Hindawi
Advances in Preventive Medicine
Volume 2019, Article ID 8392348, 11 pages
https://doi.org/10.1155/2019/8392348

Research Article

A Cox-Based Risk Prediction Model for Early Detection of Cardiovascular Disease: Identification of Key Risk Factors for the Development of a 10-Year CVD Risk Prediction

Xiaona Jia, Mirza Mansoor Baig, Farhaan Mirza, and Hamid GholamHosseini

School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand

Correspondence should be addressed to Mirza Mansoor Baig; mirzamansoor01@gmail.com

Received 7 November 2018; Accepted 26 March 2019; Published 9 April 2019

Academic Editor: Gerardo E. Guillén Nieto

Background and Objective. Current cardiovascular disease (CVD) risk models are typically based on traditional laboratory-based predictors. The objective of this research was to identify key risk factors that affect the CVD risk prediction and to develop a 10-year CVD risk prediction model using the identified risk factors. Methods. A Cox proportional hazard regression method was applied to generate the proposed risk model. We used the dataset from Framingham Original Cohort of 5079 men and women aged 30-62 years, who had no overt symptoms of CVD at the baseline; among the selected cohort 3189 had a CVD event. Results. A 10-year CVD risk model based on multiple risk factors (such as age, sex, body mass index (BMI), hypertension, systolic blood pressure (SBP), cigarettes per day, pulse rate, and diabetes) was developed in which heart rate was identified as one of the novel risk factors. The proposed model achieved a good discrimination and calibration ability with C-index (receiver operating characteristic (ROC)) being 0.71 in the validation dataset. We validated the model via statistical and empirical validation. Conclusion. The proposed CVD risk prediction model is based on standard risk factors, which could help reduce the cost and time required for conducting the clinical/laboratory tests. Healthcare providers, clinicians, and patients can use this tool to see the 10-year risk of CVD for an individual. Heart rate was incorporated as a novel predictor, which extends the predictive ability of the past existing risk equations.

1. Introduction

Cardiovascular disease (CVD) describes various conditions that affect the functioning of heart/cardiovascular [1]. Due to the high rate of disease morbidity, CVD has become the leading cause of mortality around the world [2–4]. In New Zealand, statistics on CVD mortality in 2017 suggests that the percentage of deaths caused by CVD is 33% [4].

Majority of cardiovascular-related deaths are premature and preventable and can be improved by effective health management by employing effective diet plans, lifestyle interventions, and drug intervention [5]. To prevent CVD, a useful approach is to assess CVD risk regularly and then introduce new lifestyle adjustments or clinical treatments accordingly.

In the past decades, a great deal of research has been done on the CVD risk estimation such as the Framingham risk scores from the Framingham Heart Study (FHS) [6, 7], the QRISK equations [8], the Europe SCORE risk equations [9], the ASSIGN scores from the Scottish Heart Health Extended Cohort (SHHEC) [10], the Prospective Cardiovascular Master (PROCAM) equations [11], and the CUORE Cohort Study formulas [12]. These CVD risk prediction models have proved their effectiveness in the health and disease management for clinicians and individuals [13–15]. The new PREDICT CVD risk assessment equation developed for primary health care among the population in New Zealand has been integrated to the electronic health records (EHRs) and a web-based software called PREDICT has been developed to support general practices manage the CVD risk in primary care [13]. The PREDICT has got 400,728 patients assessed with the CVD risk and is becoming a useful tool for decision support and health management for general practitioners.

However, challenges and issues regarding the development of CVD risk estimation models still exist. CVD risk

Table 1: CVD event distribution in male and female.

	Count.	CVD Events	Age Range
Male	2294	1560	30 - 74
Female	2785	1629	30 - 74
Total	5079	3189	30 - 74

models [16–18] are based on single risk factor which cannot realize the influence of multiple factors simultaneously. Risk models [6, 8, 19] using statistical regression methods [20–22] prefer to use classic risk factors such as age, smoking, diabetes, sex, high blood pressure, and total cholesterol to estimate the risk score. Studies [18, 19, 23–27] applying data mining or machine learning techniques for the CVD risk estimations cannot provide an absolute risk estimation, although some of these models [18, 26] tried to incorporate novel predictors in the risk models. This research aims to identify the novel risk factors for CVD detection by conventional predictors and then enhance the risk estimation by developing a multiple-variable-based risk prediction model that targets the 5-year and 10-year CVD events.

2. Methods

2.1. Study Population. The study population selected from the Framingham Original Cohort study dataset [28, 29]. We obtained the ethics approval from NHLBI [30] and the Auckland University of Technology Ethics Committee (AUTEC) (Ref: 17/385 Early Detection and Self-Management of Cardiovascular Disease Using Artificial Intelligence-Based Model). The data from this cohort study includes a total of 5079 men and women aged 30-74 years free of CVD at the baseline, of them 3189 had CVD events eventually. Details of the CVD events distribution in male and female among the study population are summarized in Table 1.

2.2. Data Extraction. There are 32 exams in the Framingham Original Cohort study dataset, as shown in Appendix A. Data frame collected in the first exam "Exam1" was chosen to develop the CVD prediction model because it has the maximum number of samples 5209 subjects. Data from 130 subjects were removed because of the ethics protection. The other five exams are ranging from 8 to 12, marked with italic font (as shown in Table 7 of Appendix A) and will be used for the validation for the fitted model. Data of candidate risk factors (listed in Table 2) for creating the risk model was extracted.

