의학박사 학위논문

Deep learning based survival analysis model for cardiovascular risk assessment improves with a hybrid approach in combination with Cox regression: integrated data on healthcare and environmental exposure

딥러닝 기반 생존분석이 적용된 심혈관질환 위험 평가 모델 성능 향상을 위한 콕스 모형과 결합된 하이브리드 접근법: 헬스케어-환경 연계 데이터 활용 연구

2020년 08월

서울대학교 대학원

의과학과 의과학 전공

김 규 웅

A Thesis of the Doctor of Philosophy in
Medical Science

딥러닝 기반 생존분석이 적용된 심혈관질환 위험 평가 모델 성능 향상을 위한 콕스 모형과 결합된 하이브리드 접근법: 헬스케어-환경 연계 데이터 활용 연구

Deep learning based survival analysis model for cardiovascular risk assessment improves with a hybrid approach in combination with Cox regression: integrated data on healthcare and environmental exposure

August 2020

Department of Biomedical Sciences

Seoul National University Graduate School

Kyuwoong Kim

Deep learning based survival analysis model for cardiovascular risk assessment improves with a hybrid approach in combination with Cox regression: integrated data on healthcare and environmental exposure

by

Kyuwoong Kim

A Dissertation Submitted to the Department of Biomedical Sciences in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Medical Science at Seoul National University Graduate School

July 2020

Approved by Thesis Committee:

Chair	Hyung-jin Yoon
Vice Chair	Sang Min Park
Examiner	Young Ho Yun
Examiner	Ji Yeob Choi
Examiner	Kyung-Hee Park

딥러닝 기반 생존분석이 적용된 심혈관질환 위험 평가 모델 성능 향상을 위한 콕스 모형과 결합된 하이브리드 접근법: 헬스케어-환경 연계 데이터 활용 연구

지도교수 박 상 민

이 논문을 의학박사 학위논문으로 제출함

2020년 05월

서울대학교 대학원

의과학과 의과학 전공

김규웅

김규웅의 박사 학위논문을 인준함

2020년 07월

위 원 장 윤 형 진 (인)

부위원장 박 상 민 (인)

위 원 윤 영 호 (인)

위 원 최 지 엽 (인)

위 원 박 경 희 (인)

ABSTRACT

Deep learning based survival analysis model for cardiovascular risk assessment improves with a hybrid approach in combination with Cox regression: integrated data on healthcare and environmental exposure

Kyuwoong Kim

Department of Biomedical Sciences

The Graduate School

Seoul National University

Background and aims: The contribution of different cardiovascular disease (CVD) risk factors for the risk evaluation and predictive modeling for incident CVD is often debated. Also, to what extent data on CVD risk factors from multiple data categories should be collected for comprehensive risk assessment and predictive modeling for CVD risk using survival analysis is uncertain despite the increasing availability of the relevant data sources. This study aimed to evaluate the contribution of different data categories derived from integrated data on healthcare and environmental exposure to the risk evaluation and prediction models for CVD risk using deep learning based survival analysis in combination with Cox proportional hazards regression and Cox proportional hazards regression.

Methods: Information on the comprehensive list of CVD risk factors were collected from systematic reviews of variables included in the conventional CVD risk assessment tools and observational studies from medical literature database

(PubMed and Embase). Each risk factor was screened for availability in the National Health Insurance Service-National Sample Cohort (NHIS-NSC) linked to environmental exposure data on cumulative particulate matter and urban green space using residential area code. Individual records of 137,249 patients more than 40 years of age who underwent the biennial national health screening between 2009 and 2010 without previous history of CVD were followed up for incident CVD event from January 1, 2011 to December 31, 2013 in the NHIS-NSC with data linkage to environmental exposure. Statistics-based variable selection methods were implemented as follows: statistical significance, subset with the minimum (best) Akaike Information Criteria (AIC), variables selected from the regularized Cox proportional hazards regression with elastic net penalty, and finally a variable set that commonly meets all the criteria from the abovementioned statistical methods. Prediction models using Cox proportional hazards deep neural network (DeepSurv) and Cox proportional hazards regression were constructed in the training set (80% of the total sample) using input feature sets selected from the abovementioned strategies and progressively adding input features by data categories to examine the relative contribution of each data type to the predictive performance for CVD risk. Performance evaluations of the DeepSurv and Cox proportional hazards regression models for CVD risk were conducted in the test set (20% of the total sample) with Uno's concordance statistics (C-index), which is the most up-to-date evaluation metrics for the survival models with right censored data.

Results: After the comprehensive review, data synthesis, and availability check, a total of 31 risk factors in the categories of sociodemographic, clinical laboratory test and measurement, lifestyle behavior, family history, underlying medical conditions, dental health, medication, and environmental exposure were identified

in the NHIS-NSC linked to environmental exposure data. Among the models constructed with different variable selection methods, using statistically significant variables for DeepSurv (Uno's C-index: 0.7069) and all of the variables for Cox proportional hazards regression (Uno's C-index: 0.7052) showed improved predictive performance for CVD risk, which was a statistically significant increase (*p*-value for difference in Uno's C-index: <0.0001 for both comparisons) compared to the models with basic clinical factors (age, sex, and body mass index), respectively. When all and statistically significant variables in each data category from sociodemographic to environmental exposure were progressively added as input features into DeepSurv and Cox proportional hazards regression for predictive modeling for CVD risk, the DeepSurv model with statistically significant variables pertaining to the sociodemographic factors, clinical laboratory test and measurement, and lifestyle behavior data showed the notable performance that outperformed Cox proportional hazards regression model with statistically significant variables added up to the medication category. Extensive data linkage to environmental exposure on cumulative particulate matter and urban green space offered only marginal improvement for the predictive performance of DeepSurv and Cox proportional hazards regression models for CVD risk.

Conclusion: To obtain the best predictive performance of DeepSurv model for CVD risk with minimum number of input features, information on sociodemographic, clinical laboratory test and measurement, and lifestyle behavior should be primarily collected and used as input features in the NHIS-NSC. Also, the overall performance of DeepSurv for CVD risk assessment was improved with a hybrid approach using statistically significant variables from Cox proportional hazards regression as input features. When all the data categories in the NHIS-NSC

linked to environmental exposure data are available, progressively adding variables in each data category could incrementally increase the predictive performance of DeepSurv model for CVD risk with the hybrid approach. Data linkage to the environmental exposure with residential area code in the NHIS-NSC offered marginally improved performance for CVD risk in both DeepSurv model with the hybrid approach and Cox proportional hazards regression model.

Keywords: cardiovascular disease; healthcare data; environmental exposure; deep learning based survival analysis; Cox proportional hazards regression

Student number: 2016-21973

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

1. Background

Cardiovascular disease (CVD) is a class of serious medical conditions occurring in the blood vessels and heart (e.g. myocardial infarction [MI], stroke, heart failure, and other conditions of the circulatory system) that is one of the leading causes of morbidity and death in the world[1,2]. According to the World Health Organization report, CVD was responsible for an estimated 17.9 million deaths in 2016, which accounted for 31% of all deaths worldwide[3]. Despite the efforts to prevent CVD through interventions and providing information on well-established risk factors in the high-risk regions of CVD, countries in the high-risk regions still account for approximately 75% of CVD mortality in the world[3]. Also, patients diagnosed with CVD often face substantial disease burden due to high healthcare cost and possibility of post-event disability[4-6]. The global burden of CVD continues to rise every year[7] regardless of the widely available CVD risk assessment tools and preventive strategies.

In the past decades, most of the conventional CVD risk assessment tools were developed in the U.S and Europe to estimate future CVD risk based on the easily accessible patient-level data[8,9]. These conventional risk assessment tools are widely used to assess CVD risk in the epidemiologic studies despite the variations in study populations. Also, there are multiple categories of traditional and non-traditional risk factors that are reported to be associated with CVD risk based on the evidence from previous studies such as health behavior (e.g. cigarette smoking and lack of physical activity), dietary factors (e.g. red and processed meat consumption), non-traditional biomarkers (e.g. C-reactive protein), oral

hygiene/dental health (e.g. chronic periodontitis), and environmental factors (e.g. exposure to ambient air pollution[10-12]) (Figure 1).

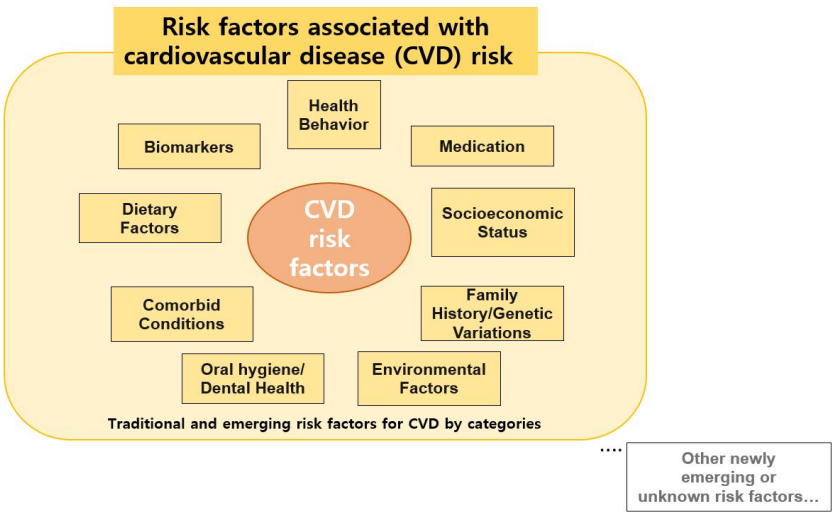

Figure 1. Risk factors for cardiovascular disease (CVD) at-a-glance

However, information on most of the non-traditional risk factors are usually difficult to collect or simply unavailable due to technical challenges on data integration. Whether CVD risk assessment could be improved with additional use of non-traditional risk factors from multiple categories remains uncertain. A recent study reported that the use of information on certain biomarkers added to basic clinical risk factors could contribute to improving the predictive performance of the machine learning (ML) models for identifying atrial fibrillation[13]. Rather than using only conventional risk factors for CVD risk assessment in the dataset for research in health science, developing CVD risk prediction models from multiple categories of data on the risk factors is of importance in preventive cardiology if the use of these additional information could provide a more comprehensive and improved evaluation of future CVD risk.

