

Original Article



# Development and External Validation of a Deep Learning Algorithm for Prognostication of Cardiovascular Outcomes

In-Jeong Cho , MD<sup>1,2,3</sup>, Ji Min Sung , PhD<sup>3</sup>, Hyeon Chang Kim , MD, PhD<sup>3,4</sup>, Sang-Eun Lee , MD<sup>3</sup>, Myeong-Hun Chae , BS<sup>5</sup>, Maryam Kavousi , MD, PhD<sup>6</sup>, Oscar L. Rueda-Ochoa , MD, MSc<sup>6,7</sup>, M. Arfan Ikram , MD, PhD<sup>6,8</sup>, Oscar H. Franco , MD, PhD<sup>6</sup>, James K Min , MD<sup>9</sup>, and Hyuk-Jae Chang , MD, PhD<sup>3,10</sup>

OPEN ACCESS

**Received:** Mar 29, 2019

**Revised:** Jun 10, 2019

**Accepted:** Aug 7, 2019

**Correspondence to**

**Hyuk-Jae Chang, MD, PhD**

Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.  
E-mail: hjchang@yuhs.ac

Copyright © 2020. The Korean Society of Cardiology

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

**ORCID iDs**

In-Jeong Cho   
<https://orcid.org/0000-0002-1209-5129>  
Ji Min Sung   
<https://orcid.org/0000-0003-1958-7596>  
Hyeon Chang Kim   
<https://orcid.org/0000-0001-7867-1240>  
Sang-Eun Lee   
<https://orcid.org/0000-0001-6645-4038>  
Myeong-Hun Chae   
<https://orcid.org/0000-0001-6525-4827>  
Maryam Kavousi   
<https://orcid.org/0000-0001-5976-6519>  
Oscar L. Rueda-Ochoa   
<https://orcid.org/0000-0003-2684-9005>  
M. Arfan Ikram   
<https://orcid.org/0000-0003-0372-8585>

<sup>1</sup>Division of Cardiology, Department of Internal Medicine, Ewha Womans University Seoul Hospital, Ewha Womans University College of Medicine, Seoul, Korea

<sup>2</sup>Ewha Womans University Graduate School, Seoul, Korea

<sup>3</sup>Division of Cardiology, Severance Cardiovascular Hospital, Yonsei University College of Medicine, Seoul, Korea

<sup>4</sup>Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, Korea

<sup>5</sup>AI R&D Lab. of Selvas AI Inc., Seoul, Korea

<sup>6</sup>Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands

<sup>7</sup>School of Medicine, Faculty of Health, Universidad Industrial de Santander UIS, Bucaramanga, Colombia

<sup>8</sup>Department of Radiology, Erasmus MC, Rotterdam, The Netherlands

<sup>9</sup>Department of Radiology and Medicine, Weill Cornell Medical College, Dalio Institute of Cardiovascular Imaging, New York-Presbyterian Hospital, New York, NY, USA




<sup>10</sup>Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, Korea

## ABSTRACT

**Background and Objectives:** We aim to explore the additional discriminative accuracy of a deep learning (DL) algorithm using repeated-measures data for identifying people at high risk for cardiovascular disease (CVD), compared to Cox hazard regression.

**Methods:** Two CVD prediction models were developed from National Health Insurance Service-Health Screening Cohort (NHIS-HEALS): a Cox regression model and a DL model. Performance of each model was assessed in the internal and 2 external validation cohorts in Koreans (National Health Insurance Service-National Sample Cohort; NHIS-NSC) and in Europeans (Rotterdam Study). A total of 412,030 adults in the NHIS-HEALS; 178,875 adults in the NHIS-NSC; and the 4,296 adults in Rotterdam Study were included.

**Results:** Mean ages was 52 years (46% women) and there were 25,777 events (6.3%) in NHIS-HEALS during the follow-up. In internal validation, the DL approach demonstrated a C-statistic of 0.896 (95% confidence interval, 0.886–0.907) in men and 0.921 (0.908–0.934) in women and improved reclassification compared with Cox regression (net reclassification index [NRI], 24.8% in men, 29.0% in women). In external validation with NHIS-NSC, DL demonstrated a C-statistic of 0.868 (0.860–0.876) in men and 0.889 (0.876–0.898) in women, and improved reclassification compared with Cox regression (NRI, 24.9% in men, 26.2% in women). In external validation applied to the Rotterdam Study, DL demonstrated a C-statistic of 0.860 (0.824–0.897) in men and 0.867 (0.830–0.903) in women, and improved reclassification compared with Cox regression (NRI, 36.9% in men, 31.8% in women).

Oscar H. Franco   
<https://orcid.org/0000-0002-4606-4929>  
 James K Min   
<https://orcid.org/0000-0001-8812-7388>  
 Hyuk-Jae Chang   
<https://orcid.org/0000-0002-6139-7545>

**Trial Registration**

ClinicalTrials.gov Identifier: [NCT02931500](https://clinicaltrials.gov/ct2/show/study/NCT02931500)

**Funding**

This research was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (No. 2017-0-00255, Autonomous Digital Companion Development and No.2018-0-00861, Intelligent SW Technology Development for Medical Data Analysis).

The Rotterdam Study is funded by Erasmus MC and Erasmus University, Rotterdam, The Netherlands; The Netherlands Organisation for Scientific Research (NWO); The Netherlands Organisation for the Health Research and Development (ZonMw); The Research Institute for Diseases in the Elderly (RIDE); The Ministry of Education, Culture and Science; the Ministry for Health, Welfare and Sports; The European Commission (DG XII); and The Municipality of Rotterdam. M. Kavousi is supported by the NWO VENI grant (VENI, 91616079). Oscar L. Rueda is supported by a scholarship by COLCIENCIAS and Universidad Industrial de Santander from Colombia. O.H. Franco works in ErasmusAGE, a center for aging research across the life course funded by Nestlé Nutrition (Nestec Ltd.); Metagenics Inc.; and AXA. The funding sources had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review or approval of the manuscript; and decision to submit the manuscript for publication.