2.3. Statistical Analysis. Cox proportional hazard regression analysis [22] was selected for developing the proposed risk model (one of the most accurate method belonging to the semiparametric statistical method). This research aims to develop a prediction model using multiple parameters to estimate the probability of developing CVD for an individual. There are mainly three statistical approaches in *survival analysis*, i.e., nonparametric, semiparametric, and parametric [31]. The nonparametric approaches can only perform univariate analysis with single predictor and therefore are not suitable for the study of continuous variables [22, 32]. Both parametric and semiparametric approaches can perform multiple parameter analysis. They assume that the predictors and the log hazard rate have a linear relationship between [33]. However, the Cox proportional hazard model has an advantage that only the rank orderings of the failure and censoring times are used to estimate and test the regression coefficients [22]. The Cox model is more efficient even though the assumption of the parametric models is met. When the assumptions are not met, the Cox regression analysis can still be used efficiently with an extended Cox regression from [34], but a parametric model such as Weibull survival distribution would be a null model.

Statistical analyses were performed in R Studio platform [35]. Missing values for candidate risk factors listed in Table 2 were imputed using *Multiple Imputation* [36]. Continuous and categorical variables were transformed and imputed using algorithms modified from Maximum Generalized Variance (MGV) in the SAS PRINQUAL procedure [37]. R function *transcan* inside the "Hmisc" package was used [35].

For candidate predictors listed in Table 2, two steps of variables selection from the list were performed. The first step was conducted in a "Forward Selection" manner [38]; i.e., the univariate Cox analysis was applied to all candidate variables. Insignificant predictors were filtered out based on a significance level p value >0.05. In the second step, all selected variables from the univariate analysis were entered into the multivariate Cox regression analysis to see how the risk factors jointly impact the incidence rate for CVD. Risk factors with a p value less than 0.05 will be finally decided.

In the validation stage, two approaches were undertaken to assess the predictive ability of our fitted model, statistical validation, and empirical validation. The statistical validation was performed with respect to both discrimination and calibration. The empirical validation was defined as an empirical comparison with a general CVD risk prediction model (the Framingham office-based risk equation [6]) in a horizontal and longitudinal perspective. The horizontal comparison was conducted by comparing with the Framingham prognostic model using data collected from multiple samples at the same time point. The longitudinal comparison was conducted by comparing with the Framingham prognostic model using data collected from specific examples at different time-points (fixed time intervals follow-up) and seeing the risk trend for an individual over time.

3. Results

3.1. Derivation of a 10-Year Risk Score for CVD. Risk factors included in the risk model are age, sex, body mass index (BMI), hypertension, systolic blood pressure (SBP), cigarettes per day, pulse rate, the status of diabetes. Characteristics of risk factors were listed in Table 3. Statistics of "Min.", "1st Qu.", "Median", "Mean", "3rd Qu.", and "Max." of these risk factors are summarized.

The regression coefficients, hazard ratios, and their corresponding upper and lower 95% confidence intervals (CI) were estimated, as presented in Table 4. Values of the baseline hazard rate where the time point is ten years were estimated

Table 2: Description of candidate predictors.

ORDERS	PREDICTORS	UNITS	TYPES
1	AGE	YEARS	CONTINUOUS
2	SEX	0001 MALE 0002 FEMALE	CATEGORICAL
3	BMI	KG/M2	CONTINUOUS
4	HYPERTENSION	0000 NEGATIVE 0001 TRANSIENT 0002 PERMANENT 0003 TYPE UNKNOWN 0008 DOUBTFUL	CATEGORICAL
5	HISTORY OF NERVOUS HEART	0000 NO 0001 YES, DEFINITE	CATEGORICAL
6	HISTORY OF PERICARDITIS	0000 NO 0001 YES, DEFINITE	CATEGORICAL
7	HISTORY OF OTHER CVD	0000 NO 0001 YES, DEFINITE	CATEGORICAL
8	PREMATURE BEATS	0000 NO 0001 YES, DEFINITE 0002 YES, DOUBTFUL	CATEGORICAL
9	HISTORY OF ATRI-OVENTRICULAR BLOCK	0000 NO 0001 YES, DEFINITE 0002 YES, DOUBTFUL	CATEGORICAL
10	HISTORY OF RHEUMATIC FEVER	0000 NONE 0001 YES 0008 DOUBTFUL	CATEGORICAL
11	HISTORY OF ALLERGY OR ASTHMA	0000 NEGATIVE 0001 ALLERGY, ALONE 0002 BRONCHIAL ASTHMA, ALONE, 0003 ALLERGY AND ASTHMA, TOGETHER	CATEGORICAL
12	HISTORY OF THYROID DISEASE	0000 NEGATIVE 0001 HYPERTHYROID ONLY 0002 HYPOTHYROID ONLY	CATEGORICAL
13	HISTORY OF SUBACUTE ENDOCARDITIS	0000 NO 0001 YES	CATEGORICAL
14	BLOOD PRESSURE SYSTOLIC	MM HG	CONTINUOUS
15	BLOOD PRESSURE DIASTOLIC	MM HG	CONTINUOUS
16	CIGARETTES PER DAY	LAPSE, FORM 8/50	CONTINUOUS
17	CIGARS PER DAY	LAPSE, FORM 8/50	CONTINUOUS
18	PIPERS PER DAY	LAPSE, FORM 8/50	CONTINUOUS
19	PULSE RATE	PER MINUTE	CONTINUOUS
20	DIABETES	0000 NO 0001 YES, DEFINITE	CATEGORICAL

as well, shown in Table 5. The 10-year baseline hazard rate is 0.1023354 at mean values of all covariates, 0.001863652 at all covariates equal to zero. Corresponding, the survival probability (exp(*basehaz*)) is 0.9027267 at mean values and 0.9981381 at all covariates equal to zero.

The Cox model has an exponential form (see Equation (1)), where t represents the time that the event occurs; $\lambda(t)$ is the hazard function for a subject at time t, determined by a set of m covariates $(X_1, X_2, \ldots, X_k)$; $\beta_1, \beta_2, \ldots \beta_k$ are the regression coefficients that measure the effect size of covariates; exp is the exponential function (exp(X) = ex); $\lambda_0(t)$ is the baseline hazard rate, an arbitrary (unknown) function, corresponding to the value of the hazard when all X_i equal zero.

$$\lambda(t) = \lambda_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right) \quad (1)$$

Table 3: Summary statistics for risk factors used in risk model.

