n* = 1), and MassARRAY (*n*= 1), as well as the other models had no information reported.

In the models reviewed, the majority (n = 12) weighted the GRS by multiplying the participants' allele score (1, 0, 1) by the SNP beta coefficient, while two models reported that they weighted the GRS by multiplying the number of risk alleles by the 'combined beta' of the CARDIoGRAMplusC4D meta-analysis and summing the products.

In the constructed models, the GRS variables were classified as tertiles, quartiles, and quantiles. Most models (n = 11) used the quartile as low GRS (quartile 1), intermediate (quartiles 2 and 3), and high (quartile 4) risk categories; tertiles were described in (n = 5) models as low GRS (tertile 1), intermediate (tertile 2), and high (tertile 3) risk categories; and only one model used quantiles as low GRS (quintile 1), intermediate GRS (quintiles 2 to 4) risk categories.

The models' follow-up period ranged from 5 to 19.4 years. Seven models projected CHD outcomes in >10-15 years, six models indicated a follow-up duration of 5-10 years, two models predicted a follow-up period of less than 5 years, and two models predicted a follow-up period of more than 15 years. The majority of the models (n = 12) reported collecting data using questionnaires, physical examinations, and laboratory diagnoses. Bootstrapping was detailed for n = 5 models, and genotype imputation was performed in five models.

The Framingham risk function developed by Wilson et al.1998 was used in the majority of the studies (*n* = 6) to predict the ten-year risk for CHD, the Framingham Adult Treatment Panel III was utilized in three models, and some models did not specify which Framingham model was used.

Genetic modeling study performance was reported in (n = 16) models, discrimination measures using the concordance index (Harrell's C statistic) or area under the receiver operating characteristic curve were reported in (n= 13) models, calibration measures were reported in (n= 10) models, Hosmer-Lemeshow goodness of fit was reported in (n = 5) models, and Grnnesby and Borgan were reported in (n= 2) models. Classification measurements were provided in (n=10) models, and calibration measures were reported in (n=10) models using the net reclassification improvement.

4.1.5 Optimal Risk Prediction Models for CHD Risk by Assessing Models Performance

Only two validation models compared performance ability in different populations, according to our review. The first model was a genetic risk modeling study based on the Framingham risk score that assessed performance using discrimination, calibration, and reclassification in three different populations: the ARIC study, the Framingham Offspring Study, and the Rotterdam Study. This work revealed conflicting results regarding model performance: the discrimination ability and reclassification showed significant improvement in the developed model but not in the validated models. The second study investigated whether adding coronary artery calcification as a measure predicting CHD risk could improve model performance, this model developed using three different groups form the Multi-Ethnic Study of Atherosclerosis, Heinz Nixdorf Recall Study, and Dallas Heart Study. The performance of these models was tested using discrimination and calibration, and the study found that adding coronary artery calcification to conventional risk factors resulted in considerable increases in risk prediction. There was also evidence of very good discrimination and calibration.

4.2 Integration of Genetic and Conventional Risk Factors for Predicting CHD/AMI Risk Among the Hungarian Populations

4.2.1 Descriptive Characteristics of the Participants

The overall characteristics of the enrolled participants were revealed; 279 from the Hungarian Roma population and 279 from the Hungarian general population were included in the current study. The mean (SD) age of the Hungarian Roma and general populations were 42.73± 12.99 and 44.14 ±12.12 years respectively, with the Roma population being younger. The findings revealed a significant difference between Hungarian Roma and the general population in terms

of sex distribution, height, weight, educational levels, job activities, and family members, with Hungarian Roma having a lower proportion of male individuals, lower education, lower economic activities, and having a large family's members. The prevalence rates of CHD, AMI, and stroke were lower among both groups, with no statistically significant difference between them, but were considerable higher among Roma. The results also indicated that SBP and glucose levels were likewise much higher in the general population, despite the fact that Roma had significantly higher prevalence of HTN-Med and HTC-Med use and significantly lower HDL-C levels than the general population. Furthermore, Roma was more likely to be a current smoker and were likely to be exposed to chronic diseases such as DM, HTN, and CKD.

4.2.2 Frequencies and associations of the individual genetic variants associated with CHD risk

Following the HWE test, one SNP (rs12413409) in the Hungarian Roma community showed deviation and was thus eliminated from further GRS computation. The finding of the allele frequencies of 30 SNPs which are associated with CHD risk in the Hungarian populations were consistent with previously published evidence.

Even after the Bonferroni adjustment, it was shown that nine SNPs had statistically differing prevalence in the two research populations: Six SNPs were more prevalent in the Hungarian general population, including rs2306374 (gene MRAS), rs9818870 (gene MRAS), rs12190287 (gene TCF21), rs10455872 (gene LPA), rs3184504 (gene SH2B3), and rs9982601 (gene KCNE2), while three SNPs were more frequently among Roma individuals including, rs17609940 (gene ANKS1A), rs2259816 (gene HNF1A), and rs12936587 (gene RASD1).