**Conflict of Interest**

Myeong-Hun Chae is an employee of Selvas AI Inc, which contributed to the development of deep learning models described in the study. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

**Author Contributions**

Conceptualization: Chang HJ; Data curation: Sung JM, Kavousi M, Franco OH; Formal analysis: Sung JM, Chae MH, Kavousi M, Rueda-Ochoa OL, Ikram MA; Funding acquisition: Chang HJ; Investigation: Cho

**Conclusions:** A DL algorithm exhibited greater discriminative accuracy than Cox model approaches.

**Trial Registration:** ClinicalTrials.gov Identifier: [NCT02931500](https://clinicaltrials.gov/ct2/show/study/NCT02931500)

**Keywords:** Cardiovascular diseases; Artificial intelligence

**INTRODUCTION**

As clustering of risk factors are associated with development of cardiovascular disease (CVD), various prediction models have been developed to identify high-risk individuals for CVD. Traditional approaches for CVD prediction models have been used Cox hazard regression-based analyses.<sup>1-3)</sup> These models identify risk factors in terms of odds or hazard ratios and then provide 10-year risks for CVD, which enables treatment strategies tailored to an individual.<sup>4)</sup>

Hazard regression-based models use pre-specified risk factors, which must meet the assumption of independence between predictors.<sup>4)</sup> In a prospective cohort, because pre-selected risk factors are measured at pre-planned time points, the collected risk factor information can be fully exploited by statistical methods. However, in clinical practice, as types and cycles of risk factor measurement vary widely, conventional statistical models cannot use all of the available risk information but use only a part of such database. The modern hospital information system (HIS) has generated a complex, time-series digitalized health dataset. However, the proper analytic method has not been clearly defined to maximize predictive performance using these numerous, repeated-measures datasets.

Deep learning is a class of machine learning algorithms<sup>5)</sup> and demonstrates excellent performance in classification. The overall transformations have multiple layers in deep learning<sup>4)</sup> and this capacity could enhance predictive model performance in complex time-varying datasets. To date, several small studies have explored the potential of deep learning for disease prediction based on data from specific time points.<sup>6-8)</sup> The purpose of this study was to evaluate the discriminative accuracy of a deep learning-based prediction algorithm to integrate repeated-measures health examination data for prediction of CVD, and to compare it with conventional Cox hazard regression analysis.

**METHODS**

This study was approved and exempted from informed consent by the Institutional Review Board of Yonsei University, Severance Hospital in Seoul, Korea (IRB No.4-2016-0383). This study used National Health Insurance Service-Health Screening Cohort (NHIS-HEALS)<sup>9)</sup> data and National Health Insurance Service-National Sample Cohort (NHIS-NCS)<sup>10)</sup> data derived from a national health screening program and the national health insurance claim database in South Korea and prospective cohort data from the Rotterdam Study.<sup>11)</sup> Data in NHIS-HEALS and NHIS-NCS were fully anonymized for all analyses and informed consent was not specifically obtained from each participant. In the Rotterdam Study, all data were collected in a standardized manner according to the pre-determined study protocol and informed consent was obtained from all participants.

IJ; Methodology: Cho IJ, Kim HC; Writing - original draft: Cho IJ; Writing - review & editing: Kim HC, Lee SE, Rueda-Ochoa OL, Ikram MA, Franco OH, Min JK, Chang HJ.

### Study population for development and internal validation

The National Health Insurance Service (NHIS) provides insurance benefits and free health screening programs for all citizens and residents of Korea. All adults over 40 years old are recommended to undergo periodic health examinations and the participation rate was as high as 74.8% in 2014.<sup>12)</sup> The NHIS constructed the NHIS-HEALS cohort consisting of data for 514,795 people (age 40–79 years), who had been randomly sampled from 10% of the source population who had undergone the NHIS health examination in 2002 or 2003. The cohort was followed up till either a participant's disqualification from health services or the end of the study period in 2013. Individuals who were free from CVD at baseline and had health examinations at least 2 times during follow-up period were included in the analysis.

### Allocation of study datasets for model development

The study population in NHIS-HEALS was randomly divided into 3 groups—development, calibration, and validation datasets. The development dataset was used to build a model for fitting the parameters of the predictors; the calibration dataset was used to tune the parameters to prevent over-fitting in the training model; and the validation dataset was used to evaluate the prediction performance of the developed model. For improving predictive accuracy of the deep learning algorithm, the imbalance between those with CVD and without CVD was adjusted by under-sampling of the majority class. That is, those with and without CVD were constituted with a ratio of 1:1 for the development and calibration datasets to build a model, which has been traditionally applied in deep learning methods for dealing with imbalanced data, because predictive accuracy is impaired when data are imbalanced.<sup>13)</sup>

### Study population for external validation

We validated the prediction models in 2 external cohorts, the NHIS-NSC and the Rotterdam Study. The NHIS-NSC is a national sample cohort representing all Korean age groups, in which 10% of the entire Koreans with health insurance were randomly sampled and followed up from 2002 until 2013. A subgroup of those aged between 40 and 79 years was used for external validation. Since the Rotterdam Study consists almost exclusively of Caucasian subjects, it is ethnically and geographically different from NHIS-HEALS (**Supplementary Data 1** for the details of the Rotterdam Study). For external validation, individuals who were free from CVD at baseline and had health examinations at least two times were enrolled according to the same criteria for the development cohort.