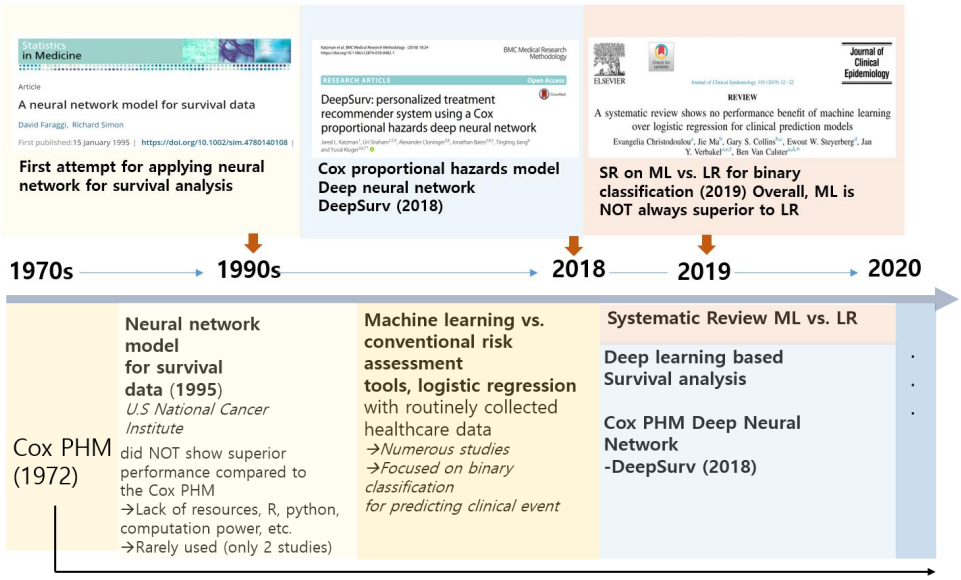

Figure 2. Trends in methodologies used for studies in predicting clinical outcomes in the past decades

Abbreviations: SR, systematic review; ML, machine learning; LR, logistic regression; Cox PHM, Cox proportional hazards model

In addition to the conventional CVD risk assessment tools, clinical event prediction using ML algorithms has been established as an important aspect in data-driven cardiovascular research, especially in the recent years with increasing availability of the data sources and advanced techniques[14]. This recent advance in cardiovascular epidemiology contributed to numerous studies that used ML techniques for predicting CVD outcome using a wide range of variables[15,16]. Most of the representative studies showed that applying ML techniques outperformed the conventional CVD risk assessment tools in predicting the CVD outcome[17,18]. However, a recent meta-analysis published in 2019 by Christodoulou *et al.*, found that ML showed no superior performance to logistic regression in clinical prediction models based on the 71 studies with 282 comparisons from the Medline literature search from January, 2016 to August, 2017. Furthermore, these studies

have focused on binary classification of the future CVD event without incorporating survival analysis with learning-based prediction algorithms. Taking the time element into account in the prediction model can provide more useful assessment in the population-level risk for future CVD risk compared to the models that simply identify the binary outcome.

Due to the recent development of Cox proportional hazards deep neural network, also known as DeepSurv, it is possible to apply survival analysis using multilayer neural networks[19]. In the past two decades, survival analysis using neural network has not been widely developed or applied after the Faraggi-Simon[20] model developed in 1995 did not show improved performance compared to the Cox proportional hazards regression[21,22] (Figure 2). The lack of adaptation of the Faraggi-Simon model was possibly attributable to the lack of computational power or publicly available packages (i.e. compare to the modern day R and Python packages) for implementing the neural network model for survival analysis. Overall, comparing the predictive performance of the DeepSurv and traditional Cox proportional hazards regression with variables derived from vast amount of available healthcare data linked to other sources for CVD risk assessment is of interest for data-driven cardiovascular health research.

2. Research problem

Despite the growing availability of the healthcare data[23,24] that can be used for comprehensively assessing the risk of CVD, majority of the studies have only evaluated risk factors for CVD without fully considering CVD risk factors from other data categories. In these studies, the extent to which the unexamined risk factors associated with CVD could have modified or produced potentially biased

risk estimation is unclear. Also, evidence regarding to what extend the data on CVD risk factors should be collected for the optimal CVD risk assessment is somewhat inconclusive as non-traditional, yet important risk factors for CVD such as ankle-brachial index (ABI), high-sensitivity C-reactive protein (hsCRP) level, coronary artery calcium (CAC) score, and dental health are often not considered in the conventional CVD risk assessment tools.

Table 1. Previous research trend, problem, and unmet need for cardiovascular risk assessment using learning-based algorithms.

Category	**Predictive modeling**	**Feature selection method**	**Model performance**	**Claim DB linked to environmental DB**
Previous research trend	Most studies comparing performance of deep learning models with logistic regression for CVD as a binary outcome	Automated feature selection (e.g. RF and LASSO) for predictive modeling for CVD as a binary outcome	Performance benefit with more data categories were observed in studies with clinical events as binary outcomes	Mostly focused on associations between environmental exposure and CVD outcome (rather than using environmental DB for predictive modeling)
Research problem	Time element (time-to-event) is not considered in the deep-learning models for CVD as a binary outcome, which is theoretically not comparable to Cox PHM	Automated feature selection methods for binary outcome cannot be directly applied for deep-learning based survival analysis models due to the theoretical difference in the outcome (binary vs. time to event)	No well-established methodology for optimizing model performance in deep learning-based survival analysis	Performance benefit of linking healthcare data to environmental data has not been extensively studied, especially in deep learning-based survival analysis
Unmet need	Predictive modeling with survival analysis	Feature selection method for performance improvement in deep learning-based survival analysis	A hybrid approach in combination with Cox PHM while expanding data categories for modeling building	Evaluation of performance benefit for data linkage to environmental data with a hybrid approach

Abbreviations: CVD, cardiovascular disease; DB, database; Cox PHM, Cox proportional hazards model; RF, random forest; LASSO, least absolute shrinkage and selection operator

Whether the information on these non-traditional and other emerging risk factors such as data on dental health and environmental exposure contribute to the predictive performance using survival analysis, especially in the predictive

modeling based on DeepSurv and Cox proportional hazards models is largely unknown.

3. Main hypothesis and objective

3.1. Hypothesis

This study aimed to test the main hypotheses that (1) the predictive performance of DeepSurv and Cox proportional hazards models for incident CVD event using input features with statistics-based variable selection methods from multiple data categories is superior to the model with basic clinical factors and (2) the overall performance for both models would steadily increase as more input features are added from multiple data categories derived from the NHIS database linked to environmental exposure data.

3.2. Objective

The main objectives of this study are as follows:

(A). Comprehensively review CVD risk factors from conventional CVD risk assessment tools and evidence from observations studies and screen for data availability in the National Health Insurance Service (NHIS) database linked to environmental exposure data.

(B). Use operational definition and statistics-based variable selection methods to collect information on comprehensive list of CVD risk factors available in the NHIS database linked to environmental exposure data. Also, conduct a population-based cohort study to check the strength of association between the selected sets of variables and incident CVD event.

(C). Evaluate and compare the predictive performance of DeepSurv and Cox proportional hazards regression for predictive modeling of incident CVD using multiple input features from (1) statistics-based variable selection methods (in comparison to the models with basic clinical factors and factors included in a conventional CVD risk assessment tool) and (2) progressively adding variables in data categories by level of feasibility and accessibility based on the NHIS data (in comparison to the previous model in each step).

II. MATERIALS AND METHODS

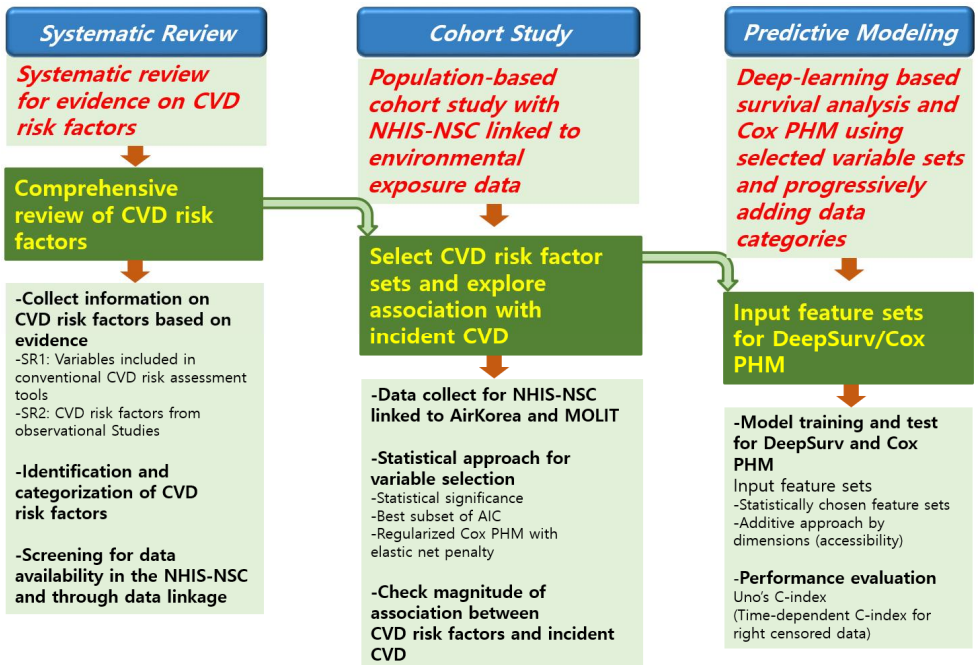

Figure 2. Overview of research methods for integrated data on healthcare and environmental exposure for CVD risk assessment using deep learning based survival analysis and Cox proportional hazards regression

Abbreviations: CVD, cardiovascular disease; SR, systematic review; NHIS-NSC, National Health Insurance Service-National Sample Cohort; MOLIT, Ministry of Land, Infrastructure, and Transport Cox PHM, Cox proportional hazards model

1. Comprehensive review and identification of cardiovascular disease (CVD) risk factors

1.1. Systematic review on variables included in conventional CVD risk assessment tools

To systematically review, identify the risk assessment models for CVD, and abstract data on the included variables in each model, I followed the items listed in the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS), which is the guideline proposed by the Cochrane Prognosis Methods group[25]. Literature search was conducted in PubMed and Embase to identify the published articles from 1 January 1970 to 22 January

2020. After identifying the articles, I manually retrieved articles that had the most up-to-date information on CVD risk assessment tools (e.g. Qrisk3 instead of Qrisk1 and Qrisk2)[26] and reviewed the estimated outcomes, study population, risk factors/predictors (variables) that are included in each model. I excluded the articles that reported the results on external validation or comparing the prognostic value of different models. Prior to qualitative synthesis of the articles that provide information on the CVD risk assessment models, an additional reviewer was recruited in the review process in case of doubt.