The connection based on genotyped alleles counts for predicted linkage disequilibrium was studied for the 30 SNPs predisposed to CHD in Hungarian general (n=279), and Roma Hungarian populations (n=279). The nonrandom association between the alleles in the 30 CHD loci was measured using an alternative LD color scheme standard (D)/LOD. The results revealed no correlation between the SNPs.

The mean GRS and wGRS were greater in the general population than in the Roma community (27.27 3.43 vs. 26.68 3.51, p value=0.046 and 3.52 0.68 vs. 3.33 0.62, p value=0.001, respectively).

4.2.3 Multivariable regression analyses for CHD/AMI

The odds ratio of CRFs according to SCORE-based models for CHD/AMI risk prediction revealed that age and elevated cholesterol or therapy for high cholesterol level were associated significantly with CHD/AMI risk in the Hungarian general population ((OR = 1.18, p-value = 0.046, 95% CI 1.00-1.17) and (OR= 4.90, p-value = 0.032, 95% CI 0.24-0.22)), respectively, while HTN-Med showed a significant association with CHD/AMI risk only among the Roma population (OR = 7.85, p-value = 0.001, 95% CI 2.41-25.55). When we combined the study population and used ethnicity as a possible predictor: age (OR = 1.05, p-value = 0.030, 95% CI 1.01-1.10), HTC (OR = 3.55, p-value = 0.007, 95% CI 1.42-8.90), and HTN-Med (OR = 4.79, p-value = 0.001, 95% CI, 1.90-12.07) were associated significantly with CHD/AMI in the Hungarian populations.

Although ethnicity did not prove to be a significant predictor, Roma seems to have a higher risk of developing CHD/AMI independently from conventional risk factors. Although no significant association was observed between CHD/AMI and GRS/wGRS in these models, the ORs revealed a proportionally increased risk in the second and third tertiles in the general population.

The addition of GRS and wGRS tertiles to CRFs revealed that age, HTC-Med, and HTN-Med were significantly associated with CHD/AMI in the combined population independent of the effect of ethnicity. Although ethnicity was not shown to be a significant and independent risk factor by itself, the results suggest that the Roma population has a higher risk of developing AMI/CHD. By including DM as a reasonable new predictor to the updated SCORE-based genetic models, age, HTC-Med, and HTN-Med remained significant risk factors for developing CHD/AMI. However, the GRSs still did not show a significant association with the CHD/AMI. Several models were developed utilizing the data from the lipids profiles of, but these studies did not provide useful information. Furthermore, due to confounding, it was not useful to include these predictors for CHD/AMI risk prediction in the Hungarian populations. HDL-C was show to have a significant association with CHD/AMI in the Hungarian general population (OR=0.95, p value 0.009, 95% C (0.91-0.99)) in Framingham model, with no statistically significant association of HDL-C predictor among Roma, however, HTN-Med revealed statistically significant (OR=8.16, p value 0.001, 95% C (2.40-27.67)). TC was likewise associated with CHD/AMI risk in the Hungarian general population (OR=0.47, p value 0.028, 95% C (0.24-0.92)). TG was not associated with CHD/AMI in all population in the included

model in SCORE basic model, with no statistically significant association among Roma, and PROCAM risk model indicated similar results.

4.2.4 ROC curve analyses

The model performances (discrimination, calibration, and risk classification) revealed that; The AUC curve estimates LROC for CHD/AMI risk prediction when integrating the CRF basic, CRFs with GRS, and CRFs with wGRS models, which were 0.8149, 0.8346, and 0.8160 in the Hungarian general population and 0.8616, 0.8549, and 0.8674 in the Roma population, respectively. When DM was added to the models, minimal improvement occurred in the AUC values in the general population, and no significant improvement was observed among Roma individuals. Considerable improvements in the AUC value were observed when the CRF basic model with DM was combined with GRS in the Hungarian general population (from 0.8299 to 0.8400). For the Roma population, the CRF basic model (without DM) showed the greatest improvements in the AUC value (0.8616 to 0.8674) compared to the CRFS+wGRS model. In the combined populations, the CRF basic model with DM and ethnicity showed the greatest AUC value (LROC=0.8525) compared to the basic model (without DM). The results of the Hosmer-Lemeshow tests indicated that all the models had a good fit (p value≥0.11).

4.2.5 Marginal Plots Analyses

In the Hungarian general population, CHD/AMI risk was low among males (margin= 0.001, p= 0.707; 95% CI: -0,001-0.001) and females (margin= 0.006, p= 0.516; 95% CI: -0,013-0.026) subjects between 18 and 45 years of age. The risk of the trait increased after the age of 46 years among females (margin= 0.028, p value= 0.048, 95% CI 0.000-0.056) and after 54 years among male subjects (margin= 0.065, p value= 0.039, 95% CI 0.003-0.126); thus, female subjects were predicted to develop CHD/AMI earlier than male subjects. For the Roma population, the marginal plot shows that CHD/AMI risk is low for younger subjects (males and females between 18-40 years of age; margin= 0.03, p= 0.397; 95% CI: -0,043-0.109 and margin= 0.003, p= 0.387; 95% CI: -0,004-0.011, respectively). The risk starts to significantly increase at the age of 41 years for both males and females (male subjects: margin= 0.080, p value= 0.047, 95% CI: 0.001-0.159; female subjects: margin= 0.024, p value= 0.045, 95% CI 0.000-0.048 for age 41 years). The prediction of CHD/AMI risk in the combined population revealed that the risk is higher for the Roma population. The risk becomes significant at 34

years of age in the Roma group (margin= 0.022, p value= 0.046, 95% CI: 0.000-0.043), while for the general population, the risk becomes significant at the age of 44 (margin= 0.020, p value= 0.046, 95% CI: 0.000-0.041).

