MODELING OF CARDIOVASCULAR DISEASES (CVDS) AND DEVELOPMENT OF PREDICTIVE HEART RISK SCORE

MIRZA RIZWAN SAJID

DOCTOR OF PHILOSOPHY

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : MIRZA RIZWAN SAJID

Date of Birth : 15 FEBRUARY 1983

Title : MODELING OF CARDIOVASCULAR DISEASES (CVDs) AND DEVELOPMENT OF PREDICTIVE HEART RISK SCORE

Academic Session : SEMESTER 2 2020/2021

I declare that this thesis is classified as:

- ☐ CONFIDENTIAL (Contains confidential information under the Official Secret Act 1997)*
- ☐ RESTRICTED (Contains restricted information as specified by the organization where research was done)*
- ☑ OPEN ACCESS I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)	(Supervisor's Signature)
CP5573202	DR. NORYANTI MUHAMMAD
New IC/Passport Number	Name of Supervisor
Date: 5 JULY 2021	Date: 5 JULY 2021

NOTE: * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

We hereby declare that we have checked this thesis and, in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy.

(Supervisor's Signature)

Full Name : DR. NORYANTI MUHAMMAD

Position : SENIOR LECTURER

Date : 5 JULY 2021

(Co-supervisor's Signature)

Full Name : DR. ROSLINAZAIRIMAH BINTI ZAKARIA

Position : ASSOCIATE PROFESSOR

Date : 5 JULY 2021

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

(Student's Signature)

Full Name : MIRZA RIZWAN SAJID

ID Number : PSS17002

Date : 5 JULY 2021

MODELING OF CARDIOVASCULAR DISEASES (CVDs) AND DEVELOPMENT OF PREDICTIVE HEART RISK SCORE

MIRZA RIZWAN SAJID

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Doctor of Philosophy

Centre for Mathematical Sciences
UNIVERSITI MALAYSIA PAHANG

JULY 2021

ACKNOWLEDGEMENTS

Praise to Allah the Almighty for bestowing us with the best strengths and qualities to be used for the benefit of mankind. As a humble servant of Allah SWT, I am trying to express my gratitude in the form of little knowledge. I am a follower of His last Messenger Muhammad (PBUH), an eternal source of knowledge and guidance towards humanity's truth and is the gold standard for all intelligence.

Firstly, I am immensely grateful to my main supervisor, Dr. Noryanti Muhammad, and co-supervisor Associate Professor Dr. Roslinazairimah Zakaria for their academic guidance and unlimited support throughout my humble struggle to complete this thesis. Dr. Noryanti's patience, advice, thought-provoking ideas, objective-oriented approaches, and unconditional support have been invaluable to me. I am also thankful to Dr. Ahmad Shahbaz from Punjab Institute of Cardiology for his valuable help in data collection, especially for the standardized ethical considerations.

I am highly indebted to my dearest parents, wife, and my children (Muhammad Shahaan Alam and Emaan Fatima) for their moral support and patience throughout this work. It was really tough for them to manage their things alone, which were supposed to be my duties. Special thanks to my sisters, their husbands, and brother, whose prayers have always been sources of strength and inspiration for me.

I also owe my thankfulness to all my teachers, friends and colleagues, especially Dr. Waris Ali Khan, Dr. Nadeem Shafique Butt, Dr. Asif Hanif, Asim Butt, Arshad Ali Khan, Dr. Syed Ahmad Chan Bukhari, Jawad Asif, Dr. Waqas Sami, Dr. Fayyaz Ahmad, Mehfooz Ullah Dar, Muhammad Bilal, Dr. Asad Ullah Khan and Muhammad Ilyas for their encouragement, appreciation and support during my research.

My final thanks to the University of Gujrat (UoG) for providing me the opportunity to pursue my studies at Universiti Malaysia Pahang (UMP). I am also thankful to the Centre for Mathematical Sciences, College of Computing & Applied Sciences for offering me an excellent academic and conducive environment, which had enabled me to study smoothly.