1.2. Systematic review on traditional and emerging CVD risk factors from observational studies

I conducted a literature search on PubMed and Embase using a broad search queries adopted from the previous meta-analyses and systematic reviews in the relevant topics (details of the search terms are provided in Table 1). Because the purpose of this study was not focused on quantitative analysis of the selected articles, I did not consider the Meta-Analysis of Observational Studies in Epidemiology (MOOSE) guideline when checking the items reported in each study.

Table 2. Search queries for a comprehensive review of cardiovascular disease (CVD) risk factors

Research database for healthcare and medicine	Search terms
PubMed (https://www.ncbi.nlm.nih.gov/pubmed) 2020.01.22	((“association” [tiab] OR “risk” [tiab] OR “predictor” [tiab] OR “relationship” [tiab]) AND (“myocardial infarction” [tiab] OR “myocardial infarct” [tiab] OR “cardiac infarct” [tiab] OR “heart attack” [tiab] OR “myocardium infarct” [tiab] OR “subendocardial infarct” [tiab] OR “transmural infarct” [tiab] OR “ventricle infarct” [tiab] OR “ventricular infarct” [tiab] OR “stroke”[tiab]) “ischemic stroke”[tiab] OR “hemorrhagic

	stroke”[tiab] OR “cerebrovascular disease"[tiab] OR “cerebrovascular attack"[tiab] OR “cerebral infarct"[tiab] OR “intracranial hemorrhage"[tiab]))
EMBASE (https://www.embase.com) 2020.01.22	((‘association’:ab,ti OR ‘risk’:ab,ti OR ‘predictor’: ab,ti OR ‘relationship’:ab,ti AND (‘heart’:ab,ti OR ‘myocard’:ab,ti OR ‘subendocardial’:ab,ti OR ‘transmural’:ab, ti OR ‘coronary’:ab, ti OR ‘occlusion’: ab, ti OR ‘infarct’: ab,ti OR ‘attack’: ab,ti’ OR ‘stroke’:ab,ti ‘Ischemic stroke’:ab,ti OR ‘hemorrhagic stroke’:ab,ti OR ‘cerebrovascular disease’:ab,ti OR ‘cerebrovascular attack’:ab,ti OR ‘cerebral infarct’:ab,ti OR ‘intracranial hemorrhage’:ab,ti)

The studies included in this review were limited to observational cohort studies with accurate assessment of cardiovascular risk factors and cardiovascular outcomes. The following criteria were considered in the full-text review of the articles identified in the process of screening and considering eligibility for inclusion: (1) cohort design (2) reliable source of data from well-established studies (e.g. the Nurses’ Health Study in the United States) or medical research database (e.g. QResearch database in the United Kingdom) (3) Outcome of the study was clearly defined and was identified from a reliable source of data (4) reporting CVD outcome as hazard ratios or relative risk with 95% confidence intervals from validated statistical models.

Among the studies screened for each risk factor in the systematic review, the representative study was primarily chosen based on the study sample size and publication year after checking for the relevant meta-analysis. Secondary criteria for determining the final study was based on the Scientific Journal Ranking in the

relevant field or notable medical journals (e.g. BMJ, JAMA, Lancet, etc).

1.3. Integration of the comprehensive list of CVD risk factors

In the final process of full-text review for variables included in the conventional CVD risk assessment tools and observational studies, the comprehensive list of CVD risk factors were synthesized after removing duplicate variables. Each variable was assigned to a relevant categories ranging from sociodemographic factors to environmental exposure.

1.4. Screening for data availability

Based on the posteriori knowledge, comprehensive list of CVD risk factors derived from the conventional CVD risk assessment tools and observational studies was screened for availability in the National Health Insurance Service-National Sample Cohort (NHIS-NSC).

2. Cohort analysis for measuring strength of association between risk factors and incident cardiovascular disease

2.1. Study population and linkage to environmental exposure data

The database used in this study is derived from the administrative database derived from the National Health Insurance Service (NHIS) in the Republic of Korea. The National Health Insurance Act was established in 1989 by the Ministry of Health and Welfare in the Republic of Korea, and as the NHIS subsequently began to serve as a quasi-government entity that provides health insurance to the enrollees, which was approximately 97% of the population in the country. Since the NHIS was established as a single-insurer by the government policy, information on the enrollees' demographics, national health screening (health questionnaire and clinical laboratory results), medical/dental claims, medication prescription, and other relevant information had been collected and

managed by the NHIS. As a part of the implementation of the Government 3.0 initiative in the Republic of Korea, which promotes opening and sharing of the database in the public sector, some of the accumulated data in the NHIS had been released for research purpose.

The integrated data on healthcare and environmental exposure used in this study was derived from the NHIS-NSC, which is a nationally representative cohort constructed from approximately 2 % of the target population of the NHIS enrollees (~46,605,433) in 2002 using proportional allocation and random sampling methods. The raw data of NHIS-NSC includes approximately 1 million enrollees, of which their records on insurance eligibility, national health screening, and medical/dental claims could be used for epidemiologic research. Based on the data on their residential area code (administrative district codes in the Republic of Korea), environmental exposure data on cumulative particulate matter (PM 10 derived from AirKorea) and urban green space (provided by the Ministry of Land, Infrastructure and Transport, MOLIT), which excludes the natural green space, and only limited to city parks and artificial green space. The enrollees were limited to those residing in the seven major cities in the Republic of Korea (Seoul, Busan, Incheon, Daegu, Daejeon, Gwangju, and Ulsan) to minimize the confounding effect of natural green space. This integrated database was used to evaluate the predictive performance of deep learning based survival analysis and traditional survival analysis for assessment of future cardiovascular risk using multiple risk factors from claims data to the environmental exposure.

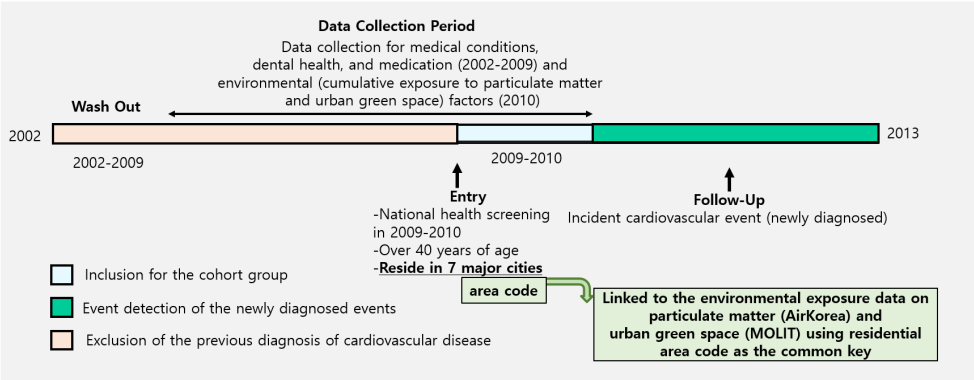

Figure 4. Longitudinal cohort study design with the National Health Insurance Service-National Sample Cohort (NHIS-NSC) linked to environmental exposure data

The study population was limited to the enrollees aged more than 40 years of age who underwent the national health screening between 2009 and 2010 without previous history of CVD and were followed up until 2013. Information on annually reviewed sociodemographic factors for insurance eligibility, medical/dental claims, medication prescription, and environmental exposure were inter-linked with unique keys (Figure 4). Accordingly, 169,942 enrollees in the NHIS-NSC who were older than 40 years of age and underwent the national health screening between 2009 and 2010 residing in the seven major cities (Seoul) with environmental exposure (cumulative exposure to PM 10 and urban green space) were identified. After excluding 32,693 enrollees who were previously diagnosed with CVD (n=31,381), died prior to follow-up (n=1,076), or missing information on health screening data (n=236), a total of 137,249 participants were included in the final analytic cohort based on the NHIS-NSC database linked to the environmental exposure data (Figure 5). Details of the cohort profile and validity of the NHIS-NSC and data linkage to environmental exposure from particulate matter and urban green space have been previously described.

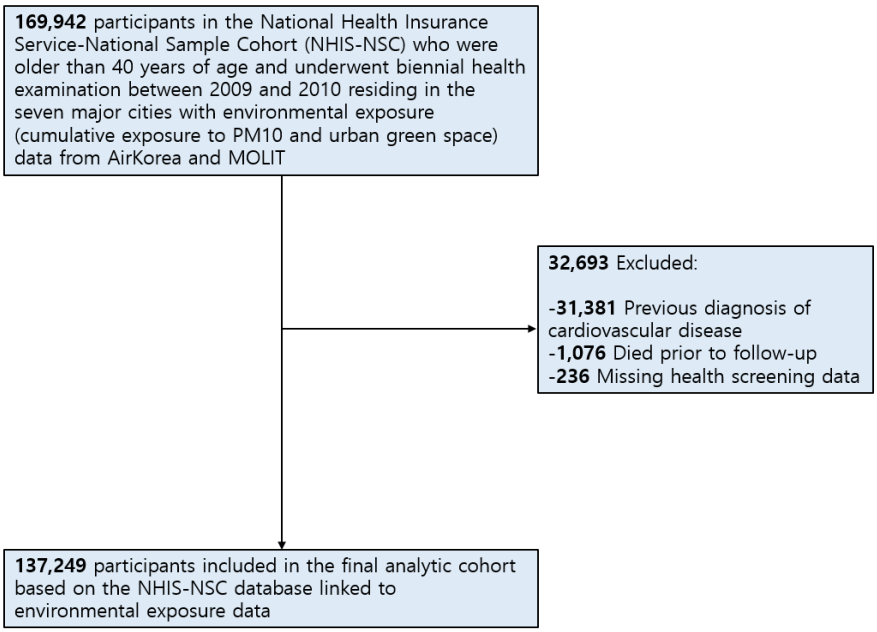

Figure 5. Flow diagram of the study population selection process from the enrollees of the NHIS-NSC linked to environmental exposure data

Abbreviations: NHIS-NSC, National Health Insurance Service-National Sample Cohort; PM, particulate matter; MOLIT, Ministry of Land, Infrastructure, and Transport

Institutional Review Board (IRB) at the Seoul National University Hospital (IRB No.: E-1802-008-918) approved this study, which adheres to the research ethics of the patient-level data and complies to the Declaration of Helsinki. The Review Board at the Big Data Steering Department in the NHIS approved this study for Kyuwoong Kim's Ph.D. dissertation. (Assigned No.: NHIS-2018-2-174). There is no additional data available other than the results reported in this study. To preserve the confidentiality of this population-based data, access to the NHIS-NSC database was only granted to Kyuwoong Kim for the research purpose for this dissertation. Unauthorized use of the NHIS in any form in this study is prohibited by the Private Information Protection Act in the Republic of Korea.