The prediction of the marginal plot revealed that CHD/AMI risk interacted by age and gender, the younger population is at low risk of developing the disease (18-40) years, adult's male of both Hungarian general and Roma are more likely to express CHD/AMI than adult's female subjects; males of Hungarian general population expected to develop CHD/AMI risk in ages (54-80) years however, Roma male develop CHD/AMI risk earlier in ages (41-60) years. Female subjects of the Hungarian Roma also have a greater risk of developing CHD/AMI compared to female subjects of the Hungarian general, the risk among Roma female begins at ages 41-70 years, and at 47-58 years among the females of the Hungarian general population.

CHAPTER FIVE

Discussion

The significant information from the systematic literature review of CHD risk prediction modelling studies among general populations, and the main finding from the original study of CHD/AMI risk prediction among the Hungarian population were both discussed in this chapter.

5.1 Discussion of the main finding of systematic literature review of CHD risk prediction modelling studies among general populations

Our comprehensive assessment of the literature reveals that a number of prognostic models have been developed for estimating the risk of CHD in the general healthy population, and the basic foundation for the predictors is the Framingham model. The Framingham model is the most widely used model for predicting CHD risk in the general population, despite the fact that it overestimated and underestimated the CHD risk in various populations. There is still no consensus about the best model(s). By include models with genetic risk components in addition to conventional models, our review offers some fresh and organized knowledge concerning CHD prognostic modelling studies.

The trend of CHD mortality and morbidity has reportedly continued to rise over time, especially in low- and middle-income countries, despite an increase in the number of publications in the medical literature on clinical prognostic modelling studies combining multiple predictors for identifying high subgroups at increased risk of CHD, the GWAS has facilitated a better understanding of the causal risk factors to CHD risk. However, translation of the acquired knowledge is lacking, and research among various populations does not support the use of genetic data in health and public health practice, it has been shown previously that data based on the use of genetic risk for calculating CHD is limited.

It was expected that incorporating a genetic risk score into conventional risk-factor-based models would increase their ability to predict the onset of CHD in the general population. We discovered that, with the exception of one notable model, the majority of the genetic modelling studies had not yet been externally validated in various populations, despite the fact that, the majority of these studies have indicated improvements in performance. Demonstrating the good performance of the generated models is not sufficient, improvements must have confirmed in different populations to ensure the generalizability. It has been previously

demonstrated that models should not be recommended for clinical use until external validity is established, because prediction model performance in new patients is frequently lower than in the development population.

Healthcare providers and policymakers feel that the prevention by identifying high risk subgroup and assisting them in making lifestyle changes prior to the onset of CHD will have a greater impact on reducing CHD mortality and disability than treatment, and this requires robust reduction of risk factors through the accurate estimate of the population at risk. Clinicians are also willing to identify appropriate guidelines for making quick decisions in everyday medical practice in order to improve patient outcomes as much as possible. They can achieve improved outcomes while lowering total healthcare expenditures by using precise prognostic and predictive model(s). Before beginning to develop a risk prognostic model, it is critical to consider whether a new model is required. One recommended solution would be the validation of the existing models instead of constructing several new models. It is difficult to translate and disseminate the results of the recognized model into clinical practice, and no prognostic application is suitable for adoption in normal daily practice.

Our review identified variance in model geographical location: most models were developed and validated in populations from the United States of America, Canada, and European. According to WHO, low- and middle-income countries account for more than three-quarters of all CHD deaths. Our review confirmed that no prognostic modelling studies for CHD originated in developing countries, for example, when we performed the search for this review, no prognostic model for people from Africa had yet been developed. Only a few studies from Asia were conducted.

The majority of participants were recruited from primary care and community settings, although the selection methods was not fully documented. Most of the studies did not specify which sampling approach was utilized, and only a few indicated the random method of the participant's selection. Sampling techniques are an important strategy for obtaining representative target populations; researchers should be more precise in this stage in order to avoid selection bias and deliver relevant information to physicians. Although few research reported consecutive participant's selection, but in some other studies, participants were selected nonconsecutively and thus increased the risk of bias due to selective sampling, it has been already demonstrated that, if the participants in the sample are not well defied or

representative, such data necessarily contain selection biases in their collection, and model building must take such difficulties into account.

Although age is the most common risk factor considered in CHD risk prediction and affects the two sexes for developing CHD, there was great variation between the developed and validated models in terms of the participants age groups; the validation studies were performed in people outside the age range of the developmental studies, and most of the developed models predicted the risk in the elderly population compared with the validated models, which may affect the number of CHD outcomes and may influence the model performance, it was previously indicated that an incorrect interpretations of basic epidemiological statistics such as age-incidence curves and hazard ratios may occur from failing to account for unobserved individual variation. Researchers should specify the age group by sex because females have a lower risk of developing CHD than males.