Predictors	Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
AGE	Age	28	37	44	44.15	51	74
SEX	Sex	1	1	2	1.548	2	2
BMI	Bmi	14.12	22.66	25.17	25.61	27.92	56.68
HYPERTENSION	Hyp	0	0	0	0.147	0	1
BLOOD PRESSURE SYSTOLIC	Bps	84	122	136	138.6	150	270
CIGARETTES PER DAY	Cgrpd	0	5	20	16.26	20	60
PULSE RATE	Pr	37	67	75	75.61	83	170
DIABETES	Dia	0	0	0	0.0197	0	1

Table 4: Regression coefficients and hazard ratios in risk model.

Predictors	Variables	coef∗	Hazard Ratio	lower .95	upper .95
AGE	log of age	2.083643	8.033686	6.4082	10.0716
SEX	sex	-0.469719	0.625178	0.5787	0.6754
BMI	log of bmi	0.608864	1.838342	1.4368	2.3521
HYPERTENSION	hyp	0.241461	1.273108	1.1342	1.429
BLOOD PRESSURE SYSTOLIC	log of bps	1.682571	5.37937	3.7938	7.6277
CIGARETTES PER DAY	cgrpd	0.009669	1.009716	1.0065	1.013
PULSE RATE	log of pr	-0.30209	0.739271	0.5879	0.9297
DIABETES	dia	1.087501	2.96685	2.3244	3.7869

∗ Estimated regression coefficient.

Table 5: Baseline hazard and survival at 10 years.

	Covariates at mean value	Covariates equal to zero
Baseline hazard estimate	0.1023354	0.001863652
Baseline survival estimate	0.9027267	0.9981381

So, the Cox model can be written as a survival function:

$$S(t) = \left[S_0(t)\right]^{\exp(\sum_{i=1}^{k} \beta_i X_i)} \quad (2)$$

A general formula for computing risk estimates has the following form:

$$\widehat{H(t)} = 1 - \left[S_0(t)\right]^{\exp(\sum_{i=1}^{k} \beta_i X_i - \sum_{i=1}^{k} \beta_i \overline{X}_i)} \quad (3)$$

where H(t) is the CVD risk estimated for an individual; S0(t) is baseline survival rate at follow-up time t, where t = 10 years (see Table 5), βi is the regression coefficient (see Table 4), X_i is the value of the i_{th} risk factor (if is continuous it is the log-transformed value), $\overline{X}_i$is the corresponding mean, and k denotes the number of risk factors. The CVD risk function could be derived from (3), using regression coefficients from Table 4 and the baseline hazard rates from Table 5; hence, we computed the probability of developing any type of CVD for an individual. A case of computing the absolute risk score in 10 years was demonstrated in Appendix C.

3.2. Nomograms. A nomogram is a two-dimensional diagram to represent a mathematical function involving several predictors [39]. It is a simple graphical illustration to approximately predict a particular event based on conventional statistical regression methods such as Cox proportional hazards model for survival analysis [40]. A nomogram is accomplishing the estimation of individual survivals in 10 years and the median survival time by years was depicted in Figure 1.

In Figure 1, each predictor has a set of n scales, and there is a mapping between each scale and the "Points" scale. The bottoms are the corresponding 10-year survival estimates, and the median survival time (years). By accumulating the total points corresponding to the specific configuration of covariates for a patient, a clinician can then manually obtain the predicted value of the event for that patient.

3.3. Validation. The validation of the proposed predictive risk model was performed using traditional statistics. C-index (also called receiver operating characteristic (ROC) area) [41] was used to assess the goodness of the risk model based on a bootstrap internal resampling validation. From the statistical validation analysis, we got a C-index (area under the receiver operator curve [AUROC]) of 0.71 indicating moderately good discrimination.

Then, we performed an empirical validation by comparing our risk model with the Framingham Heart Study model in an external dataset horizontally and longitudinally over time. In the horizontal validation process, there were 2786

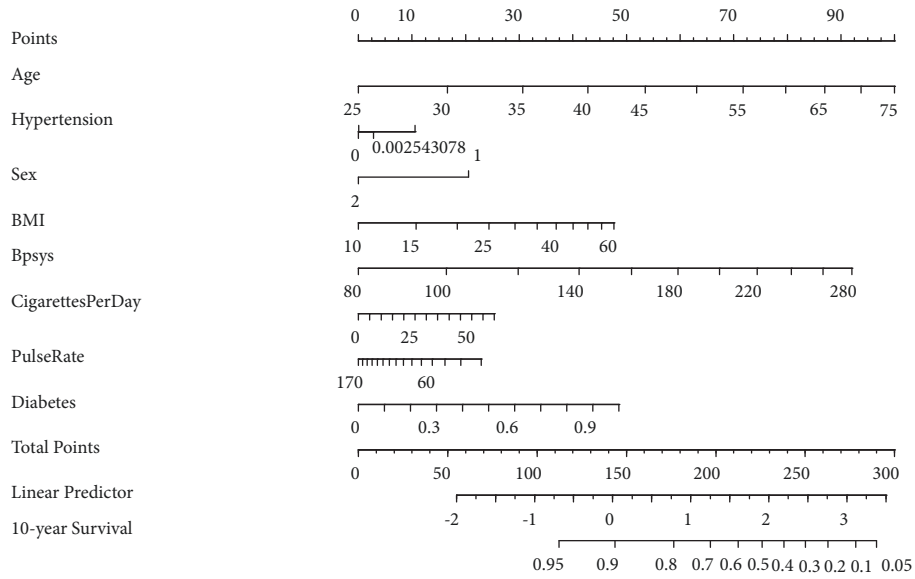

Figure 1: Nomogram for predicting overall survival in 10 years.