### Outcomes

The primary outcome was defined as the occurrence of one of the following events during the follow-up period after the baseline health examination: 1) death from CVD (International Classification of Diseases 10th edition [ICD-10] codes I-00 to I-99), 2) records of hospitalization due to myocardial infarction, coronary arterial intervention or bypass surgery, and 3) records of hospitalization due to stroke.

### Risk predictors used in model building

To develop the risk model, an a priori decision which assumed the following variables; age, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), total cholesterol (TC), fasting plasma glucose (FPG), current smoking, and exercise - as predictor variables was made. Methods for risk factor measure are shown in **Supplementary Data 2**. Details of variables included in Cox regression and deep learning are described in **Supplementary Table 1**. Variables with missing data less than 4% were included in the analysis. Multiple imputation by fully conditional specification was used in cases where the data was missing.

### Development of prediction models

As it has been known that significant differences existed in the relations between risk factors and CVD occurrence according to gender, CVD prediction models were developed separately for men and women. Data from baseline health examination and repeated measurements in periodic follow-up examinations were used to build prediction models. The time to event was defined as the time between the date of the first health examination and that of the first diagnosis of an event.

The Cox proportional hazard regression model was used to develop the statistical risk prediction model. Cox regression model used the mean, minimum and maximum values and standard deviations (SDs) for continuous variables and the mean and SDs for categorical variables, which were calculated from the periodic health screening data. Detailed method for this Cox model using longitudinal data and its improved accuracy over single-measure method has been described previously.<sup>14)</sup>

For the deep learning algorithm, a Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM)<sup>15)</sup> network was used. Deep learning algorithm was constituted by using the same variables used in Cox regression model with longitudinal data. This study used the methods combining survival analysis and deep learning for comparison of the performances between Cox hazard regression and deep learning. Some recent studies proposed replacing the linear part  $\beta^T x$  in  $f(x)$  with nonlinear deep learning neural network analyses in Cox hazard regression.<sup>16)</sup> In these studies, they proved that the methods combining two algorithms worked well in standard linear function like Cox regression but also as well in the nonlinear settings like deep learning. Thus, this study also presented the results by replacing the exponential part  $\beta^T x$  in  $f(x)$  of Cox regression with nonlinear deep learning so that it could make survival prediction from NHIS-HEALS sequential cohort data. The details of deep learning and model building process are demonstrated in **Supplementary Data 3**, and the significance of included variables are shown in **Supplementary Table 2**.

### Evaluation of prediction performance

The prediction performance of each prediction model were evaluated in the internal and external validation cohorts. Model discrimination was quantified by calculating C-statistics for the survival model. Reclassification performance was also evaluated using a reclassification table and the net reclassification improvement (NRI) index between Cox regression and deep learning model. For calculation of NRI, 3 CVD risk categories were used; <10%, 10% to <20%, and  $\geq 20\%$  in 10 years. Model calibration was assessed by comparing observed and predicted event probabilities. Observed and predicted risks were compared by plotting 2 CVD event occurrence probabilities and indicated the Hosmer-Lemeshow  $\chi^2$  statistics and Brier score which are a measure of the fit of the model. All statistical analyses were conducted with SAS (version 9.4; SAS Inc., Cary, NC, USA) and the R Statistical Package (www.R-project.org). The statistical significance criterion was set at 2-sided  $p < 0.05$ .

## RESULTS

Of the original 514,866 individuals in NHIS-HEALS, those with histories of CVD according to their questionnaires, those who had records of CVD diagnosis at baseline, or those who received health examinations less than 2 times were excluded. The remaining 412,030 individuals constituted study population for model development and internal validation.

The individuals were divided into those with CVD (25,777 individuals, 6.3%) and those without CVD (386,253 individuals, 93.7%) at any point during the follow-up period thereafter.

In individuals with CVD, they were randomly allocated to the development, calibration and validation dataset with a ratio of 6:2:2 (15,466, 5,156, and 5,155 individuals, respectively). The number of individuals with and without CVD were allocated in a ratio of 1:1 in development and calibration cohort. The ratio of 6.3:93.7, which reflects actual occurrence of CVD in the original NHIS-HEALS cohort, were used for the validation cohort. Therefore, in individuals without CVD, they were randomized as follows; 15,466 individual for development, 5,156 individual for calibration, and 77,202 individuals for validation. Therefore, a total of 30,932 individuals were used for development; 10,312 individuals for calibration; and 82,357 individuals for validation (**Figure 1**). Details of the population selection process in the NHIS-NSC and the Rotterdam Study are also described in **Figure 1**.

**Table 1** shows baseline characteristics of development and validation datasets. External validation was performed in 178,875 persons (women 51.0%) in the NHIS-NSC and 4,296 (women 50.6%) from the Rotterdam Study. Mean follow-up duration was approximately 10 years in all 3 validation datasets. Mean number of health examinations was lower in the Rotterdam Study than in the NHIS-HEALS or NHIS-NSC cohorts. The cumulative CVD event rate was higher in the Rotterdam Study (11.7%) than in the NHIS-HEALS (6.3%) and NHIS-NSC (5.2%).