Although disparities of CHD risk depend on factors such as age, gender, and geographical location of populations, only a few researchers are focusing on females and middle- and low-income countries possibly due to data availability, population needs, or priorities. Priorities for studies in developing countries are different from Europe or America. Developing regions of the world must focus on the leading causes of death such as malaria, HIV and AIDS, dengue, and tuberculosis. Developmental validation studies should compare diseases in the same age group in both models developed and validated.

The method of outcome determination should be accurate to provide proper patient risk stratification and to support personal clinical decision making with the goal of improving patient outcomes and quality of care. Incorrect outcome assessment and measurement of predictors may inflate the predictive accuracy of the predictors and that of the final prognostic model. With respect to CHD definitions reported in the models reviewed, most models defined the incidence of coronary heart disease as fatal or nonfatal myocardial infarction, and over 80 different definitions for the disease outcomes were identified. In addition, most outcomes were not fully defined, and the International Classification of Diseases codes for CHD were described in only a few models. Different outcome definitions and measurement methods may lead to differences in study results and are a source of heterogeneity across studies, and thus risk of bias. CHD definition and outcomes must be readily accessible, precisely stated, and have a low measurement errors rate in order to be valuable. Overall, heterogeneity of definitions of CHD outcome identified in this review may affect the discrimination ability of

the models: models fail to discriminate between case and non-case subjects and thus may influence the predictive accuracy of the final prognostic model. A standard consistent definition of CHD and outcome will increase transparency in reporting the predicted outcomes and may improve the quality of research. Similar to outcome predictors, several candidate predictors were reported in the models identified. Most models had the same common predictors, such as age, sex, smoking, blood pressure, and total cholesterol levels. Many novel predictors, such as genetic risk scores, biomarkers, and others have been described in a few models. Most of the novel and newly developed models show good performance in predicting CHD, but the strategies of how the predictors were selected and which approaches were used are still questionable in most studies reviewed.

Regarding the models' performance, most of the models measured discrimination ability with less commonly used calibration and classification measures. Therefore, the performance of these models has not been fully examined, and discrimination alone can be insensitive and less useful in evaluating risk prediction of future events. It will be valuable if it is used for comparing the fit of predictive models using the calibration statistic and reclassification improvement, it has been previously reported that a prediction model will constantly require reporting discrimination and calibration. If the prediction model is to be utilized to make clinical decisions. Other performance indicators may be required in specialized situations. It is essential to assess the goodness of fit and to validate the model to ensure predictive performance. The use of good strategies for model selection in addition to adequate performance and goodness-of-fit measures is needed in developing accurate predictions. It is important to ensure that the model is well-calibrated if the prognostic value is close to the true value of disease outcomes. Model calibration was assessed via Hosmer-Lemeshow goodness of fit with other measures, such as adjusted R square, cross-validation, and Akaike's and the Bayesian information criterion for small numbers of predictors. If there are many predictors, then forward stepwise regression can be used.

Our review assessed prognostic modelling studies in community-based settings, aiming to summarize the available evidence about the optimal model in predicting CHD events in healthy populations and to explore whether the inclusion of the genetic risk score in conventional risk factors improves the ability of these models. MESA risk score with coronary artery calcification, which is described by McClelland et al. (2015) seems to be an optimal model for predicting CHD risk in the general population. This model is a conventional modelling study

including age, gender, CAC, ethnicity, DM, smoking, family history of heart attack, TC, HDLC, SBP, and treatment predictors. We would recommend this model be used in low-income countries. Genetic risk score in the model of Brautbar et al. (2012) might help in predicting CHD risk when integrated to Framingham, despite the fact that the improvement only occurred in one population and failed to persist in the other included populations.

These findings can help clinicians and decision makers improve the quality of interventions and improve the health of the population at risk. Future validation studies for genetic modelling studies are needed to ensure the quality and transparency of the developed model. Methodological assessment of genetic models is required. Most of the genetic risk scores incorporated into the conventional risk factors improved the discrimination and reclassification ability in the derivation models. Most genetic modelling studies were developed using only Caucasian populations; thus, the generalizability of the existing prognostic models is questionable. Genetic modelling studies might be used to target the prevention of CHD if the individual's genetic risk is comprehensively evaluated. An accurate assessment of an individual's risk is fundamental to future efforts in personalized medicine for the primary prevention and proper management of CHD.

Even though, genetic analysis is expensive for assessing SNPs associated to CHD/AMI risk, an individual's genetic risk is thoroughly examined by a comprehensive collection of genetic association data the result may be utilized to target primary prevention of CHD. But unfortunately, most of the genetic variation were not fully examined, nor validated in different populations other than Caucasian. Understanding an individual's genetic risk through genetic modeling studies could be beneficial in targeting interventions for preventing coronary heart disease at an early stage. The costs associated with genetic assessments can vary widely. Factors influencing costs include the type of genetic testing, the number of genes analyzed, and the technologies used. Generally, genetic testing has become more accessible and affordable over time, but it still may incur expenses. Implementing genetic assessments on a population level is a complex task. Challenges include cost, logistics, ethical considerations, and the need for a robust infrastructure to handle and interpret large-scale genetic data. Additionally, the practicality and ethical implications of population-wide genetic screening should be carefully considered. The use of genetic information in targeting primary prevention has potential public health relevance. Identifying individuals at higher genetic risk for CHD could allow for more targeted public health interventions, education, and resources. However, the ethical and social

implications of using genetic information on a large scale need careful consideration. The approach aligns with the principles of preventive medicine by emphasizing the importance of understanding and mitigating risk factors before the onset of disease.