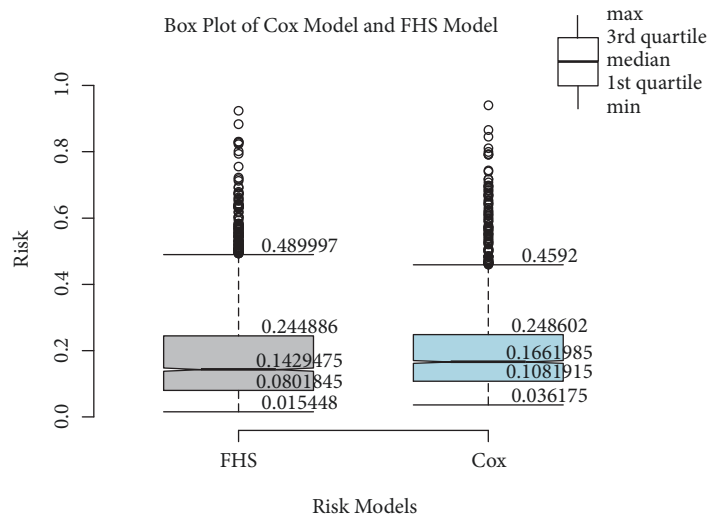

Figure 2: Horizontal comparison between Cox model and FHS model.

samples in the external dataset, and 1693 samples have got a CVD event. Risk scores using the FHS model and the proposed risk model were computed separately. Statistics of *min (lower whisker)*, *1st quartile (the lower hinge)*, *median*, *3rd quartile (the upper hinge)*, and *max (the extreme of the upper whisker)* of estimated risks for all samples are depicted in Figure 2. This box-whisker graph in Figure 2 shows that the risks assessed by our Cox model are higher than the risk calculated by the Framingham model, but the error for five statistics (min, 1st Qu, median, mean, 3rd Qu., max) is within 0.02. For example, the median values of the FHS model and the Cox model are 0.1429475 and 0.1661985, respectively. For subjects with CVD event, the Cox model is much more accurate than the FHS model whereas for subjects without CVD, the Cox risk model overestimates the risk rate. Overall, the risk scale of the Cox model is consistent with the Framingham model, which highlights that the proposed Cox model is par with the FHS model.

Table 6: Data summary for samples in the longitudinal validation.

Samples	Gender	CVD	Diabetes
Sample 1	Male	✘	✘
Sample 2	Male	√	√
Sample 3	Female	✘	✘
Sample 4	Female	√	√

In the longitudinal validation process, we selected four sex-specific subjects with or without CVD at the end of the Framingham Study. A summary of these four subjects is listed in Table 6 to confirm the longitudinal validation of the predicted CVD event.

Table 7: Exams in the Framingham Original Cohort study data set.

Exams	Exam Date Range	Age Range	Mean Age	Attendees
Exam 1	1948 - 1953	28 - 74	44	5209
Exam 2	1950 - 1955	31 - 65	46	4792
Exam 3	1952 - 1956	32 - 67	48	4416
Exam 4	1954 - 1958	34 - 69	50	4541
Exam 5	1956 - 1960	37 - 70	52	4421
Exam 6	1958 - 1963	38 - 72	54	4259
Exam 7	1960 - 1964	40 - 74	55	4191
Exam 8	*1962 - 1966*	*42 - 76*	*57*	*4030*
Exam 9	*1964 - 1968*	*44 - 78*	*59*	*3833*
Exam 10	*1966 - 1970*	*46 - 80*	*61*	*3595*
Exam 11	*1968 - 1971*	*49 - 81*	*62*	*2955*
Exam 12	*1971 - 1974*	*50 - 83*	*64*	*3261*
Exam 13	*1972 - 1976*	*53 - 85*	*66*	*3133*
Exam 14	1975 - 1978	55 - 88	68	2871
Exam 15	1977 - 1979	57 - 89	69	2632
Exam 16	1979 - 1982	59 - 91	70	2351
Exam 17	1981 - 1984	61 - 93	72	2179
Exam 18	1983 - 1985	63 - 94	74	1825
Exam 19	1985 - 1988	65 - 96	75	1541
Exam 20	1986 - 1990	67 - 97	77	1401
Exam 21	1988 - 1992	69 - 99	79	1319
Exam 22	1990 - 1994	72 - 101	80	1166
Exam 23	1992 - 1996	73 - 101	81	1026
Exam 24	1995 - 1998	76 - 103	83	831
Exam 25	1997 - 1999	78 - 104	84	703
Exam 26	1999 - 2001	79 - 103	86	558
Exam 27	2002 - 2003	82 - 104	87	414
Exam 28	2004 - 2005	84 - 104	89	303
Exam 29	2006 - 2007	85 - 102	91	218
Exam 30	2008 - 2010	88 - 102	92	141
Exam 31	2010 - 2011	90 - 99	92	91
Exam 32	2012 - 2014	93 - 106	96	40

For each sample, data with fixed time intervals (approximately two years) from longitudinal time follow-up are extracted. The data from five exams (Exam 8, Exam 9, Exam 10, Exam 11, and Exam 12) are extracted for comparison. Data summary for sample 1, sample 2, sample 3, and sample 4 are listed in Appendix B. For each sample, the risks of developing CVD in 10 years related to the selected five exams data are separately computed using the Cox model and the Framingham model. Then the trend of risk over the years with 5% error is depicted, as shown in Figure 3. This figure shows that the trend of risks of these two models are consistent and risks for a specific sample increase over time, the dotted trend lines in each graph represent the increase in the CVD risk over time. Also, samples (both male and female) with diabetes that developed CVD will have a higher risk than the ones with no developed CVD.