2.2. Variable selection and data processing

Among the final variables included in the comprehensive list of CVD risk

factors from the conventional CVD risk assessment tools and observational studies after systematic review, data synthesis, and availability screening, operational definition based on the previous study with NHIS database for each variable was determined. Prior to applying statistics-based variation selection methods, multiple collinearity test based on variance inflation factor (VIF) was conducted with a cut-off value for multiple collinearity set to VIF>5. Since age could be highly correlated with the underlying conditions identified as CVD risk factors, multiple collinearity for age and underlying conditions was additionally checked.

After checking multiple collinearity, three statistics-based variation selection methods were implemented. First, Cox proportional hazards model was fitted adjusting for all of the available risk factors, and only those that were statistically significant (cut-off point set to $p<0.05$) were selected. Second approach was obtaining the best subset based on the minimum (best) Akaike Information Criteria (AIC) in a full stepwise fashion using significance level for entry (SLENTRY) and significance level for stay (SLSTAY) value close to 1 (SLENTRY=0.99 in and SLSSTAY=0.995). In this process, subset of the explanatory variables (risk factors) with the minimum (best) AIC was chosen. Third, fitting the Cox proportional hazards model regularized by penalty terms with elastic net (combination of $L1$ and $L2$ penalties from the Ridge method and Least Absolute Shrinkage and Selection Operator, LASSO) was used. With the elastic net regularization, penalized regression coefficients in the Cox proportional hazards model shrinked to zero and only the variables with non-zero coefficients were retrieved and selected[27]. Also, the variables meeting all of the three criteria (models with the variables selected from statistical significance, subset with minimum AIC, and elastic net regularization) were additionally considered as a variable selection

method, which was similar to the variable selection approach used by the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Research Group for determining risk factors for CVD among patients with type 1 diabetes[28].

3.3. Population-based cohort analysis

Based on the previously published literature using the NHIS database, the incident CVD event that occurred during the follow-up period (January 1, 2011 to December 31, 2013) in this study were defined using the *International Classification of Diseases, Tenth Revision (ICD-10)* for coronary heart disease (ICD-10: I20-I25) and total stroke (ICD-10: I60-I69) with at least 2 days (48 hours) of hospitalization[29-31]. This operational definition for incident CVD event in the NHIS database using the medical claim records have been reported. To statistically test the proportionality assumption of the Cox regression model, partial residual of each explanatory variable from the model and follow-up time for the individuals with incident CVD event were computed and checked for the correlation independent of change in time (Schoenfeld residual). Additional assessment for the proportionality assumption was graphically tested with log-log plot for age to check if the survival estimates largely differ by age. To examine the strength of association between the risk factors selection from the statistics-based variable selection methods, hazard ratio (HR) and 95% confidence intervals (95% CI) were computed using Cox proportional hazards regression for each variable adjusting for all the other variables in the selected set. Data collection and statistical analyses for statistics-based variable selection and population-based cohort analyses were conducted with 9.4 (SAS Institute, Cary, NC, USA) and R software, version 3.6.3 (R foundation). Statistical test was two-sided and statistical significance was

defined as p values <0.05 for all analyses.

3. Predictive modeling using survival analysis: DeepSurv and Cox proportional hazards regression

3.1. Model development

Predictive modeling with survival analysis for incident CVD in the NHIS-NSC linked to environmental exposure data was conducted with Cox proportional hazards deep neural network (DeepSurv)[19] and Cox proportional hazards regression[32]. DeepSurv is a feed-forward neural network, of which the output is the predicted log-hazard function from the input features that are parametrized by the weights of the network through fully connected and dropout layers (Figure 6). For the model development process with DeepSurv, early stopping (which stops training when the validation loss stops improving) was conducted to avoid overfitting in the training set. DeepSurv was implemented with *Pycox* package in Python 3.7.4. Random hyper-parameter optimization search was adopted for DeepSurv[33]. Cox proportional hazards model is a semiparametric (estimates log-risk function using a linear function without estimating the baseline hazard function) survival model that consists of baseline hazard function and log-risk function.

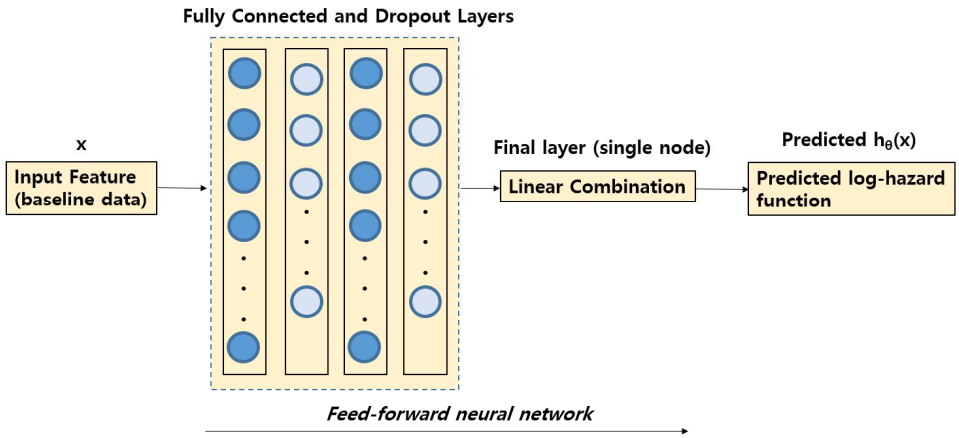

Figure 6. Structure of Cox proportional hazards deep neural network (DeepSurv)

DeepSurv is a feed-forward neural network consists of fully connected and dropout layers where the output of the network is the predicted log-hazard function, which is a part of Cox proportional hazards model, with given input features.

To develop predictive models based on survival analysis using DeepSurv and Cox proportional hazards regression, two approaches were selected for input features. First, the following variable sets were used as input feature sets for DeepSurv and Cox proportional hazards regression: (1) basic clinical factors that can be simply collected from demographics and anthropometric measurement (age, sex, and body mass index) based on a previous study; (2) European Society of Cardiology Systematic Coronary Risk Evaluation (ESC SCORE) factors (age, sex, total cholesterol, systolic blood pressure, and cigarette smoking), which requires minimal information on demographics, lipid profile, blood pressure measurement, and health questionnaire survey on cigarette smoking; (3) Model 1 (all variables), Model 2 (statistically significant variables from Cox proportional hazards model), Model 3 (subset of best AIC), Model 4 (variables selected from regularized Cox proportional hazards regression with elastic net penalty), and Model 5 (variables meeting criteria in Model 2-4). For DeepSurv models, Model 2-5 are considered

hybrid approach since the selected variables used as input features was adopted from the Cox proportional hazards model (i.e. unlike automated variable selection using variable importance in Random Forest or LASSO for machine learning models for binary classification). For sensitivity analysis, number of underlying conditions was replaced for all of the associated health conditions and used as input a feature for predictive modeling. Second, progressively adding 11variables included in each data category ranging from sociodemographic factors to environmental exposure was implemented for predictive modeling with DeepSurv and Cox proportional hazards regression. For variables pertaining to each data category, all variables and statistically significant variables were progressively added as input features.

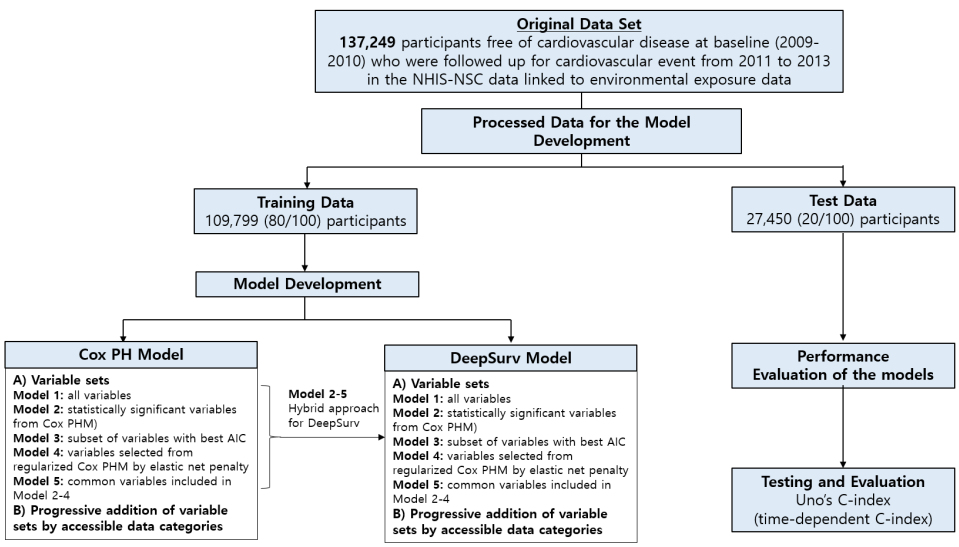

Figure 7. Framework for the model development and performance evaluation with survival analysis using DeepSurv and Cox proportional hazards regression for CVD risk using NHIS-NSC linked to environmental exposure data

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; AIC, Akaike Information Criteria; Cox PHM, Cox proportional hazards model

3.2. Evaluation of the predictive performance of the models

For the evaluation of the predictive performance of each model derived from the Cox proportional hazards regression and DeepSurv, the Uno's method (2011) was chosen because it takes the censoring distribution into account for weighing the uncensored observations in the estimation[34]. Unlike the Harrell's method[35,36] that simply discards the pairs that could not be compared due to the event of censoring, Uno's method is censoring-independent, and thus considered the most up-to-date method for evaluating the area under the curve (AUC) for the survival analysis models with right-censored data. The AUC value presented as Uno's concordance statistics (C-index) can be interpreted as the probability of an individual with the incident CVD event has a higher risk score than an individual without the event[1]. In addition, *p*-value for difference in Uno's C-index was computed to compare the predictive performance of the DeepSurv and Cox proportional hazards regression models with different input features. The *p*-value for difference in Uno's C-index was computed for the models with ESC SCORE factors and models with multivariable factors (Model 1-5) in reference to the model with basic clinical factors. For the models with variables progressively added from sociodemographic factors to environmental exposure, *p*-value for difference in Uno's C-index was compared in reference to the model in the previous step.