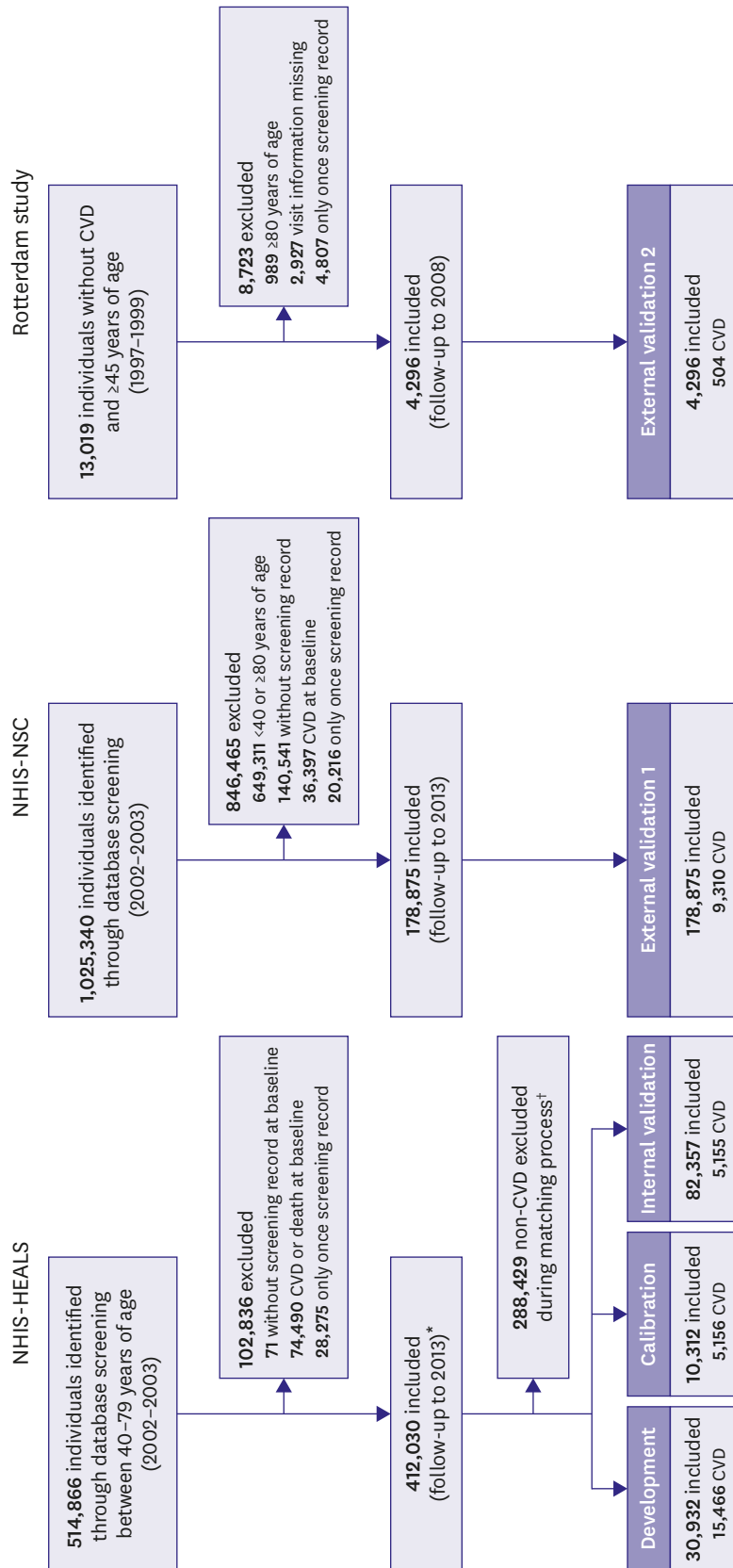
Relative risk estimates for the Cox model predictors were described in **Supplementary Table 3**. Age, BMI, DBP, current smoking and exercise were associated with CVD in men while age, SBP, TC and exercise were associated with CVD in women. Moreover, SDs of CVD risk factors were more important compared to their mean values in Cox models using repeated measures.

Performance indicators of each model are demonstrated in **Tables 2 and 3**. C-statistics (95% confidence interval [CI]) for Cox regression and deep learning were 0.813 (0.803–0.823) and 0.896 (0.886–0.907) in men and 0.837 (0.825–0.849) and 0.921 (0.908–0.934) in women in the internal validation dataset, suggesting improved performance from Cox regression to deep learning. From Cox regression to deep learning, the NRI was 24.8% ( $p < 0.001$ ) in men and 29.0% ( $p < 0.001$ ) in women. In external validation with NHIS-NSC (External validation 1), deep learning had a C-statistics of 0.868 (95% CI, 0.86–0.88) in men and 0.889 (0.88–0.91)

**Table 1.** Baseline characteristics of the development, internal validation and external validation cohorts

Variable	Development		Internal validation		External validation 1		External validation 2	
	Men (n=18,009)	Women (n=12,923)	Men (n=44,694)	Women (n=37,663)	Men (n=87,687)	Women (n=91,188)	Men (n=1,686)	Women (n=2,610)
Age (years)	54.6±9.9	56.3±10.2	51.3±8.9	52.7±9.3	49.4±8.8	49.9±9.1	66.2±6.2	66.7±6.3
Hypertension	6,094 (25.6)	5,594 (31.6)	7,468 (16.7)	7,725 (20.5)	13,685 (15.6)	16,491 (18.1)	1,073 (63.6)	1,670 (64.0)
Diabetes	2,926 (12.3)	1,684 (9.5)	3,774 (8.4)	2,019 (5.4)	8,071 (9.2)	5,089 (5.6)	185 (11.0)	236 (9.0)
Smoking	8,126 (45.1)	418 (3.2)	19,300 (43.2)	1,005 (2.7)	39,748 (45.3)	2,893 (3.2)	400 (23.7)	469 (18.0)
Exercise	8,712 (48.4)	3,985 (30.8)	22,782 (51.0)	12,520 (33.2)	43,614 (49.7)	33,651 (36.9)	1,286 (76.3)	1,934 (74.1)
BMI (kg/m <sup>2</sup> )	24.0±2.9	24.1±3.1	24.0±2.9	23.9±3.1	24.0±2.9	23.9±3.1	26.7±3.3	27.3±4.4
SBP (mmHg)	131.2±18.2	127.7±19.5	128.1±17.1	124.2±18.5	128.3±17.2	124.5±18.1	142.8±20.7	140.7±20.6
DBP (mmHg)	82.4±11.7	78.8±12.0	81.1±11.4	77.2±11.7	80.8±11.2	77.2±11.4	79.3±11.1	76.1±10.4
FPG (mg/dL)	103.1±40.3	100.1±43.5	99.5±35.9	95.0±30.7	100.5±33.9	95.5±27.8	107.1±25.4	103.9±23.3
TC (mg/dL)	200.6±39.6	205.4±40.2	198.8±38.0	201.9±39.5	198.2±37.9	202.3±38.3	218.9±36.5	233.5±35.6
Follow-up (years)	8.5±2.8	8.9±2.6	10.3±1.7	10.4±1.4	9.6±2.9	9.3±2.2	10.9±2.3	11.4±2.2
No. of health screening	5.0±2.7	4.4±2.0	6.1±2.9	5.2±2.1	5.2±2.7	4.4±1.9	2.9±0.5	2.6±0.5