4. Discussion

It is widely accepted that CVD has become one of the significant public health issue globally [42, 43] and contributes significantly to the annual deaths globally. Previous studies have noted the importance of identifying associated risk factors and the early detection and intervention of CVDs [44–48] and investigated reducing the risk of developing CVD in early stages. Consequently, CVD risk prediction tools based on a single variable or multiple variables have been devised to yield estimates of the CVD risk [6, 8, 9, 14, 49–51].

Motivated by the objective of early detection and risk estimation of CVD, the present study was designed to identify novel CVD risk factors, determine the effect of these factors, and then develop a risk prediction model based on the identified factors. Although risk factors could vary from one specific CVD component to another, there is sufficient evidence that different types of CVD have commonalities of risk factors. We developed and validated a 10-year risk equation for CVD risk using follow-up data rigorously measured by the Framingham Heart Study.

This investigation extends the number of risk factors by the previous general CVD risk formulations, incorporating heart rate to estimate absolute CVD risk. The approach used in this research is based on advanced statistical techniques that allow reducing the bias in the assessment of true CVD risk. The whole process of data analysis strictly follows the guideline of regression modelling strategies and survival analysis [34, 52].

We use continuous variables (age, BMI, SBP, and pulse rate) to generate the model that performs better than other similar models developed using categorical variables. Compared with simpler approaches that try to make inferences of 5-year and 10-year risk models such as the model based on logistic regression analysis [53] and the CVD risk model using Kaplan-Meier and log-rank test [46], the proposed Cox risk model is more adequate and will avoid severe errors of underestimation or overestimation [22, 34]. Moreover, this model was developed based on a more substantial number of samples and events, suggesting a valid estimation of the real risk.

4.1. Comparison with Other CVD Risk Prediction Tools. The old version Framingham general CVD risk function [53] is useful for identifying persons at high risk of CVD, but it was based on a limited number of risk factors (serum cholesterol, SBP, smoking history, electrocardiogram, and glucose intolerance). The new Framingham laboratory-test-based formula [6] included HDL cholesterol in the risk function. The QRISK study investigators incorporated family history as a novel risk factor by the Framingham general formulas [8]. Although researchers have published risk scores [6, 8, 53] for predicting general CVDs, these functions did not include heart rate in the risk model.

Risk models formulated by using machine learning or data mining techniques have incorporated heart rate as a risk factor but tools that can predict CVD absolute risk are fewer. For example, a prediction tool [54] focuses on the classification of CVD event by employing the ANN and the Bayesian classifier based on heart rate variability. The diagnosis CVD model [27] categorizes the CVD risk as different levels but an absolute risk score cannot be obtained. Even though a supportive tool [19] will generate the estimate

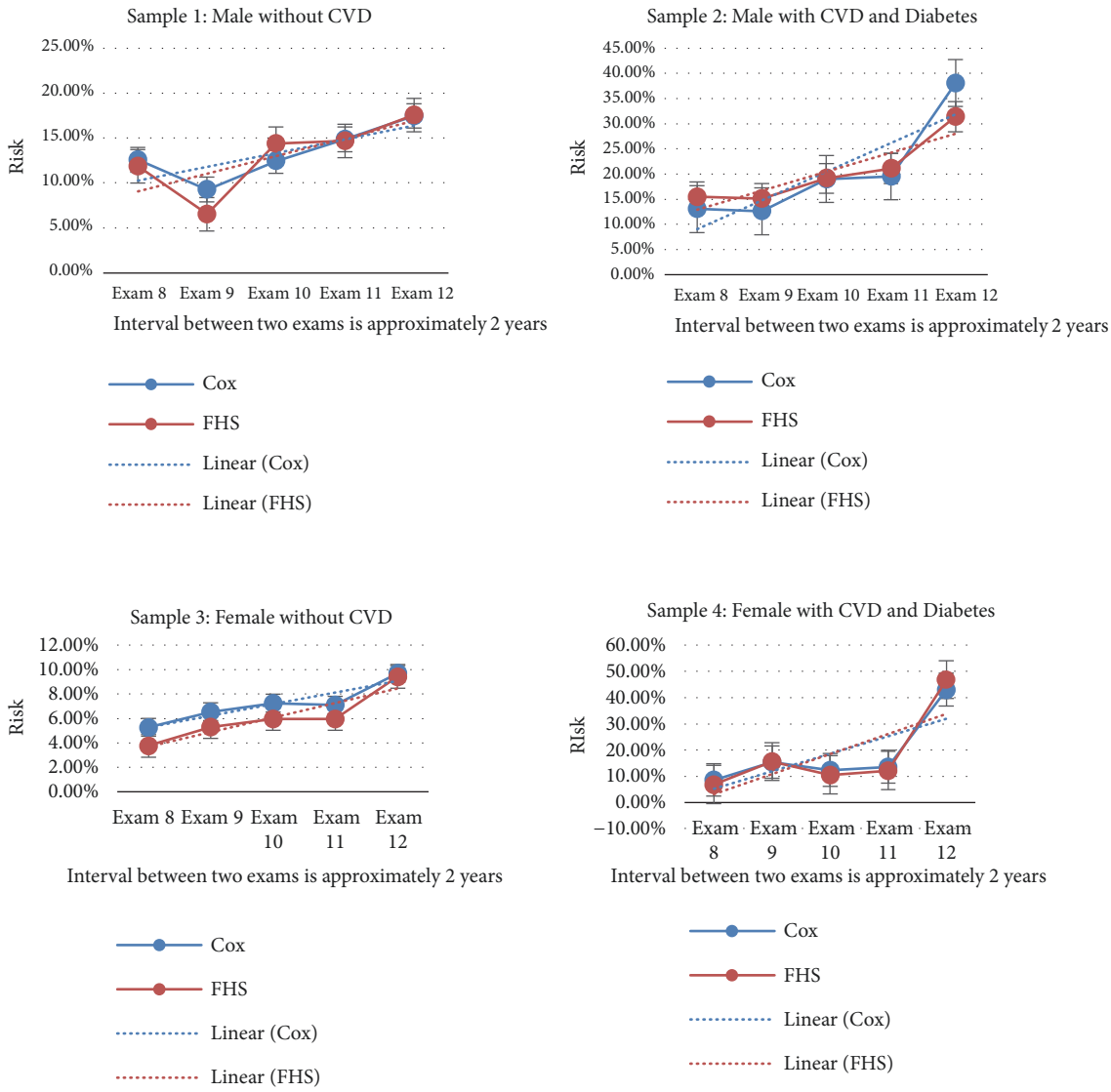

Figure 3: Longitudinal validation.

of a risk score, but the user can not know how many years the score is targeting.