III. RESULTS

1. Identification and categorization of cardiovascular disease risk factors

Overall, the total number of CVD risk models was 13 and they were developed based on population-based cohort data from North America (n=5), Europe (n=5), and Asia (n=3) (Table 2). There were no models developed using data from South America and Africa. Most of the models used cohort data collected from a single country whereas European Society of Cardiology Systematic Coronary Risk Evaluation (ESC SCORE), Asia Pacific Asia Cohort Studies, and Keys used cohort data from multiple countries. Among the models that used cohort data from a single country, all of them included different ethnic groups in their data except for Chien (Taiwan) and Korean Risk Prediction Model (Republic of Korea). There was a large variation in the estimated CVD outcomes.

Although most of the models provided risk estimate for incident and fatal CVD events (including heart disease and stroke), some models (PROCAM, Chien, Chambless, Keys, and the Heart Score) were developed for only assessing adverse coronary events such as myocardial infarction. While other models did not particularly provide the information on time frame for the estimated CVD risk, the Framingham Risk Score and the Atherosclerotic Cardiovascular Disease (ASCVD) risk estimator are specified as 10-year risk of fatal CVD and 10-year risk of ischemic heart disease and stroke, respectively. Also, the CVD risk model developed from the Asia Pacific Cohort studies stated that the model provides risk estimation for 8-year risk of CVD. In addition, variables that were used to develop the CVD risk estimation largely differed by the models. While the ESC SCORE

contained the smallest number of variables (n=5), Q-risk3 had the largest number (n=21) of variables that were used to develop the model.

Table 3. List of conventional CVD risk assessment tools developed in North America, Europe, and Asia.

Model/ Developer	Country/ Region	Reference Article	Estimated Outcomes	Study population	Risk factors/predictors (variables)
Framingham Risk Score	USA	D'Agostino et al., (2008)[2]	10-year risk of fatal CVD	Framingham Heart Study	Age, sex, TC, HDL-C, systolic BP smoking, diabetes
ASCVD Risk Estimator (ACC/AHA pooled cohort equation)	USA	Goff et al., (2013)[3]	10-year risk of IHD and stroke	ARIC Study, Cardiovascular Health Study, CARDIA study, Framingham Heart Study and Framingham Offspring study	Age, sex, race, systolic and diastolic BP, TC, HDL-C, LDL-C diabetes, smoking, hypertension treatment, medication use (statin, aspirin)
ESC SCORE	Europe	Conroy et al., (2003)[4]	Fatal CVD events	Pooled dataset from 12 European countries[a]	Age, sex, smoking, TC, systolic BP
Q-risk3	UK (England and Wales)	Hippisley-Cox et al., (2017)[5]	Incident CVD events	QResearch Database (version 41)	Age, sex, ethnicity, systolic BP, BMI, family history of CHD, TC/HDL ratio, Townsend deprivation score, treated hypertension, rheumatoid arthritis, atrial fibrillation, type 2 diabetes, chronic renal disease (including nephrotic syndrome, chronic glomerulonephritis, chronic pyelonephritis, renal dialysis, and renal transplant), chronic kidney disease (stage 3,4, and 5), systolic BP variability, diagnosis of migraine, corticosteroid use, systemic lupus

					erythematosus, antipsychotic use (including amisulpride, aripiprazole, clozapine, lurasidone, olanzapine, paliperidone, quetiapine, risperidone, sertindole, or zotepine), severe mental illness (including psychosis, schizophrenia, or bipolar affective disease), Diagnosis of HIV or AIDS, Diagnosis of erectile dysfunction or treatment for erectile dysfunction (BNF chapter 7.4.5 including alprostadil, phosphodiesterase type 5 inhibitors, papaverine, or phentolamine)
PROCAM	Germany	Assmann., (2002)[6]	Acute coronary events	PROCAM Study	Age, history of diabetes, smoking, family history of MI, LDL-C, HDL-C cholesterol, TG, SBP
Chien	Taiwan	Chien et al., (2012)[7]	Coronary artery disease	Chin-San Community Cardiovascular Cohort Study	Age, sex, BMI, Systolic BP, TC, HDL, LDL-C
Friedland	USA	Friedland et al., (2009)[8]	Incident CVD events	Patient records from the Medical College of Wisconsin and Froedtert Hospital	Age, smoking, hyperlipidemia, diabetes, hypertension, and audiometric patterns (strial, mid-sloping, low-sloping, high-sloping)
ASSIGN Score	UK (Scotland)	Woodward et al., (2007)[9]	Incident CVD events	Scottish Heart Health Extended Cohort Study	Age, TC, HDL-C, systolic BP, diabetes, smoking social deprivation (SIMDSC10), family history of heart disease or stroke

Asia Pacific Cohort Studies Collaboration	Asia	Barzi et al., (2007)[10]	8-year risk of CVD	Asia Pacific Cohort Studies, Chinese cohorts, and Framingham Study	Age, sex, TC, systolic BP, smoking
Chambless	USA	Chambless et al., (2003)[11]	Incident CHD events	ARIC Study	Age, sex, race, TC, HDL-C, systolic BP, antihypertensive medication, smoking, diabetes, IMT, PAD
Keys	US and Europe	Keys et al., (1972)[12]	Incident CHD Events	International Cooperative Study on the Epidemiology of Cardiovascular Disease (Middle-aged men)	Age, systolic BP, TC, smoking, and BMI
The HEART Score	The Netherlands	Brady et al., (2018)[13]	Adverse outcomes from acute coronary syndrome such as myocardial infarction	Community hospital in the Netherlands	Age, ECG, initial troponin, history, risk factors (currently treated diabetes mellitus, current or recent (<one month) smoker, diagnosed hypertension, diagnosed hypercholesterolaemia, family history of coronary artery disease and obesity)
Korean Risk Prediction Model	Republic of Korea	Jung et al., (2015)[14]	ASCVD events	The Korean Heart Study	Age, sex, HDL-C, treated systolic BP, untreated systolic BP, smoking, diabetes

[a]Finland, Russia, Norway, UK, Denmark, Sweden, Belgium, Germany, Italy, France, Spain,

Abbreviations: CVD, cardiovascular disease; ASCVD, atherosclerotic cardiovascular disease; ACC/AHA, American College of Cardiology; ESC SCORE, European Society of Cardiology Systematic Coronary Risk Evaluation; PROCAM, Prospective Cardiovascular Münster; CVD, cardiovascular disease; IHD, ischemic heart disease; CHD, coronary heart disease; MI, myocardial infarction; TC, total cholesterol; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglyceride; BP, blood pressure; BMI, body mass index; ECG, electrocardiogram; ARIC Atherosclerosis Risk in Communities; CARDIA, coronary artery risk development in young adults; ASSIGN, assessing cardiovascular risk using SIGN guidelines to assign potential patients to preventive treatment); SIGN, Scottish Intercollegiate Guidelines Network; SIMDSC10, Scottish Index of Multiple Deprivation score divided by 10; IMT, intima-media thickness; PAD, peripheral artery disease

After abstracting information on the variables included in the conventional CVD risk assessment tools, the comprehensive list of the variables was summarized by the following data categories: sociodemographic factors, clinical laboratory test and measurement, lifestyle behavior, family history, underlying medical conditions, and medication. Subsequently, each variable was assessed for availability in the NHIS-NSC linked to the environmental exposure data (Table 3). Since the NHIS-NSC is based on healthcare claims database and were not linked to electronic health records, patient-level data on electrocardiogram, intima-media thickness, and audiometric patterns were not available. Also, none of the conventional CVD risk assessment tools used environmental exposure data or applied neural network for survival analysis. In addition, the NHIS-NSC is limited to a single ethnic group (South Koreans) and thus data on diverse ethnicity was inherently absent. Due to the changes in the reporting standards of the clinical laboratory tests along with the national health screening questionnaires, information on lipid profiles (HDL-cholesterol, LDL-cholesterol, and triglyceride) are available from 2009 in the NHIS-NSC.