BMI = body mass index; DBP = diastolic blood pressure; FPG = fasting plasma glucose; SBP = systolic blood pressure; TC = total cholesterol.



**Figure 1.** Formation of the development, internal validation, and external validation cohorts. CVD = cardiovascular disease; NHIS-HEALS = National Health Insurance Service-Health Screening Cohort; NHIS-NSC = National Health Insurance Service-National Sample Cohort. \*Individuals with CVD and non-CVD were defined according to CVD occurrence during the mean 9.8 ± 2.2 years follow-up period to 2013; †123,601 out of 412,030 individuals were randomly selected as the final dataset to deal with the imbalanced data between CVD and non-CVD. During the matching process for the development and validation datasets, 288,429 non-CVD were excluded.

**Table 2.** Predictive performance of FRS and prediction models of men in internal and external validation cohorts

Statistics	Men					
	Internal validation		External validation 1		External validation 2	
	Cox regression	Deep learning	Cox regression	Deep learning	Cox regression	Deep learning
C-statistic (95% CI)	0.813 (0.803–0.823)	0.896 (0.886–0.907)	0.801 (0.793–0.809)	0.868 (0.860–0.876)	0.779 (0.742–0.816)	0.860 (0.824–0.897)
$\chi^2$ test (p value)	5.02 (0.833)	5.28 (0.809)	10.42 (0.318)	9.70 (0.375)	4.30 (0.891)	5.96 (0.744)
Brier score	0.058	0.040	0.056	0.041	0.114	0.087
	Cox regression to deep learning		Cox regression to deep learning		Cox regression to deep learning	
Difference in C-statistic (95% CI)	0.083 (0.069–0.097)		0.067 (0.055–0.079)		0.081 (0.030–0.132)	
p value	<0.001		<0.001		0.002	
Reclassification						
Cases move to higher	1,223 (39.1%)		2,195 (39.7%)		106 (42.2%)	
Cases move to lower	619 (19.8%)		1,079 (19.5%)		53 (21.1%)	
Controls moved to higher	2,812 (6.8%)		8,189 (10.0%)		197 (13.7%)	
Controls moved to lower	5,074 (12.2%)		12,041 (14.7%)		424 (29.5%)	
NRI (%)	24.76		24.88		36.93	
p value	<0.001		<0.001		<0.001	

CI = confidence interval; FRS = Framingham risk score; NRI = net reclassification improvement.

**Table 3.** Predictive performance of FRS and prediction models of women in internal and external validation cohorts

Statistics	Women					
	Internal validation		External validation 1		External validation 2	
	Cox regression	Deep learning	Cox regression	Deep learning	Cox regression	Deep learning
C-statistic (95% CI)	0.837 (0.825–0.849)	0.921 (0.908–0.934)	0.836 (0.826–0.846)	0.889 (0.879–0.898)	0.792 (0.755–0.829)	0.867 (0.830–0.903)
$\chi^2$ test (p-value)	2.58 (0.979)	3.28 (0.952)	6.39 (0.700)	4.46 (0.879)	2.15 (0.989)	2.40 (0.984)
Brier score	0.045	0.031	0.038	0.027	0.082	0.064
	Cox regression to deep learning		Cox regression to deep learning		Cox regression to deep learning	
Difference in C-statistic (95% CI)	0.084 (0.066–0.102)		0.053 (0.039–0.067)		0.075 (0.022–0.128)	
p value	<0.001		<0.001		0.005	
Reclassification						
Cases moved to higher	837 (41.3%)		1,602 (42.4%)		110 (43.5%)	
Cases moved to lower	316 (15.6%)		734 (19.4%)		49 (19.4%)	
Controls moved to higher	1,437 (4.0%)		5,166 (5.9%)		201 (8.5%)	
Controls moved to lower	2,604 (7.3%)		7,987 (9.1%)		383 (16.2%)	
NRI (%)	29.02		26.18		31.83	
p value	<0.001		<0.001		<0.001	

CI = confidence interval; FRS = Framingham risk score; NRI = net reclassification improvement.