Future validation studies should include genetic application in different geographic locations, and fully independent validation by independent investigators using alternative measurements of these risk factors in different population settings may improve the prognosis of the disease. Incomplete reporting of information in both conventional and genetic modelling studies was observed regarding the following methodologies: sampling technique, subject selection criteria, categories and blinding, genetic information, construction of the final models, classification measures, duration of follow-up, and missing values of the participants and technique used for handling this issue. Therefore, simply excluding the participants with missing values from the analysis reduces the effective sample size and may also lead to inaccurate estimates of the predictor outcome associations and the predictive performance of the final model. The performance of a predictive model is overestimated when simply determined on the sample of subjects that were used to construct the model, and statistical techniques such as shrinkage and bootstrapping are available to attempt to reduce over-optimism at the model-building stage. Comprehensive and valid information on conventional and genetic models is needed. Researchers should enhance the quality of their reports by describing and highlighting this important information.

To ensure the generalizability of the prognostic model and the ability of the model to predict CHD in populations with different characteristics, an external validation study is needed to evaluate the model's performance and to avoid overfitting in prognostic modelling studies. The shrinkage and penalization method should also be applied to reduce overfitting by readjusting the regression coefficients.

Our review reveals that only one conventional model was considered a good prognostic model for CHD in the general population and applicable for use in clinical practice. One genetic modelling study (Brautbar et al., 2012), was externally validated in three different populations and performed decision analysis but had limitations regarding classification improvement in the comparator model.

For conventional and genetic modelling studies, the identified previous models might be considered a good and optimal prognostic for CHD risk in the general population, and applicable for use in clinical practice. However, the model of Brautbar et al. (2012) was

externally validated in three different populations and performed decision analysis, but this model had limitations regarding the discrimination and classification improvement that occurred in the developmental group without significant improvement in the comparator groups.

There are several reasons why the performance of a prognostic model needs to be evaluated before its results can be used; most of the models fail to satisfy certain statistical notions of correctness, fail to be useful in a clinical setting, or have invalid prognostic information. Furthermore, the same model might fail according to one clinical criterion and pass according to another. There were two definitions of the validation prognostic modelling study. First, a statistically validated model passes all appropriate statistical checks, including the goodness of fit on the original data set and unbiased prognosis on a new data set. Second, a clinically validated model performs satisfactorily on a new data set according to context-dependent statistical criteria.

5.2 Discussion of CHD/AMI risk prediction among the Hungarian population

The Hungarian cross-sectional study's main finding indicated a significant difference between populations in terms of gender distribution, height, weight, education levels, economic activities, family member, SBP, HDL-C, and glucose level. Roma had lower educational levels, a higher proportion of unemployed subjects and a bigger family size than the general population. Despite the fact that the frequency of CHD, AMI, and stroke was lower and non-significant among the included participants, clinical risk variables such as DM, CKD, HTN-Med, HTC-Med were greater among Roma. Although Smoking and decreased HDL-C levels were higher among Roma, SBP and glucose levels were significantly higher in the Hungarian general population. These findings revealed that the Roma population is at high risk for cardiometabolic illnesses, and our findings were consistent with previously Roma research.

When we investigated the effect size of the 30 SNPs chosen for predicting CHD/AMI risk in Hungarian populations. The result show that, six SNPs, including rs2306374 (gene MRAS), rs9818870 (gene MRAS), rs12190287 (gene TCF21), rs10455872 (gene LPA), rs3184504 (gene SH2B3), and rs9982601 (gene KCNE2), showed strong association with CHD/AMI risk after Bonferroni correction in the Hungarian general population, and three SNPs, rs17609940

(gene ANKS1A), rs2259816 (gene HNF1A), and rs12936587 (gene RASD1), were shown to be significant associated with CHD/AMI risk among the Roma.

Our findings also indicated that the general population's mean weighted GRS and GRS were greater. Despite the fact that the vast majority of the chosen SNPs were not shown to be independently linked to CHD/AMI, it appears that the general population is more likely to acquire CHD/AMI. Before looking at how factors like allele frequency can be inherited beside the CRFs for CHD/AMI risk, we ran HWE to discover any deviations of markers that failed to be at HWE. One SNP (rs12413409) indicated that was not in HWE was eliminated, leaving 29 SNPs integrated with CRFs to determine the weighted and unweighted GRSs burden for CHD/AMI.