Table 4. List of variables and assessment of their availability in the National Health Insurance Service-National Sample Cohort (NHIS-NSC) abstracted from the 13 conventional CVD risk assessment tools identified from the literature search

Variables used in the conventional CVD risk assessment tools[a] (n=13)	**Availability in the NHIS-NSC**
Sociodemographic factors	
Age	O
Sex	O
Social deprivation	X
Ethnicity (race)	X
Clinical laboratory test and measurement	
Systolic blood pressure	O
Diastolic	O

blood pressure	
Total cholesterol	O
HDL-cholesterol	△
LDL-cholesterol	△
Triglyceride	△
Body mass index	O
Hypertension	O
Type 2 diabetes	O
Hyperlipidemia	O
Electrocardiogram	X
Intima-media thickness	X
Audiometric patterns	X
Lifestyle behavior	
Cigarette smoking	O
Family history	
Family history of CVD	O
Underlying medical conditions	
Atrial fibrillation	O
Peripheral artery disease	O
Chronic kidney disease	O
Diagnosis of migraine	O
Systemic lupus erythematosus	O
Severe mental illness	O
Rheumatoid arthritis	
Diagnosis of HIV or AIDS	X
Erectile dysfunction	X
Medication	
Aspirin	O
Statin	O
Antihypertensive medication	O
Antipsychotic use	O
Corticosteroid use	O

NOTE: O: Available, △, partially available (health screening data from 2009), X: Not available

[a]Conventional CVD risk assessment tools are as follows: Framingham risk score, ASCVD risk , ESC SCORE, Q-risk3, PROCAM, Chien, Friedland, ASSIGN Score, Asia Pacific Cohort Studies Collaboration, Chambless, Keys, The Heart Score, and the Korean Risk Prediction Model

Abbreviations: CVD, cardiovascular disease; HDL-cholesterol, high-density lipoprotein cholesterol ; LDL-cholesterol, low-density lipoprotein cholesterol; HIV, human immunodeficiency virus; AIDS, acquired immune deficiency syndrome

After screening for previous meta-analyses on the relevant topics and screening published literature using search terms for identifying comprehensive CVD risk factors in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[37], initial search on medical literature database (PubMed and Embase) resulted in a total of 95,875 records (PubMed 31,493 and Embase 64,382) along with 5 records from other sources were included. After removing duplicates, 45,782 articles were remaining and the titles and abstracts of these articles were screened for relevance. After this screening process, 45,396 articles were excluded and 386 articles were determined to be eligible for full-text review. Finally, 339 articles were removed after full-text review and 47 articles were selected for review of comprehensive CVD risk factors (Figure 8).

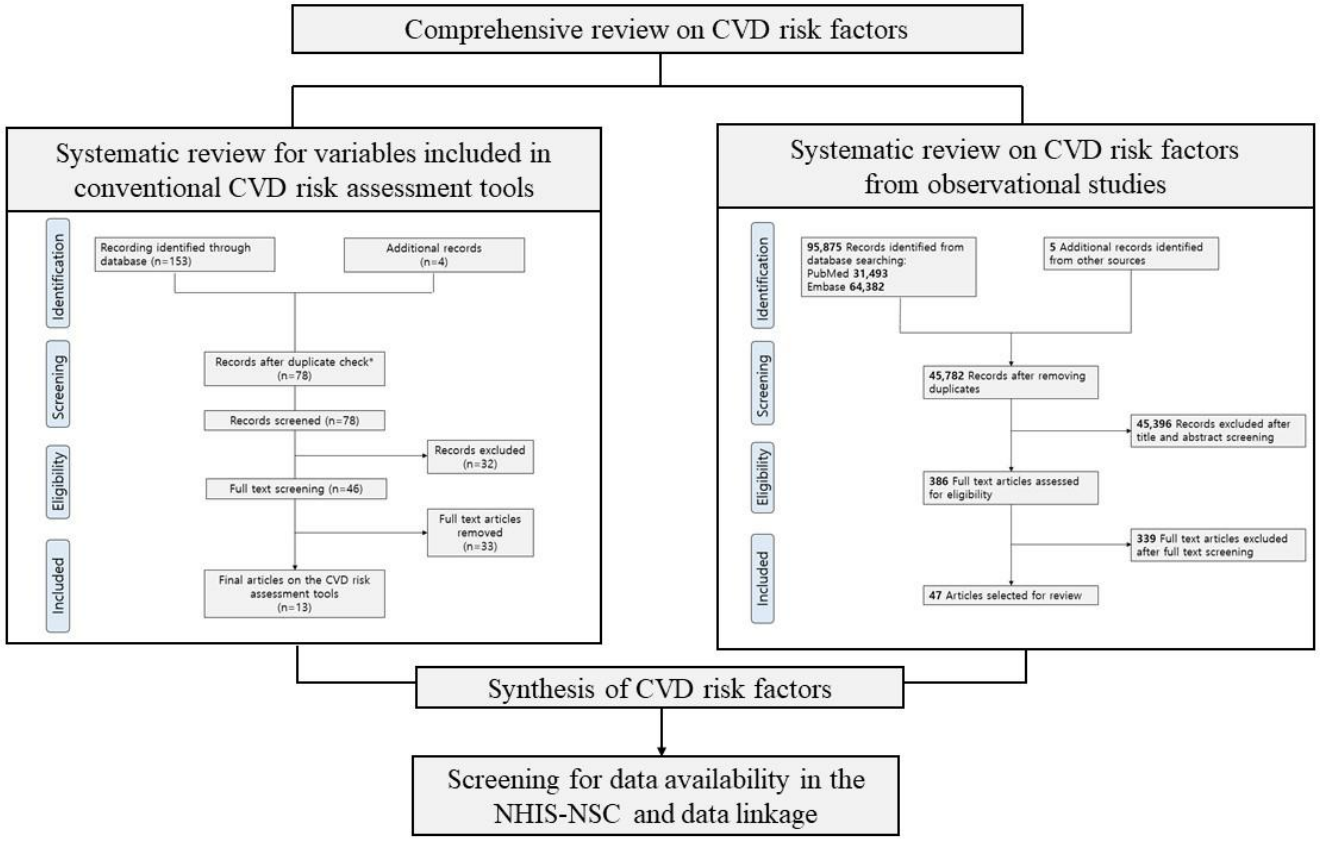

Figure 8. Flow diagram for a comprehensive review on CVD risk factors identified from conventional CVD risk assessment tools and observational studies in PubMed and Embase database

Comprehensive list of CVD risk factors from the literature search were limited to the factors that were not included in the conventional CVD risk assessment models. Risk factors identified from the literature search were derived from 47 observational studies and they were summarized by factors, representative meta-analysis, reference article, data source, sample size, and outcome used in the study and were further reviewed for relevant data categories and availability in the NHIS-NSC (Table 4). Due to the limited sources of data (administrative, national health screening, medical/dental claims) collected for constructing the NHIS-NSC, most of the patient-level data on circulatory system, physical fitness, biomarkers, and dietary factors were not available. Environmental exposure data on particulate matter and urban green space could be merged using the residential area code of the NHIS enrollees as the common key for according to the previous studies with the NHIS-NSC. Other environmental factors such as arsenic exposure and household fuel use could not be obtained and linked to the NHIS-NSC. However, data on medical conditions, dental disease, and medication use that were identified as CVD risk factors could be collected from the claims and prescription records in the NHIS-NSC.

Table 5. Comprehensive list of CVD risk factors identified from the observational studies in the systematic review

Factors associated with CVD	No. of studies screened	Representative meta-analysis	Representative study	Data Source of the Representative study	Sample size of the representative study	Outcome	Availability in the NHIS-NSC[a]
ABI	42	Ankle Brachial Index Collaboration (2008)[16]	Alzamora et al., (2013)[38]	Primary Health Centers in Barcelona, Spain	3,786	Cerebrovascular events	X
hsCRP	62	Emerging Risk Factors Collaboration (2010)[39]	Blaha et al., (2011)[40]	MESA JUPITER	950	CVD	X
CAC	27	Pletcher et al., (2004)[41]	Blaha et al., (2011)[40]	MESA JUPITER	950	CVD	X
Apo B	39	Sniderman et al., (2011)[42]	Hwang et al., (2017)[43]	TNT and IDEAL	10,001 (TNT) 8,888 (IDEL)	MCVE	X
GGT	28	Du et al., (2013)[44]	Yang et al., (2018)[45]	NHIS-NSC	456,100	Stroke	O
Physical inactivity (MVPA)	79	Wahid et al., (2016)[46]	Kim et al., (2019)[47]	NHIS (nationwide cohort for elderly)	1,119,925	CVD	△
Short sleep duration	34	Cappuccio et al., (2011)[48]	Chandola et al., (2010)[49]	The Whitehall II cohort	10,308	CHD	X
Shift-work	51	Vyas et al., (2012)[50]	Hublin et al., (2010)[51]	The Finnish twin cohort	20,142	CVD deaths and	X

						disability due to CVD	
Cardiorespiratory fitness (peak exercise oxygen consumption)	62	Kodama et al., (2009)[52]	Laukkanen et al., (2004)[53]	KIHD (Finland)	2,361	CVD death	X
Handgrip strength	23	Chainnani., (2016)[54]	Celis-Morales et al., (2018)[55]	UK Biobank	502,293	CVD mortality	X
Push-up exercise capacity	1	N/A	Yang et al., (2019)[56]	Male firefighter cohort in the USA	1,104	CVD	X
Retinal vein occlusion	24	Khan et al., (2013)[57]	Rim et al., (2015)[58]	NHIS-NSC	5,074	Stroke and AMI	O
Retinal artery occlusion	12	Zhou et al., (2016)[59]	Rim et al., (2016)[60]	NHIS-NSC	2,403	Stroke	O
Chronic kidney disease	35	Palmer et al., (2011)[61]	Angelantonio et al., (2010)[62]	The Reykjavik study (Iceland)	16,958	MACE	O
NAFLD	56	Targher et al., (2016)[63]	Zeb et al., (2016)[64]	MESA	6,814	Incident cardiac events	O

Anemia	29	N/A	Zakai et al., (2005)[65]	CHS	1,205	CVD mortality	O
Parkinson's disease	98	Alves et al., (2020)[66]	Huang et al., (2013)[67]	NHI claims database (Taiwan)	2,204	Ischemic stroke	O
Chronic periodontitis	12	Lafon et al., (2014)[68]	Hansen et al., (2016)[69]	The Danish Nationwide Cohort Study	17,691	CVD mortality	O
Dental caries	17	N/A	Park et al., (2019)[70]	NHIS-HEALS	247,696	CVD	O
Red meat	69	Kim et al., (2017)[71]	Larsson et al., (2011)[72]	COSM	40,291	Stroke	X
Processed meat	72	Kim et al., (2017)[71]	Bernstein et al., (2012)[73]	HPFS	84,010	Stroke	X
White meat	53	Kim et al., (2017)[71]	Haring et al., (2015)[74]	ARIC	11,601	Stroke	X
Fish[b]	49	Larsson et al., (2011)[75]	Mozaffarian et al., (2005)[76]	CHS	4,775	Stroke	X
Fried food	36	Gadiraju et al., (2015)[77]	Guallar-Castillón et al., (2012)[78]	Spanish EPIC	40,757	CHD	X
Fruit and vegetable[c]	56	Wang et al., (2014)[79]	Larsson et al., (2013)[80]	SMC and COSM	74,961	Stroke	X
Sugar and artificially sweetened beverages	78	Narain et al., (2016)[81]	Pase et al., (2017)[82]	FHS	2,888	Stroke	X
Coffee	37	Sofi et al., (2007)[83]	Kleemola et al., (2000)[84]	Cohort of eastern Finish men and women	20,179	CHD	X
Milk	48	Guo et al.,	Bergholdt et al.,	Copenhagen General	33,625	IHD	X