ABSTRAK

Penyakit kardiovaskular (CVDs) adalah penyebab utama kematian dengan 31% kematian global. Tujuan kajian ini adalah untuk membangunkan model lintasan yang sah secara statistik yang mempertimbangkan kemungkinan lintasan bukan linear dan ciri binari endogenos, dan pengantara keduanya bagi status CVDs. Kajian ini menumpukan pembangunan pelbagai bentuk model ramalan risiko tempatan dan penggunaan skor risiko jantung yang ringkas dengan menggunakan ciri bukan makmal dan algoritma pembelajaran mesin (*machine learning*) (*ML*). Walau bagaimanapun, penukaran bentuk algoritma *ML* yang kompleks kepada model statistik yang ringkas menjadi perhatian utama didalam kajian ini. Kajian kawalan kes yang sesuai dengan jantina dilakukan di Institut Kardiologi Punjab, Pakistan, di mana 460 individu sebagai sampel dipilih melalui persampelan bersistematik. Kaedah *warp-partial least square* digunakan untuk menggangar pelbagai lapisan model lintasan yang dihipotesiskan. Model ini menganggarkan pekali *warp* menggunakan keseluruhan aliran linear yang terdapat dalam segmen linear daripada hubungan bukan linear. Model yang dibangunkan ini merupakan laluan pintasan yang novel di mana ciri demografi dan sosioekonomi menjadi pemacu utama bagi ciri tingkah laku, yang membawa kepada status CVD secara langsung dan tidak langsung melalui metabolic sindrom. Dalam membangunkan model ramalan risiko, dua algoritma *ML*, iaitu sokongan linear mesin vektor (*linear support vector machine*) dan *artificial neural network* mengatasi model konvensional iaitu analisis regresi logistik (*Logistic Regression Analysis*) (LRA). Prestasi model yang dibangunkan ini dinilai melalui pelbagai matriks yang ditetapkan dengan menggunakan pengesahan bersilang 10-kali ganda (*10-fold cross-validation*). Kemudian, satu novel metodologi dibangunkan dan digunakan untuk mengira skor risiko jantung yang ringkas berdasarkan ciri bukan makmal yang dipanggil *Non-Laboratory based Heart Risk Score* (*NLHRS*). Metodologi ini menyusun dan mengumpul algoritma *ML* yang terbaik dan digunakan sebagai asas untuk mengira indeks pemberat ciri relatif (*relative feature weights*). Indeks pemberat ini disebut sebagai *NLHRS*, yang selanjutnya digunakan sebagai kovariat di model LRA ringkas untuk menganggarkan kemungkinan CVD berlaku. Perubahan yang berlaku iaitu dari model algoritma yang bersifat kotak-hitam komplek kepada model statistik ringkas yang menghasilkan model yang tidak memerlukan sistem berautomatik untuk pelaksanaannya. *NLHRS* yang berasaskan algoritma *ML* dan model yang bersangkutan yang dibangunkan ini menunjukkan prestasi yang lebih baik daripada model sediada yang berdasarkan skor risiko semi-kuantitatif dari segi penilaian diskriminasi dan penentukuran. Akhirnya, keupayaan model ramalan *NLHRS* juga diuji dan disesuaikan mengikut strata penduduk. Kajian ini menyimpulkan beberapa perkara. Pertama, penggunaan pendekatan kaedah yang fleksibel dalam anggaran dapat memodelkan ciri binari bagi status CVD dan lintasan tidak linear didalam lintasan model yang kompleks. Model lintasan CVD yang dianggarkan dapat digunakan sebagai strategi penangguhan penyakit dalam pengaturan klinikal. Kedua, model algoritma *ML* menawarkan model ramalan risiko yang lebih baik dan konsisten berbanding model yang berasaskan LRA. *NLHRS* dan model yang berkaitan dengannya yang dibangunkan merupakan hasil metodologi novel yang memberikan bentuk skor risiko yang sah dan ringkas dan dapat digunakan tanpa sistem berautomatik.