Based on prior research, age is an important non-modifiable risk factor for CHD/AMI in multivariable analytic prediction models. We used this risk factor to predict CHD/AMI in Hungarian populations. The fundamental reason for this is because CHD is a condition caused by a multitude of risk factors, the majority of which are thought to be age related. In the multivariable regression model, age was related with the development of CHD/AMI, particularly in the general population, this result was like some previous reports on the risk factors of CHD. Furthermore, the marginal plot analyses revealed that the risk of CHD/AMI is obviously low among the younger populations in general; however, both populations (general and Roma) were susceptible to premature onset of CHD/AMI, with the risk expected to begin earlier among Roma individuals.

In this investigation, there was no significant association between male subjects and CHD/AMI risk, this conclusion may be due to of the small fraction of Hungarian males who participated in this study. However, the marginal plot predictions revealed that the risk of CHD/AMI increased among male subjects of the Hungarian populations compared to female subjects. However, Roma individuals tend to develop CHD/AMI earlier than the Hungarian general population. According to previous studies, male sex is also an independent predictor of CHD and plays a role in increasing CHD risk, despite the fact that the mortality rate of CHD is remain greater among females than males.

Hyperlipidemia was a major risk factor for CHD/AMI in the general population, and hypertension showed a significant association with the trait only in the Roma population.

Despite widespread agreement on the importance of lowering blood cholesterol levels, particularly LDL-C, in lowering the risk of CHD, this predictor was not employed in derived mode. In several models, including Framingham, SCORE, and PROCAM, LDL-C, HDL-C, and TG were found to be underestimating CHD/AMI risk in Hungarian populations, possibly due to the low prevalence of CHD/AMI among included subjects, or due to the complex interaction with other predictors. There have been no previous studies that predict CHD/AMI using increased total cholesterol rather than LDL-C, our study contrasted with these previous studies. Other research indicated, hypertension, dyslipidemias, diabetes, obesity, and smoking account for more than 90% of the population's attributable risk of acute myocardial infarction. DM was found to have no association with CHD risk in all models, probably due to its low prevalence, or interaction with other variables. DM had a substantial correlation with CHD/AMI in both populations in a bivariate analysis. Our study was consistent with previous study that indicated that SCORE was less precise for estimating risk in patients with DM. Smoking is known to be the strongest contributor to CHD/AMI in all available models, such as the Framingham, SCORE, QRISK1, QRISK2, and PROCAM models, all of which have been developed and validated to predict CHD risk in general populations. Herein, smoking was found to be a nonsignificant predictor of an elevated risk of CHD in the Hungarian populations, which could be owing to selection bias or interaction as a confounding factor with other predictors. One reason for this could be the presence of other highly correlated predictor variables in the model that have a stronger association with CHD. For example, high blood pressure, high cholesterol levels were also well-established risk factors for CVD and may have a stronger association with the outcome variable in the multiple regression model. In this case, the effect of smoking may be overshadowed by the effects of these other variables, thus making smoking a nonsignificant factor.

Based on previous studies, the combination of the CRFs and genetic components might improve the model performance and predict AMI/CHD better than CRFs alone. Although GRS/wGRS were not significantly associated with CHD/AMI risk in study populations, the utility of genetic factors is still questionable in risk prediction. We examined the predictive ability of the combined models where CRFs and genetic risk scores were integrated. A basic model that included DM showed good discrimination improvement compared to the model without DM. In addition, the models with GRS integration showed the greatest discrimination

improvement in general. The highest improvement for the Hungarian general occurred when we added the GRS to DM however the greatest improvement in the Hungarian Roma occurred when we added the wGRS to the basic SCORE.

5.3 Conclusions

Although the GWAS has improved our understanding of CHD etiology by identifying and validating genetic variants associated with complex human traits, the use of genetic architecture for clinical assessment of CHD is still uncertain, as the majority of the SNPs' functions for CHD risk prediction are unknown or are poorly understood. McClelland et al. (2015) provided an optimal model for predicting CHD in the general population by incorporating CAC marker to the CRFs basic model, this model can predict CHD risk in different populations.

This study can help clinicians and decision makers to improve the quality of interventions using CAC predictor in CRFs for CHD/AMI risk prediction and can help in improving the health of the population at risk. Future genetic modeling validation studies are required to assure the quality and transparency of the created models. Genetic models must be evaluated methodologically. Because most genetic models for CHD include Caucasian populations, there is no generalizability of the available genetic prognostic models. If an individual's genetic risk is thoroughly examined by a comprehensive collection of genetic association data, including numerous studies using graphical displays and extensive textual material. Genetic modelling studies may be utilized to targeted primary prevention of CHD. An accurate assessment of an individual's risk is fundamental to future efforts in personalized medicine.

We attempted to develop a genetic scoring model with a careful selection of CHD/AMI-associated SNPs that could provide a somewhat valid estimate of the genetic load in our study populations; however, due to the small number of CHD/AMI patients in the study groups, the GRSs were not significantly associated with the trait in the regression models, but the predictive accuracy of the models with a genetic component was remarkable.