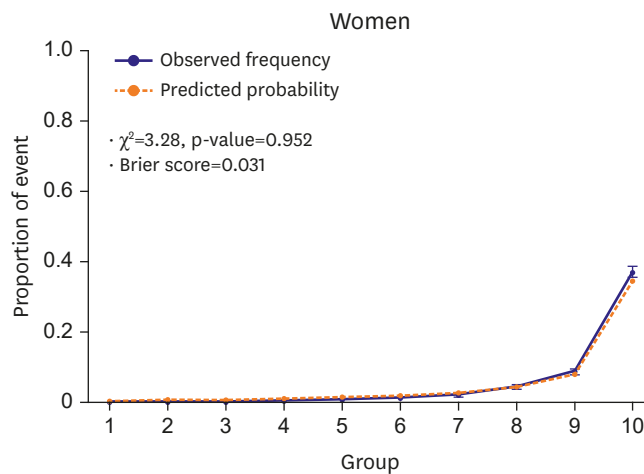
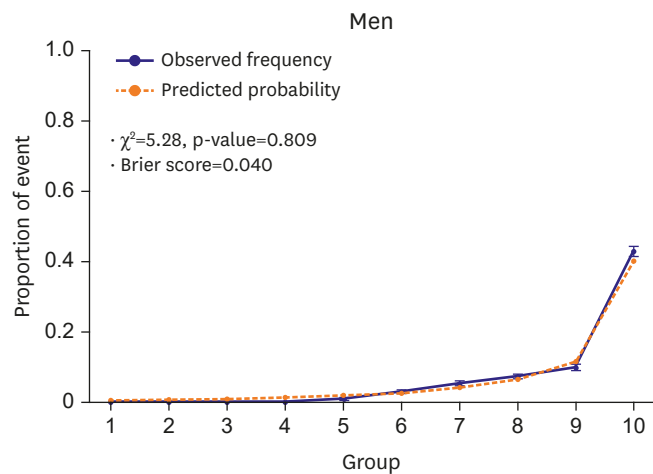
in women and improved reclassification compared with Cox regression (NRI, 24.9% in men, 26.2% in women, all  $p < 0.001$ ). In external validation with Rotterdam Study (External validation 2), deep learning had a C-statistics of 0.860 (95% CI, 0.82–0.90) in men and 0.867 (0.83–0.90) in women and improved reclassification compared with Cox regression (NRI, 36.9% in men, 31.8% in women, all  $p < 0.001$ ).

The Brier scores in each deep learning model indicated good calibration between the estimated predicted risk and observed risk (Tables 2 and 3). A calibration plots for deep learning models also confirmed good agreement between the estimated predicted risk and observed risk, grouped by decile of risk (Figure 2).

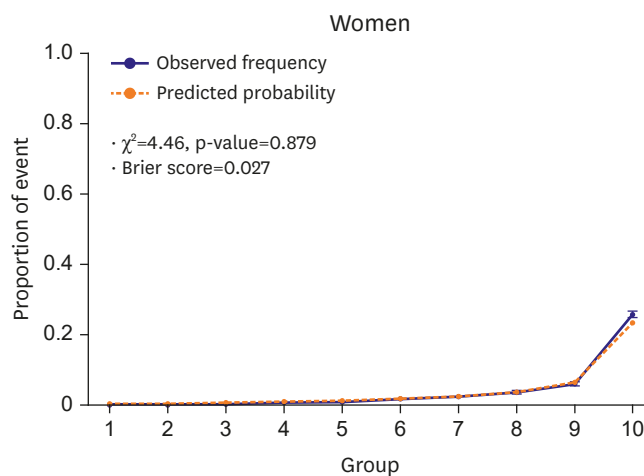
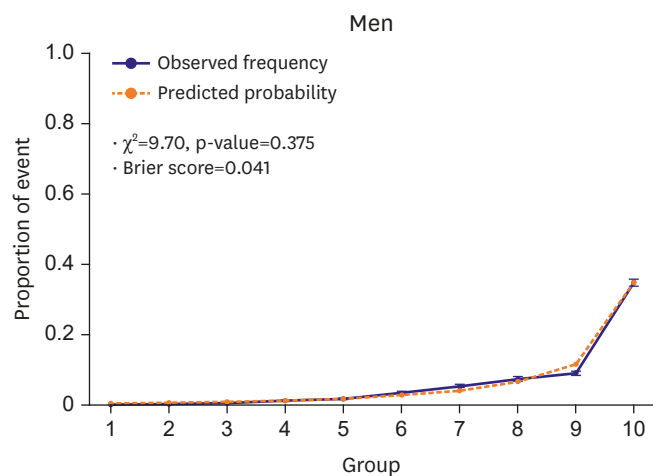
## DISCUSSION

The principal findings of this study were as follows: 1) A deep learning algorithm significantly improved predictive performance over the conventional statistical approach when analyzing a large repeated-measures data for prediction of CVD. 2) Better performance of the deep

**A. Internal validation**



**B. External validation 1**



**C. External validation 2**

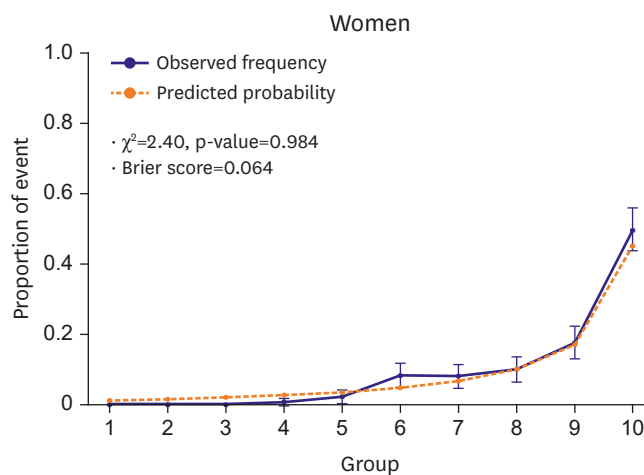
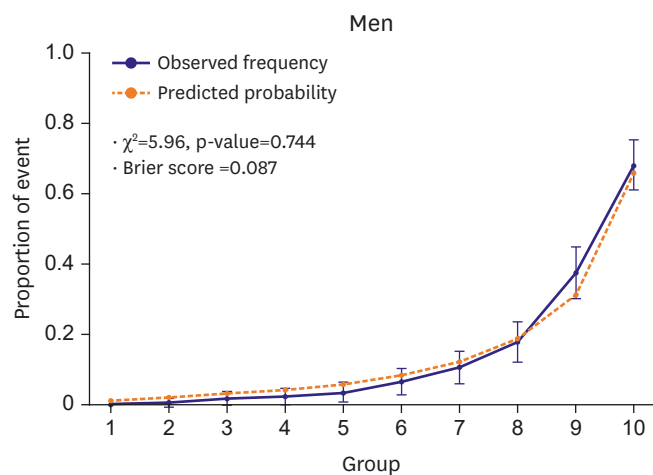


Figure 2. Predicted vs. observed probability of cardiovascular disease by deep learning in the internal validation and external validation cohorts.



learning algorithm over Cox regression analysis was confirmed by external validation in the Rotterdam Study, a different ethnicity, as well as in a different South Korean cohort.