ABSTRACT

Cardiovascular diseases (CVDs) are the leading cause of death, with 31% of global mortality. The purpose of this study is two folds such as the development of a statistically valid path model which considered the possible non-linear paths, mediators, and binary endogenous feature of CVDs status. Further, it focuses on the development of various forms of local risk prediction models and simple heart risk scores using non-laboratory features and machine learning (*ML*) algorithms. However, the conversion of a complex form of *ML* algorithms into a simple statistical model is the prime concern. A gender-matched case-control study was conducted in Punjab Institute of Cardiology, Pakistan, in which a sample of 460 individuals was selected through systematic sampling. The warp-partial least square method was utilized to estimate the multi-layer hypothesized path model. This model estimated warped coefficients using the overall linear trend found in linear segments of non-linear relationships. This model found novel pathways in which demographic and socioeconomic features are the main drivers of behavioral features, leading to CVDs status directly and indirectly through metabolic syndrome. In developing risk prediction models, two *ML* algorithms, linear support vector machine and artificial neural network outperformed the existing conventional logistic regression analysis (LRA) model. The performance of the models was assessed through various established matrices using 10-fold cross-validation. A novel methodology was used to compute simple heart risk scores called non-laboratory based heart risk score (*NLHRS*). The methodology is proposed as stacking ensemble *ML* and the best *ML* algorithms are used as a base learner to compute relative feature weights. The index of these weights is referred to as *NLHRS*, which was further used as a covariate in the simple LRA model to estimate the likelihood of CVDs. This conversion from a complex black-box nature of *ML* algorithms into simple statistical models yielded such models, which do not require automated systems for their implementation. *ML*-based *NLHRS* and their associated models outperformed the existing semi-quantitative risk score-based model in terms of discrimination and calibration assessments. Finally, the predictive capability of valid *NLHRS* models has also been tested and adjusted for different strata of the population. Firstly, the study concludes that the adoptions of the flexible approach in estimation can model the binary feature of CVDs and non-linear paths in the complex path models. The estimated CVDs path model can be implemented as a disease delay strategy in clinical settings. Secondly, the *ML* models offer better and consistent risk prediction models as compared to LRA-based model. The *NLHRS* and their associated models which are the outputs of novel methodology provide valid and simple forms of risk scores and can be used without automated systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

A	Attributes for classification
B	Matrix of regression coefficients of y_i variables on y_i variables where $i \neq i$
C	Set of classes in the data set
$\mathbf{C}$	Correlation Matrix
c	Class
c_t	Cost function in kernels
f	Map from n-dimensional to m-dimensional
F	Weight vector
H	Hosmer and Lemeshow Test
H_0	Null Hypothesis
H_1	Alternative Hypothesis
$H(S)$	Entropy of data set S
$H(t)$	Entropy of subset t
I	Identity matrix
k_r	Kernel
$k-1$	A method of cross-validation
log_2	The logarithm of base 2
M	Margin
N	Size of population
N_i	Number of items
n	Sample size
P	Probability for favorable events
p	Proportion
Q	Bias
R^2	Coefficient of determination

S	Data set
S^2	Sample Variance
SE	Standard error
s_p	Split point in decision tree
T	Subsets creating from splitting set S
t	Target class
v_i	Training example
W_O	Moderator
x	First input feature
x'	Second input feature
x_j	Independent variable
y_i	Dependent variable
Z	Z-statistic
z_i	Logit function or logit index
σ	Free parameter in kernel function
Γ	Matrix of regression coefficients of y_i on x_i
ϕ_{ij}	Covariance between exogenous variables
ϵ	Error term in regression analysis
ζ_i	Error vector
β_0	Intercept in regression analysis
φ_l	Regression coefficient of interaction term
α	Level of significance
α_j	Regression coefficient of mediator
ρ	Reliability coefficient
ρ_O	Regression coefficient of moderator
γ_{11}	Regression coefficients of x_j on y_i
β_j	Regression coefficient
β_{ii}	Regression coefficient of y_i with another y_i where $i \neq i$
λ	Eigen value