		(2017)[85]	(2015)	Population Study			
Egg	31	Geiker et al., (2018)[86]	Qin et al., (2018)[87]	CKB	461,213	CVD, IHD, and MCE	X
Green and roasted teas	52	Bohn et al., (2012)[88]	Tanabe et al., (2008)[89]	Tokamachi–Nakasato cohort in Japan	6,358	Stroke, cerebral infarction and cerebral hemorrhage	X
Nuts	23	Mayhew et al., (2016)[90]	Bao et al., (2013)[91]	NHS and HPFS	76,464	Heart disease and stroke mortality	X
Alcohol (light-to-moderate)	84	Ronksley et al., (2011)[92]	Smyth et al., (2015)[93]	PURE	155,875	MI, Stroke	△
Dietary fiber	62	Threpleton et al.,(2013)[94]	Kokubo et al., (2011)[95]	The Japan Public Health Center-based prospective study	86,387	CVD	X
Folic acid, vitamin B6, and vitamin B12	27	Zhou et al., (2011)[96]	Albert et al., (2008)[97]	WAFACS	5,442	CVD	X
Vitamin C supplement	53	Chen et al., (2013)[98]	Osganian et al., (2003)[99]	NHS	85,118	CHD	X
Dietary and supplemental	93	Bolland et al., (2014)[100]	Messenger et al., (2012)[101]	MrOS	3,904	CVD	X

Vitamin D							
Dietary sodium	60	Mozafaarian et al., (2014)[102]	Cook et al., (2007)[103]	TOPH I and TOPH II	744	MI, Stroke, Coronary revasculari zation, CVD death	X
Dietary and supplemental Calcium	97	Chung et al., (2016)[104]	Hemerijck et al., (2013)[105]	NHANES III linked to the NDI	20,024	CVD death	X
Dietary potassium	82	D'Elia et al., (2011)[106]	Umesawa et al., (2008)[107]	JACC Study for Evaluation of Cancer Risks	58,730	CVD death	X
Omega-3 fatty acids	49	Zhang et al., (2016)[108]	Amiano et al., (2014)[109]	Spanish EPIC	41,091	Coronary events	X
Cadmium exposure	32	Larsson et al., (2016)[110]	Tellez-Plaza et al., (2013)[111]	Cohort of American Indians in the Strong Heart Study	3,348	CVD incidence and mortality	X
Lead exposure	57	Navas-Acien et al., (2007)[112]	Lanphear et al., (2018)[113]	NHANES III linked to the NDI	14,289	CVD mortality	X
Arsenic exposure	61	Moon et al., (2017)[114]	Chen et al., (2011)[115]	HEALS	11,746	CVD mortality	X
Household fuel use	15	N/A	Mitter et al., (2016)[116]	Golestan Cohort Study	50,045	CVD mortality	X
Urban	27	N/A	Seo et al.,	NHIS-NSC	351,409	CVD	O

green space			(2019)[117]				
PM 2.5	36	Fu et al., (2015)[118]	Crouse et al., (2012)[119]	Canadian cohort of nonimmigrants	2,145,400	Nonaccidental and CVD mortality	△ (only 3 cities from 2009)
PM 10	45	Fu et al., (2015)[118]	Arthur-Hvidtfeldt et al., (2019)[120]	The Danish, Diet, Cancer and Health cohort	49,564	CVD mortality	O
Black carbon	12	N/A	Arthur-Hvidtfeldt et al., (2019)[120]	The Danish, Diet, Cancer and Health cohort	49,564	CVD mortality	X
Nitrogen dioxide	47	Mustafic et al., (2012)[121]	Arthur-Hvidtfeldt et al., (2019)[120]	The Danish, Diet, Cancer and Health cohort	49,564	CVD mortality	X

NOTE: O **:**Available; **△:** partially available from 2009 or not well defined prior to 2009 survey; X: Not available;

[a]Including the data source that can be merged into the NHIS-NSC

[b]This study found that broiled or baked fish consumption was associated with higher risk of stroke whereas fried fish or fish sandwich consumption was associated with lower risk of stroke.

[c]In particular, apples, pears, and green leafy vegetables

[d]Cannot be linked due to the limited information on the administrative/location code in the NHIS-NSC

Abbreviations: CVD, cardiovascular disease; MI, myocardial infarction; AMI, acute myocardial infarction; CHD, coronary heart disease; MCE, major coronary event; IHD, ischemic heart disease; MACE, major adverse cardiovascular events; MCVE, major cardiovascular event; ABI, ankle-brachial index; hsCRP, high-sensitivity C-reactive protein; CAC, coronary artery calcium; GGT, Gamma-glutamyl transferase; Apo B; Apolipoprotein B; NHI, the National Health Insurance (Taiwan); HEALS, Health Effects of Arsenic Longitudinal Study; JACC, the Japan Collaborative Cohort; MESA, the Multi-Ethnic Study of Atherosclerosis; JUPITER: Justification for the Use of Statins in Primary Prevention: An Intervention Trial Evaluating Rosuvastatin trial; TNT, Treating to New Targets trial; IDEAL, Incremental Decrease in End points through Aggressive Lipid lowering trial; KIHD, Kuopio Ischaemic Heart Disease Risk Factor Study; CKB, the Chinese Kadoorie Biobank study; SMC, the Swedish Mammography Cohort; COSM, the Cohort of Swedish men; FHS, the Framingham Heart Study; HPFS, the Health Professionals Follow-Up Study; NHS, the Nurses' Health Study; PURE, the Prospective Urban Rural Epidemiological study; MrOS, the Osteoporotic Fractures in Men study; TOPH, trials of hypertension prevention; NHANES, National Health and Nutrition Examination

Survey ; NDI, the National Death Index; ARIC, the Atherosclerosis Risk in Communities Study; CHS, the Cardiovascular Health Study; EPIC, the European Prospective Investigation into Cancer and Nutrition; NHIS-NSC, the National Health Insurance Service-National Sample Cohort; WAFACS, the Women's Antioxidant and Folic Acid Cardiovascular Study; PM, particulate matter

Risk factors for CVD identified from the conventional CVD risk assessment tools and literature search of observational studies were synthesized and categorized into the following data categories: sociodemographic factors, lifestyle behavior, measureable health status and anthropometric measurement, environmental factors, underlying medical conditions, dental/oral health conditions, physical fitness, medication, dietary factors, and biomarkers (Figure 9). After screening for availability in the NHIS-NSC linked to the environmental exposure data, the following categories (variables) were selected for the comprehensive list of CVD risk factors: sociodemographic factors (age, sex, income status [adopted instead of social deprivation]), clinical laboratory test and measurement (hypertension, type 2 diabetes, hyperlipidemia, gamma-glutamyl transferase, [GGT], body mass index [BMI]), lifestyle behavior (cigarette smoking, alcohol consumption, physically inactive), family history (family history of CVD), underlying medical conditions (atrial fibrillation, peripheral artery disease, retinal vein/artery occlusion, anemia, non-alcoholic fatty liver disease [NAFLD], chronic kidney disease [CKD], migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus, rheumatic arthritis, dental health (chronic periodontitis, dental caries), medication (aspirin, corticosteroid, antipsychotics), and environmental exposure (high cumulative exposure to PM 10, low urban green space coverage).

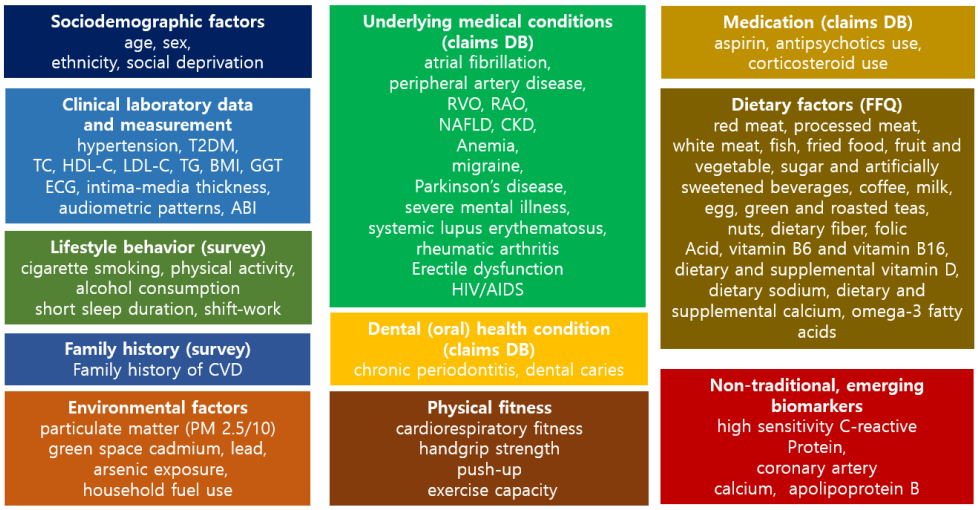

Figure 9. Risk factors for CVD by different data categories identified and synthesized from the comprehensive review on conventional CVD risk assessment tools and observational studies

Abbreviations: CVD, cardiovascular disease; DB, database; FFQ, food frequency questionnaire

Based on the established evidence from previous studies, the following operational definitions were used to identify and abstract information on the relevant variables in the NHIS-NSC linked to the environmental exposure data (Table 5).