Aging, elevated total cholesterol, and hypertension may all interact to raise the chance of developing CHD/AMI. In the future, verifying the genetic risk prediction model with independent datasets or cohorts to examine its accuracy, dependability, and generalizability would aid in determining the model's resilience and external validity. It would also be useful to assess the clinical usability and possible impact of the genetic risk prediction model in

clinical settings. This could include conducting prospective studies or implementation trials to evaluate the efficacy of our approach in clinical practice, as well as its potential for directing individualized risk assessment, prevention, and treatment options. However, using genetic risk prediction models has the potential to enhance health outcomes by providing individualized risk estimates and assisting in early diagnosis and management. However, careful assessment and discussion of these models' limitations and problems is required for their effective adoption in clinical practice.

5.4 Strengths and Limitations of This Study

The systematic review's strengths are that the search was done in several databases, including Embase, PubMed, Cochrane, Web of science, and Scopus with a solely human filter. In addition to that one reviewer extracted the data, which was then thoroughly reviewed by another, and the individual study characteristics were provided. Quality assessment was performed in duplicate using CHARM, and GRIPS statements and because of the heterogeneity and the huge number of different predictors identified, this data were unsuitable for conducting meta-analysis which may limit the scope of this investigation.

A single point data collection was likely performed for the cross-sectional study in order to identify the risk factors of CHD/AMI, and understand the role of the modifiable and nonmodifiable risk factors in developing CHD/AMI, then described the features among the general and Roma populations using genetics and conventional risk factors. The weighted and unweighted GRSs were developed based on careful selection of 30 SNPs for CHD/AMI risk prediction; however, due to the small number of CHD/AMI patients in the study groups, the GRSs were not significantly associated with the trait in the regression models. Likely, the predictive accuracy of the models with a genetic component showed a remarkable AUC improvement. Age, and medication of both high total cholesterol and hypertension all increased the risk for CHDAMI, indicating a synergistic interaction between these predictors. Our findings show that SCORE risk prediction dose not estimated the actual risk of CHD/AMI as many of the risk factors like male sex, SBP, DBP, LDL-C, HDL-C, smoking, DM, and GRSs, some predictors were not incorporated. Other limitations of this study include the lack of prospective or retrospective follow-up, making it unsuitable for predicting CHD/AMI, the small sample size, selection, and recall bias, and small number of CHD/AMI events. In the future, verifying the genetic risk prediction model accuracy, reliability, and generalizability

with independent datasets and cohorts could aid in identifying the model's resilience and external validity.

CHAPTER SIX

Novelty

6.1 Our systematic review identified CAC as the most effective marker for predicting CHD risk. We also outlined the ideal model for investigating CHD prognostic modeling studies in general populations. This represents the initial comprehensive analysis of studies exploring genetic and traditional risk factors for CHD risk prediction, employing the CHARM and GRIPS statements.

6.2 In this study, the GRSs and the allele frequencies of the 30 SNPs associated with CHD/AMI in Hungarian populations were calculated and compared for the first time.

6.3 This study is contemporary and pioneering as no genetic risk prediction model for the Roma population has been developed, no studies combining genetic and traditional risk factor modeling have been undertaken, and no assessments of model performance have been conducted.

6.4 This study is the inaugural examination of CHD/AMI risk prediction to explicitly address the condition of premature onset in both general and Roma populations.

CHAPTER SEVEN

Summary

CHD is recognized as a major cause of illness and mortality, as well as disability, in central Europe, developing countries, and some developed countries. Despite this, the incidence and death rate of CVD has been reduced in several countries as a result of successful technologies developed by GWASs. Preventive intervention is also available through effective medication of lifestyle risk factors and comprehensive medication with statins. In many nations, successful legislation for a significant killer (smoking) has also been developed and implemented. Identifying high-risk individuals who may develop CHD risk remains challenging; a complete and precise assessment of CHD risk is required to identify those high-risk groupings. Obesity, diabetes, and financial hardship are just a few of the risk factors for coronary heart disease that are on the rise throughout countries, organizations, and communities. The bulk of these factors have been proven to work synergistically with other factors such as obesity, smoking, hyperlipidaemia, and hypertension to accelerate the atherosclerosis process. Early identification of persons at high risk of developing CHD is crucial for health promotion and prevention initiatives, taking into consideration both hereditary and environmental risk factors. This method of identification has the potential to reduce mortality and morbidity while increasing cost-effectiveness. Treatment can benefit from a well-structured strategy for identifying persons at moderate risk of CHD before symptoms develop. Framingham, SCORE, QRISK, and ASSIGN prediction models are crucial in assessing the risk of CHD among population groups. Our systematic review identified CAC as the most effective marker for predicting CHD risk. We also outlined the ideal model for investigating CHD prognostic modeling studies in general populations. This represents the initial comprehensive analysis of studies exploring genetic and traditional risk factors for CHD risk prediction, employing the CHARM and GRIPS statements. In this study, the GRSs (weighted and unweighted) and the allele frequencies of the 30 SNPs associated with CHD/AMI in Hungarian populations were calculated and compared for the first time. The majority of these models, including Hungary, overstated and underestimated the frequency of CHD in the general population. However, several models have been developed in an attempt to assess the risk of CHD using individual risk factors. We developed a novel model for CHD risk prediction across Hungarian