LIST OF ABBREVIATIONS

AFVIF	Average Full Collinearity Variance Inflation Factor
AMI	Acute Myocardial Infarction
AMOS	Analysis of Moments Structures
ANN	Artificial Neural Network
ANN-RS	Artificial Neural Network based Risk Score
ANOVA	Analysis of Variance
AO	Abdominal Obesity
APC	Average Path Coefficient
ARS	Average R-Square
AUC	Area under the Curve
BMI	Body Mass Index
BS	Brier Score
CB-SEM	Covariance based Structure Equation Modeling
CE	Combined Exposure
CART	Classification Analysis and Regression Trees
CI-TC	Corrected Item Total Correlation
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CHD	Coronary Heart Disease
CI	Confidence Interval
CVDs	Cardiovascular Diseases
CVM	Cardiovascular Mortality
DM	Diabetes Mellitus
DT	Decision Trees
DV	Dependent Variable
FFQ	Food Frequency Questionnaire

FN	False Negative
FH	Family History
FP	False Positive
GFI's	Goodness of Fit Indices
GI	Gini Index
GoF	Goodness of Fit
HDL	High-density Lipoprotein
HICs	High-income Countries
HTN	Hypertension
ICT	Islamabad Capital Territory (Capital city of Pakistan)
ICR	Internal Consistency Reliability
IDF	International Diabetes Federation
ID3	Iterative Dichotomoizer
IV	Independent Variable
KMO	Kaiser-meyer-olkin Test
KS	Kolmogorove Smirnove
LMICs	Low-middle-income countries
LL	Log liklihood
LRA	Logistic Regression Analysis
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
MS	Metabolic Syndrome
MSA	Metabolic Syndrome Abnormalities
NDH	Negative Dietary Habits
NCDs	Non-communicable Diseases
NLHRS	Non-laboratory based Heart Risk Score
OLS	Ordinary Least Square
OR	Odds Ratio

PA	Physical Activity
PI	Prognostic Index
PI_A	Prognostic Index based on ANN-RS
PI_S	Prognostic Index based on SVM-RS
PI_R	Prognostic Index based on RFS
PLS	Partial Least Square
PLS-SEM	Partial Least Square based Structure Equation Modeling
PPLUA	Percentage Population Living in Urban Areas
RBF	Radial Basis Function
RFS	Risk Factors based Risk Score
RMSE	Root Mean Square Error
RMSEA	Root Mean Square Error Approximation
ROC	Receiver Operating Characteristics
RR	Risk Ratio
RT	Regression Trees
SDH	Social Determinants of Health
SEM	Structure Equation Modeling
SE	Socio-economic
SFW	Subjective financial Well-being
ST	Sleep Satisfaction
SPSS	Statistical Packages for Social Sciences
SS	Self-reported Subjective Stress
SVM	Support Vector Machine
SVM-RS	Support Vector Machine based Risk Score
TN	True Negative
TP	True Positive
WC	Waist Circumference
WHO	World Health Organization
WHR	Waist to Hip Ratio

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Cardiovascular diseases (CVDs) are a group of diseases, usually referred to as conditions that involve narrowed or blocked blood vessels leading to heart attack and other related problems. Recent estimates of the World Health Organization (WHO) showed that 31% of cause-specific mortality occurs due to this disease group (Organization et al., 2018a). Global cardiovascular mortality (CVM) has established an exponential trend where 40.8% increase was observed from 1990 to 2013 (Roth et al., 2015). This rise has made it the most important and the largest cause of noncommunicable diseases (NCDs) at over 50% (McAloon et al., 2016). However, the nature of disease burden is diverse in different regions of the world. It usually occurs in low-middle-income countries (LMICs), contributing to 80% of the annual deaths (Organization et al., 2019). These statistics of CVDs related deaths reflect the enormity of disease, which is continuously growing. Therefore, it has become a public health challenge, especially for LMICs and needs to be duly addressed.

Global burden of disease (GBD) reported that age-standardized CVDs mortality in high-income countries (HICs) is decreased by 21% in the last two decades (Lozano et al., 2012). The adoption of population and individual-based preventive strategies recommended by WHO is the main reason for this substantial decline in HICs (Bonita et al., 2013). In population-based strategies, the focus is on developing such public health policies, which transform the distribution of primary risk factors of CVDs in the population. However, individual-based strategies tend to focus on the early assessment of the likelihood of CVDs events, identifying high-risk individuals and their management through modifiable risk factors. Besides, it is more suitable in clinical settings as they