Various risk prediction models have been proposed for the purpose of CVD prediction.<sup>1-3)17)18)</sup> However, since the accuracy of the CVD prediction models based on risk factor or statistics is not satisfactory in all circumstances,<sup>19)</sup> various attempts to increase the predictability are continuing. The most common approach is to add new biomarkers as predictors to improve disease predictability.<sup>20)21)</sup> However, apart from costs, biomarkers such as coronary artery calcium score has demonstrated only modest increase in predictive accuracy when added to a traditional risk factor model.<sup>22)</sup>

Since electronic health record was introduced several decades ago, huge amount of data has been accumulating in the medical field. In the current study, deep learning model using periodic risk factor measures showed better predictive accuracy over traditional Cox hazard regression approach. We used longitudinal Cox regression model incorporating mean and variability information into the Cox model, which showed better performance compared to the single measured method.<sup>14)</sup> Still, Cox regression demonstrated lower predictive accuracy than deep learning. A nationwide repeated health screening system like that used in South Korea may not be applicable to all health care systems. However, accumulation of large-scale data is accelerating in the medical field as HIS is advancing as a healthcare platform. The deep learning model provided good discrimination and calibration using these repeated data and can, therefore, may be a valuable tool for risk predictions in the era of electronic health record.

Deep learning, which have shown high value for many classification problems,<sup>23)</sup> is different from Cox regression-based statistics in many respects. A Cox regression model needs assumptions of proportional changes in the hazards being predicted and independence among pre-specified variables and does not reflect variable changes over time. In contrast, deep learning is agnostic in any assumptions and fully uses variable that are constantly changing into their models. Therefore, deep learning is a more proper method for analyzing data from daily clinical practice, where numerous confounding factors exist and risk factors for each individual change continuously. Moreover, the risk factors are closely related to each other and their interactions are complex. In the current study, traditional risk factors derived from prior cohort studies such as BMI and TC were not significant predictors for CVD in regression models, which corresponds to previous findings derived from analysis of various hospital data that showed many traditional risk factors were less significant factors for CVD occurrence.<sup>24)</sup> In this regard, deep learning may be more suitable for analyzing complex time-varying data derived from standard clinical practice, which may differ greatly from data derived from prospective controlled trials.

In both the NHIS-NSC of a different Korean population and the Rotterdam Study of a Europeans, deep learning approach showed improved discrimination to a considerable extent compared to the traditional statistical approach, implying its robust predictive power to be highly generalizable in geographically-disparate ethnically-diverse settings. Application of the developed model in this present study to other large-scale ethnic cohorts—such as blacks and Hispanics—now appears warranted. Models derived from machine learning, including deep learning, are fundamentally dynamic, and can incorporate new data for the continuous update and optimization of its algorithm, which improves its predictive performance over time.<sup>25)26)</sup> More importantly, models derived from machine learning algorithm can be locally retrained in diverse populations to maximize accuracy in different populations of patients

with varying clinical and demographic profiles.<sup>27)</sup> The paradigm shift of methodology for building a prediction model from traditional hypothesis-driven statistical analysis to these self-training deep learning methods can help offer insights for precision medicine into personalized disease prediction. Regarding the importance of a continuous learning system and informatics tools to assist health-care providers in interpreting data and tailoring decisions of treatment to a patient in precision medicine,<sup>28)</sup> the deep learning algorithm represents new opportunities for physicians to engage in precision medicine by providing precise information of CVD risk in each individual.

This study had several limitations. First, the level of risk factors can be modified by drug or non-drug treatment during the course of follow-up, thereby changing the risk of CVD, and these modifications may be unpredictable based upon physician and patient behavior. While predict models may be influenced by such confounders, it is likely to be more affected by single point measures. This study showed that deep learning methods can further improve CVD prediction ability using repeatedly-measured information, suggesting the strengths of deep learning for these data. Second, outcome events were ascertained from health insurance claims data and diagnoses were not adjudicated by medical records or laboratory tests as in other large data studies. However, evaluation of the deep learning approach in a prospectively enrolled cohort study nevertheless demonstrates the robustness of this model in populations in which events were prospectively ascertained. Third, some risk factors some risk predictors included in the prediction model did not satisfy the assumptions required for proportional hazard analysis (**Supplementary Table 4**). However, we used proportional hazard analysis, which has been the most widely used for CVD prediction, because our main purpose was to assess whether we could improve the predictive power of the models by using repeatedly measured data. Fourth, incorporation of more complex information can enhance the predictive power of deep learning, since the benefit of deep learning is its capacity to deal with large complex data without any assumptions. However, since the purpose of the current study was to confirm the superior analytic performance of deep learning to that of Cox regression, we used the same variables for both models. Further studies using a large number of variables would be needed to validate predictive performance of deep learning with increase in number of variables. Fifth, the number of health screening was relatively small in Rotterdam Study compared to NHIS-HEALS and NHIS-NSC. This might be one of the reasons for the lower C-statistics in Rotterdam Study. Sixth, the problem of imbalanced data was adjusted by under-sampling of the majority class. Further studies with more complex validation scenario and experiments with different class ratios are needed.