Some equations only focused on specific CVD outcomes. The Europe SCORE project equations were developed for the fatal cardiovascular event [9]. These risk estimation tools [7, 14, 30] are just for coronary heart disease. Also, there are some risk models aiming stroke [16, 55]. Compared with these disease-specific models to estimate the risk of developing specific CVD outcomes, the present study generated a general CVD risk tool that could predict a global CVD risk as well as the risk of developing individual components.

Moreover, compared with the laboratory-based algorithms, the present research proposed a more straightforward way to estimate 10-year CVD risk based on risk factors. An individual can assess his or her CVD risk during an office visit or his monitoring of the combination of risk factors in the risk model, either manually or use some devices like wearable sensors.

4.2. Implication. The CVD risk prediction model could be implemented at the primary care for population analysis and identifying the high-risk individual. This would be a transformation in healthcare management of CVD at an individual as well as at a population level. However, with a small event size of diabetes, caution must be applied to the practice of this risk model. Even though we have used multiple imputation methods to impute the missing values for diabetes, the original feature of data in-balance, which decides that the imputed data frame for the "diabetes" might still have a data in-balance there. Advanced imputation methods need to be considered in the future for avoiding unexpected outcome caused by the diabetes data in-balance.

Our research aims to provide a CVD prediction model based on key risk factors, so that it can be used at the point-of-care for better and informed decision making. Thus, risk factors based on a clinical test such as total cholesterol, HDL cholesterol were not included, but some of these risk factors

Table 8: Exam data for Sample 1: male without CVD.

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	44	26.386894	120	82	40	0	0	0	1
Exam 9	45	26.826676	120	80	0	0	0	0	0
Exam 10	47	27.467643	118	70	20	0	0	0	1
Exam 11	49	28.222249	110	76	44	0	0	0	1
Exam 12	52	28.675012	110	80	50	0	0	0	1

Table 9: Exam data for Sample 2: male with CVD and diabetes.

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	45	27.74258	132	83	20	0	0	0	1
Exam 9	47	26.26118	124	80	20	0	0	0	1
Exam 10	49	27.664352	130	78	20	0	1	0	1
Exam 11	51	27.121914	130	90	20	0	1	0	1
Exam 12	53	24.816551	122	82	20	0	0	1	1

Table 10: Exam data for Sample 3: female without CVD.

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	44	20.776333	110	70	20	0	0	0	1
Exam 9	46	20.265439	120	70	20	0	0	0	1
Exam 10	48	22.312012	118	73	20	0	0	0	1
Exam 11	50	21.797119	114	82	20	0	0	0	1
Exam 12	52	21.797119	130	76	20	0	0	0	1

Table 11: Exam data for Sample 4: female with CVD and diabetes.

Exams	age	bmi	bps	pr	cgrpd	trt	hyp	dia	smk
Exam 8	46	21.793044	130	65	3	0	1	0	1
Exam 9	48	21.967388	170	75	16	0	1	0	1
Exam 10	50	22.494583	140	60	8	0	1	0	1
Exam 11	53	22.31746	140	63	8	0	1	0	1
Exam 12	54	23.380197	160	58	2	1	1	1	1

have a substantial effect on the development of CVD. We have provided a valid framework for creating a risk model using the Cox regression model; future work should consider risk factors not included in our model at this moment. Thus, expanding more predictors into the risk model is an important issue for future research.

5. Conclusion

The proposed study devised a risk prediction model based on multivariable predictors. A novel risk factor "heart rate" was incorporated into this risk equation by conventional risk factors. A satisfying predictive ability with C-index (AUROC) of 0.71 was obtained, which ensures the accuracy of estimating risk scores. Compared with studies focusing on specific diseases, the proposed algorithm can be applied to measure the 10-year risk of CVD. Health care professionals, public health physicians, practice managers, and individuals can run the proposed model to quantify risk at a population level, during patient consultation and identify high-risk individuals for further preventive health care for the entire practice.

Appendix

A. Exams in the Framingham Original Cohort Study Dataset

See Table 7.

B. Data Summary for Samples

See Tables 8–11.

C. Computation of Absolute Risk

Here, we take a specific subject to illustrate the process of risk score calculation. This sample is a 44-year-old man not having diabetes and hypertension. He has a systolic blood

Table 12: Data summary for the subject 15018644.

PREDICTORS	VALUES	UNITS
AGE	44	YEARS
SEX	1	MALE
BMI	26.38689413	KG/M2
HYPERTENSION	0	NO
TREATMENT OF HYPERTENSION	0	NO
BLOOD PRESSURE SYSTOLIC	120	MM HG
CIGARETTES PER DAY	40	LAPSE
SMOKING	1	YES
PULSE RATE	82	PER MINUTE
DIABETES	0	NO
COX MODEL RISK	*12.57%*	
FHS MODEL RISK	*11.86%*	

pressure of 120 mm Hg, pulse rate of 82 per minute, BMI of 26.38689413 kg/m_2 and is a current smoker smoking 40 lapses per day, as shown in Table 12.