communities by hypothesizing that Hungarian Roma had greater genetic diversity and environmental risk factors for CHD than the overall population. According to our data, Hungarian general has a larger burden of GRSs than the Roma community, which may predispose them to CHD development. Hungarian Roma, on the other hand, are more vulnerable to environmental risk factors that may interact synergistically to accelerate CHD, such as low socioeconomic deprivation including educational gap, lower economic activities, and living in a large family member, having low HDL-C levels, being more smokers, and having a higher rate of diabetes, hypertension, chronic kidney disease, and stroke. When CRFs and GRSs were integrated in multivariable models, age, medication of elevated total cholesterol and hypertension were found to be strongly and independently associated with CHD/AMI, other predictors overstated CHD/AMI risk in Hungarian populations, probably due to the low prevalence of CHD/AMI in the sample. This study is contemporary and pioneering as no genetic risk prediction model for the Roma population has been developed, no studies combining genetic and traditional risk factor modeling have been undertaken, and no assessments of model performance have been conducted. This study is the inaugural examination of CHD/AMI risk prediction to explicitly address the condition of premature onset in both general and Roma populations. We urged the Hungarian Roma population to address their modifiable risk factors in order to protect themselves from this hazardous disease because it is asymptomatic, and the majority of CHD patients are unaware that they are at risk. Subjects with proven risk factors for premature CHD/AMI or familial hypercholesteraemic should be prioritized for prompt intervention and therapy.

CHAPTER EIGHT

Recommendations

8.1 Application of the optimal model for predicting CHD in the general population by incorporating CAC marker to the CRFs basic model, this model can predict CHD risk in different populations including Hungarians. Clinicians should use CAC biomarkers in addition to the lipid levels and comorbidities to predict CHDAMI risk in suspected patients.

8.2 The Hungarian Roma population should improve their modifiable risk factors by adopting a healthy diet in order to increase the lower HDL-C, lowering their cholesterol levels (LDL-C), controlling HTN, DM and stopping smoking. To avoid complications and CHDAMI development, hypertensive patients should consult their doctor on a regular schedule, physical activity can help in lowering the blood pressure, and increasing insulin sensitivity, and lowering the LDL-C.

8.3 Targeted screening to identify members of the Roma and general Hungarian populations who have a family history of familial hypercholesteremic disease or CHD/AMI in order to attempt lifestyle modification or medication as a preventive measure.

8.4 Validation of this models in a large group of Hungarian populations will help in confirming the results and provide more information about CHDAMI risk.

Activate health education programs for Roma communities to improve their knowledge about CHD/AMI risk.

CHAPTER NINE

Keywords

Coronary heart disease, Systematic review, Prognostic modelling studies, Genetic risk factors, Conventional risk factors, Developmental models, Validation models, Coronary artery calcification, SCORE, Hungarian Populations, Discrimination, Calibration.

CHAPTER ELEVEN

Acknowledgment

First and foremost, I would like to express my gratitude to Allah Almighty, the most merciful and compassionate in His endless compassion, who made it possible for me to fulfill this study despite the limitations of my family and humanity.

I would like to extend my heartfelt thanks to my research supervisor Dr. Fiatal Szilvia, for the continuous support of my Ph.D. study and research, and for letting me work on this project. I am very grateful to her.

I am deeply grateful to Prof. Janos Sándor, who gave me the golden opportunity to complete this research. I would also like to sincerely thank my school principal, Prof. Róza Ádány, Prof. Margit Balázs.

I would like to thank my colleagues at Debrecen University for helping me in editing the thesis. Finally, I would like to thank my family, brothers, and sisters for keeping me motivated and helping me whenever I called them. I am thankful to my husband and friends for their immense support and help during this project. Without their help completing this project would have been very difficult.

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/195/2023.PL
Subject: PhD Publication List

Candidate: Nayla Mohamed Gomaa Nasr
Doctoral School: Doctoral School of Health Sciences

List of publications related to the dissertation

1. **Nasr, N. M. G.**, Soltész, B., Sándor, J., Ádány, R., Fiatal, S.: Comparison of Genetic Susceptibility to Coronary Heart Disease in the Hungarian Populations: Risk Prediction Models for Coronary Heart Disease.
   Genes. 14, 1-16, 2023.
   DOI: http://dx.doi.org/10.3390/genes14051033
   IF: 3.5 (2022)
2. **Nasr, N. M. G.**, Soltész, B., Sándor, J., Ádány, R., Fiatal, S.: Prognostic Modelling Studies of Coronary Heart Disease: A Systematic Review of Conventional and Genetic Risk Factor Studies.
   JCDD. 9 (9), 1-21, 2022.
   DOI: http://dx.doi.org/10.3390/jcdd9090295
   IF: 2.4

List of other publications

3. Dahlia, D., Artanti, K. D., Hargono, A., Martini, S., **Nasr, N. M. G.**, Li, C. Y.: Death risk among COVID-19 patients with diabetes mellitus.
   J Public Health Afr. 13 (s2), 1-5, 2022.
   DOI: http://dx.doi.org/10.4081/jphia.2022.2399
   IF: 0.8

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *

Total IF of journals (all publications): 6,7
Total IF of journals (publications related to the dissertation): 5,9

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.

10 July, 2023