In conclusion, a time-series deep learning algorithm analysis of periodic health screening data resulted in predictive models for CVD outcomes that had greater discriminative accuracy than conventional statistical approaches. However, the utility of this model in clinical care requires further research.

## ACKNOWLEDGEMENTS

This study used NHIS-HEALS data (NHIS-2016-2-132) and NHIS-NSC data (NHIS-2017-2-300) from the National Health Insurance Service (NHIS). The authors declare no conflicts of interest with the NHIS.

## SUPPLEMENTARY MATERIALS

### Supplementary Data 1

Study design of Rotterdam Study

[Click here to view](#)

### Supplementary Data 2

Methods for measurement of risk factors

[Click here to view](#)

### Supplementary Data 3

Model building and training in the recurrent neural network

[Click here to view](#)

### Supplementary Table 1

Variables used in each prediction model

[Click here to view](#)

### Supplementary Table 2

Attribute ranking

[Click here to view](#)

### Supplementary Table 3

Hazard ratios for cardiovascular disease risk factors in the Cox regression model in the development dataset

[Click here to view](#)

### Supplementary Table 4

Test of the proportional hazards assumption in Cox regression model

[Click here to view](#)

## REFERENCES

1. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24:987-1003.  
[PUBMED](#) | [CROSSREF](#)
2. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335:136.  
[PUBMED](#) | [CROSSREF](#)
3. D'Agostino RB Sr, Grundy S, Sullivan LM, Wilson P; CHD Risk Prediction Group. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001;286:180-7.  
[PUBMED](#) | [CROSSREF](#)

4. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J* 2017;38:1805-14.  
[PUBMED](#) | [CROSSREF](#)
5. Deo RC. Machine learning in medicine. *Circulation* 2015;132:1920-30.  
[PUBMED](#) | [CROSSREF](#)
6. Narain R, Saxena S, Goyal AK. Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach. *Patient Prefer Adherence* 2016;10:1259-70.  
[PUBMED](#) | [CROSSREF](#)
7. Khatibi V, Montazer GA. A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. *Expert Syst Appl* 2010;37:8536-42.  
[CROSSREF](#)
8. Kukar M, Kononenko I, Grošelj C, Kralj K, Fettich J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artif Intell Med* 1999;16:25-50.  
[PUBMED](#) | [CROSSREF](#)
9. Seong SC, Kim YY, Park SK, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. *BMJ Open* 2017;7:e016640.  
[PUBMED](#) | [CROSSREF](#)
10. Seong SC, Kim YY, Khang YH, et al. Data resource profile: the National Health Information Database of the National Health Insurance Service in South Korea. *Int J Epidemiol* 2017;46:799-800.  
[PUBMED](#) | [CROSSREF](#)
11. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol* 2015;30:661-708.  
[PUBMED](#) | [CROSSREF](#)
12. National Health Insurance Service. *National Health Screening Statistical Yearbook 2014*. Wonju: National Health Insurance Service; 2015.
13. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf Sci* 2013;250:113-41.  
[CROSSREF](#)
14. Cho JJ, Sung JM, Chang HJ, Chung N, Kim HC. Incremental value of repeated risk factor measurements for cardiovascular disease prediction in middle-aged Korean adults: results from the NHIS-HEALS (National Health Insurance System-National Health Screening Cohort). *Circ Cardiovasc Qual Outcomes* 2017;10:004197.  
[PUBMED](#) | [CROSSREF](#)
15. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735-80.  
[PUBMED](#) | [CROSSREF](#)
16. Liao L, Ahn HI. Combining deep learning and survival analysis for asset health management. *Int J Progn Health Manag* 2016;7:020.
17. Lloyd-Jones DM, Leip EP, Larson MG, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation* 2006;113:791-8.  
[PUBMED](#) | [CROSSREF](#)
18. Pencina MJ, D'Agostino RB Sr, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the Framingham heart study. *Circulation* 2009;119:3078-84.  
[PUBMED](#) | [CROSSREF](#)
19. Ramsay SE, Morris RW, Whincup PH, Papacosta AO, Thomas MC, Wannamethee SG. Prediction of coronary heart disease risk by Framingham and SCORE risk assessments varies by socioeconomic position: results from a study in British men. *Eur J Cardiovasc Prev Rehabil* 2011;18:186-93.  
[PUBMED](#) | [CROSSREF](#)
20. Murphy TP, Dhangana R, Pencina MJ, D'Agostino RB Sr. Ankle-brachial index and cardiovascular risk prediction: an analysis of 11,594 individuals with 10-year follow-up. *Atherosclerosis* 2012;220:160-7.  
[PUBMED](#) | [CROSSREF](#)
21. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann Intern Med* 2009;150:65-72.  
[PUBMED](#) | [CROSSREF](#)
22. Polonsky TS, McClelland RL, Jorgensen NW, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. *JAMA* 2010;303:1610-6.  
[PUBMED](#) | [CROSSREF](#)
23. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag* 2012;29:82-97.  
[CROSSREF](#)

24. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Med Care* 2013;51:251-8.  
[PUBMED](#) | [CROSSREF](#)
25. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. *Am J Gastroenterol* 2010;105:1224-6.  
[PUBMED](#) | [CROSSREF](#)
26. Jung K, Shah NH. Implications of non-stationarity on predictive modeling using EHRs. *J Biomed Inform* 2015;58:168-74.  
[PUBMED](#) | [CROSSREF](#)
27. Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J Vasc Surg* 2016;64:1515-1522.e3.  
[PUBMED](#) | [CROSSREF](#)
28. Antman EM, Loscalzo J. Precision medicine in cardiology. *Nat Rev Cardiol* 2016;13:591-602.  
[PUBMED](#) | [CROSSREF](#)