The risk estimate based on the Cox model is calculated as follows:

$$\begin{aligned}\sum_{i=1}^{k}\beta_i X_i &= 2.083643 * \log(44) - 0.469719 * 1 \\ &\quad + 0.608864 * \log(26.386894) + 0.241461 \\ &\quad * 0 + 1.682571 * \log(120) - 0.302090 \\ &\quad * \log(82) + 0.009669 * 40 + 1.087501 \\ &\quad * 0 = 16.518741\end{aligned} \tag{C.1}$$

$$\begin{aligned}\sum_{i=1}^{k}\beta_i \overline{X}_i &= 2.083643 * 3.768 - 0.469719 * 1.548 \\ &\quad + 0.608864 * 3.230 + 0.241461 * 0.1469 \\ &\quad + 1.682571 * 4.913 - 0.302090 * 4.311 \\ &\quad + 0.009669 * 13.96 + 1.087501 * 0.02001 \\ &= 16.518741\end{aligned} \tag{C.2}$$

$$\begin{aligned}\widehat{H(t)} &= 1 - \left[S_0(t)\right]^{\exp\left(\sum_{i=1}^{k}\beta_i X_i - \sum_{i=1}^{k}\beta_i \overline{X}_i\right)} \\ &= 1 - 0.9027267^{\exp(16518741 - 16.247045)} \\ &= 0.125658 \approx 12.57\%\end{aligned} \tag{C.3}$$

Data Availability

The cardiovascular disease (CVD) data used to support the findings of this study were supplied by Framingham Heart Study-Cohort (FHS-Cohort) under license and so cannot be made freely available. Requests for access to these data should be made with Open BioLINCC Studies Group through this website https://biolincc.nhlbi.nih.gov/studies/framcohort/.

Additional Points

The main contribution of the present study is developing a risk prediction model for early detection of CVD. More specifically, the contribution can be summarized in four major respects: firstly, a novel risk factor "heart rate" was identified as significant for the development of CVD; secondly, an CVD risk prediction model aiming for early detection of CVD was developed based on various risk factors; thirdly, an absolute risk score in 10 years of CVD can be calculated using this risk model; lastly, multiple forms of the risk estimation of CVD, namely risk equation and nomogram, were also developed.

Conflicts of Interest

Authors declare no conflicts of interest.

Authors' Contributions

All authors contributed equally.

References

[1] S. Mendis, P. Puska, B. Norrving et al., *Global Atlas on Cardiovascular Disease Prevention and Control*, World Health Organization, 2011.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go et al., "Heart disease and stroke statistics update: a report from the American Heart Association," *Circulation*, vol. 131, no. 4, pp. e29–e322, 2015.

[3] W. C. Chan, C. Wright, T. Riddell et al., "Ethnic and socioeconomic disparities in the prevalence of cardiovascular disease in New Zealand," *The New Zealand Medical Journal*, vol. 121, no. 1285, 2008.

[4] Heart Foundation, *General heart statistics in New Zealand*, Heart Foundation, 2017, https://www.heartfoundation.org.nz/statistics.

[5] H. C. McGill, C. A. McMahan, and S. S. Gidding, "Preventing heart disease in the 21st century implications of the pathobiological determinants of atherosclerosis in youth (PDAY) study," *Circulation*, vol. 117, no. 9, pp. 1216–1227, 2008.

[6] R. B. D'Agostino Sr., R. S. Vasan, M. J. Pencina et al., "General cardiovascular risk profile for use in primary care: the Framingham heart study," *Circulation*, vol. 117, no. 6, pp. 743–753, 2008.

[7] D. M. Lloyd-Jones, P. W. F. Wilson, M. G. Larson et al., "Framingham risk score and prediction of lifetime risk for coronary heart disease," *American Journal of Cardiology*, vol. 94, no. 1, pp. 20–24, 2004.

[8] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study," *British Medical Journal*, vol. 335, no. 7611, pp. 136–141, 2007.

[9] R. M. Conroy, K. Pyörälä, A. P. Fitzgerald et al., "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, 2003.

[10] M. Woodward, P. Brindle, and H. Tunsfall-Pedoe, "Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)," *Heart*, vol. 93, no. 2, pp. 172–176, 2007.

[11] G. Assmann, P. Cullen, and H. Schulte, "Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Münster (PROCAM) study," *Circulation*, vol. 105, no. 3, pp. 310–315, 2002.

[12] M. Ferrario, P. Chiodini, L. E. Chambless et al., "Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation," *International Journal of Epidemiology*, vol. 34, no. 2, pp. 413–421, 2005.

[13] S. Wells, T. Riddell, A. Kerr et al., "Cohort profile: the PREDICT cardiovascular disease cohort in New Zealand primary care (PREDICT-CVD 19)," *International Journal of Epidemiology*, vol. 46, no. 1, pp. 22-22, 2017.

[14] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[15] Cardiovascular Disease Risk Assessment Steering Group and others, New Zealand primary care hand- book 2012. Wellington: Ministry of health; 2013 (2017).

[16] J. Yu, L. Dai, Q. Zhao et al., "Association of cumulative exposure to resting heart rate with risk of stroke in general population: the Kailuan cohort study," *Journal of Stroke and Cerebrovascular Diseases*, vol. 26, no. 11, pp. 2501–2509, 2017.

[17] K. H. Han, K. C. Park, M. J. Kim, Y. S. Kim, and H. Chun, "Association between heart rate variability and 10-year atherosclerotic cardiovascular disease risk score," *Atherosclerosis*, vol. 263, pp. e190–e191, 2017.

[18] L. Murukesan, M. Murugappan, M. Iqbal, and K. Saravanan, "Machine learning approach for sudden cardiac arrest prediction based on optimal heart rate variability features," *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 4, pp. 521–532, 2014.

[19] P. Unnikrishnan, D. K. Kumar, S. Poosapadi Arjunan, H. Kumar, P. Mitchell, and R. Kawasaki, "Development of health parameter model for risk prediction of CVD using SVM," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 3016245, 7 pages, 2016.

[20] A. Cannon, *Reliability Data Banks*, Springer Science & Business Media, 2012.

[21] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[22] D. R. Cox, "Regression models and life-tables," in *Breakthroughs in Statistics*, Springer Series in Statistics, pp. 527–541, Springer, New York, NY, USA, 1992.

[23] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Health Informatics Journal*, vol. 19, no. 2, pp. 121–129, 2013.

[24] J. Kim, J. Lee, and Y. Lee, "Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree," *Health Informatics Journal*, vol. 21, no. 3, pp. 167–174, 2015.

[25] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction," *Semantic Scholar*, 2011.

[26] P. Melillo, R. Izzo, A. Orrico et al., "Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis," *PLoS ONE*, vol. 10, no. 3, Article ID e0118504, 2015.

[27] S. Vaanathi, "Cardiovascular disease prediction using fuzzy logic expert system," *IUP Journal of Computer Sciences*, vol. 11, no. 3, 2017.

[28] T. R. Dawber, W. B. Kannel, and L. P. Lyell, "An approach to longitudinal studies in a community: the Framingham Study," *Annals of the New York Academy of Sciences*, vol. 107, no. 1, pp. 539–556, 1963.

[29] W. B. Kannel, M. Feinleib, P. M. Mcnamara, R. J. Garrison, and W. P. Castelli, "An investigation of coronary heart disease in families: The framingham offspring study," *American Journal of Epidemiology*, vol. 110, no. 3, pp. 281–290, 1979.

[30] R. H. Eckel, W. W. Barouch, and A. G. Ershow, "Report of the national heart, lung, and blood institute-national institute of diabetes and digestive and kidney diseases working group on the pathophysiology of obesity-associated cardiovascular disease," *Circulation*, vol. 105, no. 24, pp. 2923–2928, 2002.

[31] E. T. Lee and J. Wang, *Statistical Methods for Survival Data Analysis*, vol. 476, JohnWiley & Sons, 2003.

[32] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports*, vol. 50, no. 3, pp. 163–170, 1966.

[33] B. Efron, "The efficiency of Cox's likelihood function for censored data," *Journal of the American Statistical Association*, vol. 72, no. 359, pp. 557–565, 1977.

[34] F. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer, 2015.

[35] R. Ihaka and R. R. Gentleman, "A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[36] S. Van Buuren, *Flexible Imputation of Missing Data*, CRC Press, 2012.

[37] W. F. Kuhfeld, The prinqual procedure, SAS/STAT Users Guide 2. pp. 1265–1323. 1990.

[38] I.-G. Chong and C.-H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1-2, pp. 103–112, 2005.

[39] M. W. Kattan, "Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer," *Current Opinion in Urology*, vol. 13, no. 2, pp. 111–116, 2003.

[40] M. W. Kattan, P. W. Kantoff, M. Kattan et al., "Comparison of Cox regression with other methods for determining prediction models and nomograms," *The Journal of Urology*, vol. 170, no. 6, pp. S6–S10, 2003.

[41] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[42] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[43] D. S. Hay, *Cardiovascular Disease in New Zealand, 2004: A Summary of Recent Statistical Information*, National Heart Foundation of New Zealand, 2004.

[44] H. B. Hubert, M. Feinleib, P. M. McNamara, and W. P. Castelli, "Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study," *Circulation*, vol. 67, no. 5, pp. 968–977, 1983.

[45] L. Cupples, "Some risk factors related to the annual incidence of cardiovascular disease and death using pooled repeated biennial measurements," *Framingham Heart Study*, 1987.

[46] D. E. Weiner, H. Tighiouart, M. G. Amin et al., "Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: a pooled analysis of community-based studies," *Journal of the American Society of Nephrology*, vol. 15, no. 5, pp. 1307–1315, 2004.

[47] M. Böhm, K. Swedberg, M. Komajda et al., "Heart rate as a risk factor in chronic heart failure (SHIFT): The association between heart rate and outcomes in a randomised placebo-controlled trial," *The Lancet*, vol. 376, no. 9744, pp. 886–894, 2010.

[48] M. C. Odden, M. G. Shlipak, H. E. Whitson et al., "Risk factors for cardiovascular disease across the spectrum of older age: the Cardiovascular Health Study," *Atherosclerosis*, vol. 237, no. 1, pp. 336–342, 2014.

[49] W. De Ruijter, R. G. J. Westendorp, W. J. J. Assendelft et al., "Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study," *BMJ*, vol. 338, no. 7688, pp. 219–222, 2009.

[50] M. J. Pencina, R. B. D'Agostino, M. G. Larson, J. M. Massaro, and R. S. Vasan, "Predicting the 30-year risk of cardiovascular disease: the framingham heart study," *Circulation*, vol. 119, no. 24, pp. 3078–3084, 2009.

[51] L. Bannink, S. Wells, J. Broad, T. Riddell, and R. Jackson, "Web-based assessment of cardiovascular disease risk in routine primary care practice in New Zealand: the first 18,000 patients (PREDICT CVD-1)," *The New Zealand Medical Journal*, vol. 119, no. 1245, 2006.

[52] D. G. Kleinbaum and M. Klein, *Survival Analysis*, vol. 3, Springer, 2010.

[53] W. B. Kannel, D. McGee, and T. Gordon, "A general cardiovascular risk profile: the Framingham study," *American Journal of Cardiology*, vol. 38, no. 1, pp. 46–51, 1976.

[54] H. Kim, M. I. Ishag, M. Piao, T. Kwon, and K. H. Ryu, "A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries," *Symmetry*, vol. 8, no. 6, article 47, 2016.

[55] P. Parmar, R. Krishnamurthi, M. A. Ikram et al., "The stroke riskometerTM app: validation of a data collection tool and stroke risk predictor," *International Journal of Stroke*, vol. 10, no. 2, pp. 231–244, 2015.