Table 6. Operational definitions for CVD risk factors available in the NHS-NSC linked to environmental exposure data listed by categories and reference articles

Variables in the NHIS-NSC linked to environmental exposure	Operational definition	Reference
Sociodemographic factors		
Age	<65 (middle-aged), ≥65 (elderly)	-
Sex	Male, Female	-
Low income	Lowest quartiles in the insurance premium	-
Clinical laboratory test and measurement		
Hypertension	Systolic blood pressure≥140 mmHg or diastolic blood pressure ≥90 mmHg or with antihypertensive prescription (more than 30 days)	Korea Hypertension Fact Sheet (2018)
Type 2 diabetes	Fasting serum glucose≥126 mg/dL or with antidiabetic drug prescription (more than 30 days)	Korea Diabetes Fact Sheet (2015)
Hyperlipidemia	Total cholesterol≥240 mg/dL or with statin prescription (more than 30 days)	Jeong et al., (2018)[122]
GGT	Treated as a continuous variable in log scale	Yang et al., (2018)[45]
Body mass index	Treated as a continuous variable	Choi et al., (2018)[123]
Lifestyle behavior		
Cigarette smoking	Answered "smoker" to the current status of cigarette smoking in the self-reported questionnaire in the self-reported questionnaire (national health screening)	Kim et al., (2018)[30]
Alcohol consumption	Defined as at least light-to-moderate drinker in the self-reported questionnaire (national health screening)	Choi et al., (2019)[124]
Physically inactive	Answered "none or not all per week" to walking, moderate, and vigorous physical activity in the self-	Kim et al., (2019)[29]

	reported questionnaire (national health screening)	
Family history		
Family history of CVD	Answered "yes" to the family history of heart disease or stroke in the self-reported questionnaire (national health screening)	-
Underlying medical conditions		
Atrial fibrillation	ICD-10: I48.0-I48.4, I48.9 with at least 2 outpatient visits or hospitalization	Choi et al., (2020)[125]
Peripheral artery disease	ICD-10: I70, I70.0, I70.2, I70.8, I70.9, I79.2 at least 1 inpatient/outpatient care	Oh et al., (2017)[126]
Retinal vein occlusion	ICD-10: H34.8 with inpatient/outpatient care	Rim et al., (2015)[58]
Retinal artery occlusion	ICD-10: H34.1/H34.2 with inpatient/outpatient care	Rim et al., (2016)[60]
Anemia	Hemoglobin <13.0 and <12.0 g/dL in men and women	Lee et al., (2018)[127]
NAFLD	ICD-10: K76 with at least 2 inpatient/outpatient care	Lee et al., (2017)[128]
CKD	ICD-10: N18.3, N18.4, N18.5 with at least 1 day of hospitalization or 3 days of outpatient visits	Kim et al., (2017)[129]
Migraine	ICD-10: G43 with inpatient/outpatient care	Min et al., (2019)[130]
Parkinson's disease	ICD-10: G20 with at least 2 outpatient visits or hospitalization	Choi et al. (2019)[131]
Severe mental illness	ICD-10: F31, F32/F33, F20 with hospital admission or outpatient visit	Ko et al., (2019)[132]
Systemic lupus erythematosus	ICD-10: M32.9 with hospital admission ,drug prescription, and lab test[a]	Bae et al., (2019)[133]
Rheumatic arthritis	ICD-10: M05 with disease modifying anti-rheumatic drugs	Choi et al., (2019)[134]
Dental (oral) health		
Chronic periodontitis	ICD-10: K05.3 with relevant treatment records[b]	Choi et al., (2019)[131]

Dental caries	incipient/moderate, advanced/severe with at least 2 outpatient visits[c]	Kim et al., (2019)[135]
Medication		
Aspirin	≥30 days of prescription	Hwang et al., (2018)[136]
Corticosteroid	≥30 days of prescription	Rim et al., (2018)[137]
Antipsychotics	≥30 days of prescription[d]	Leucht et al., (2009)[138]
Environmental exposure[e]		
High cumulative exposure to PM 10	Highest quartile of cumulative PM 10 exposure	Choi et al., (2020)[139]
Low urban green space coverage	Lowest quartiles of urban green space	Seo et al., (2019)[117]

[a]hydroxychloroquine, immunosuppressants, and steroids (drugs), anti-dsDNA antibody test and complement test (laboratory test)
[b]Subgingival curettage, periodontal flap operation, gingivectomy, and odontectomy
[c]Dental caries limited to enamel (ICD-10 code: K02.0), dental caries of dentin (ICD-10 code: K02.1), dental carries of cementum, arrested dental caries (ICD-10 code: K02.3), other dental caries (ICD-10 code: K02.8), and unspecified dental caries (ICD-10 code: K02.9) were classified as incipient/moderate dental caries, and those with irreversible pulpitis (ICD-10 code: K04.0), necrosis of pulp (ICD-10 code: K04.1), and periapical abscess with sinus (ICD-10 code: K04.6) were classified as advanced/severe stage dental caries
[d]Includes the following 2nd generation antipsychotic drugs: clozapine, olanzapine, quetiapine, paliperidone, risperidone, ziprasidone, zotepine, aripiprazole
[e]Cumulative exposure to particulate matter (PM 10) was computed by taking the annual average of PM 10. High cumulative exposure to PM 10 indicates highest quartile.
Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile. Environmental data were merged with residential area code in the NHIS-NSC with AirKorea database (PM 10) and Ministry of Land, Infrastructure and Transport database (urban green space)

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; SBP, systolic blood pressure; DBP, diastolic blood pressure; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease; Alcoholism guideline; ICD-10, International Classification of Disease, 10th revision; PM, particulate matter

2. Magnitude of association between selected risk factors with incident cardiovascular disease

Prior to applying statistics-based variable selection methods, multiple collinearity was checked with VIF with a cut-off point set to VIF>5 (VIF greater than 5 indicating evidence of multiple collinearity). After computing VIF among the variables identified and synthesized from the comprehensive review on conventional CVD risk assessment tools and observational studies, no evidence of multiple collinearity was found (Table 6).

Table 7. Multicolinearity test for independent variables measured by the variance inflation factor for the variables included in the final analytic cohort derived from the NHIS-NSC linked to environmental exposure data

Variables in the final analytic cohort derived from NHIS-NSC	Variance Inflation Factor
Sociodemographic factors	
Age	1.10526
Sex	2.10542
Income status	1.02221
Clinical laboratory test and measurement	
Hypertension[a]	1.06755
Type 2 diabetes[b]	1.06702
Hyperlipidemia[c]	1.10732
GGT	1.14669
Body mass index	1.08698
Lifestyle behavior	
Cigarette smoking	1.96722
Alcohol consumption	1.36205
Physically inactive	1.04013
Family history	
Family history of CVD	1.00671
Underlying medical conditions	
Atrial fibrillation	1.00525
Peripheral artery disease	1.04940
Retinal vein occlusion	1.01199
Retinal artery occlusion	1.00911
Anemia	1.06136
NAFLD	1.01395
CKD	1.00356

Migraine	1.02595
Parkinson's disease	1.00482
Severe mental illness	1.04745
Systemic lupus erythematosus	1.01253
Rheumatic arthritis	1.03950
Dental (oral) health	
Chronic periodontitis	1.04449
Dental caries	1.03709
Medication	
Aspirin	1.15719
Corticosteroid	1.03847
Antipsychotics	1.03689
Environmental exposure[d]	
High cumulative exposure to PM 10	1.00800
Low urban green space coverage	1.00728

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as fasting serum glucose ≥ 126 mg/dL or with antidiabetic drug prescription (more than 30 days)
[c]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)
[d]Cumulative exposure to particulate matter (PM 10) was computed by taking the annual average of PM 10. High cumulative exposure to PM 10 indicates highest quartile.
Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile.

Abbreviations: NHIS-NSC; National Health Insurance Service-National Sample Cohort; GGT, gamma-glutamyl transpeptidase; CVD, cardiovascular disease; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

The global goodness-of-fit test with Schoenfeld residual using all of the variables included as the comprehensive list of CVD risk factors used in the NHIS-NSC linked to environmental exposure data resulted in p=0.316, which indicates that the proportionality assumption of the Cox proportional hazards regression was not violated. In addition, log-log plot for age group showed that the survival probability was relatively parallel across the analysis time (Figure 10).

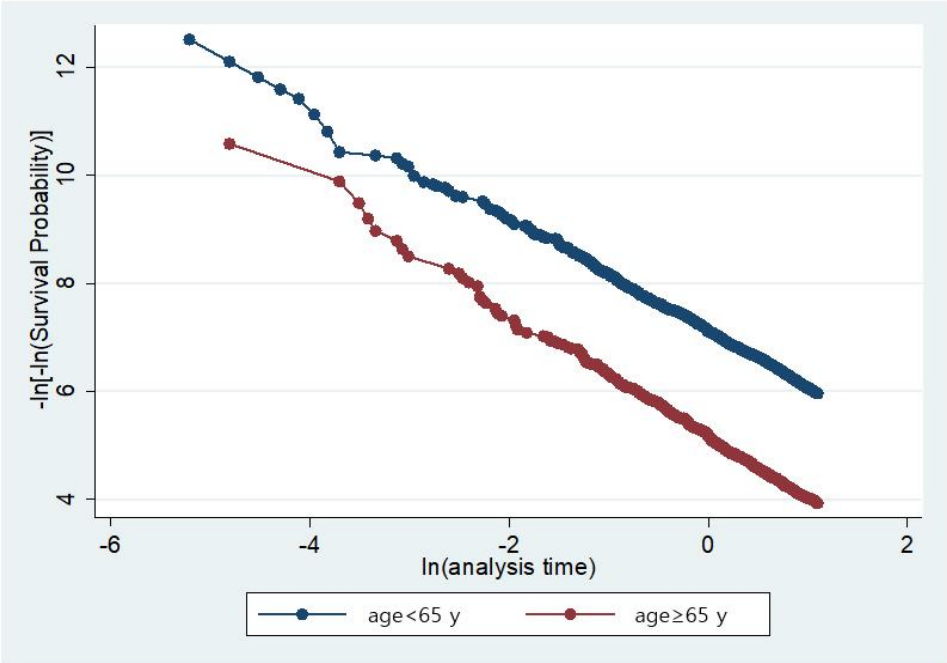

Figure 10. Log-log survival plot for age in the NHIS-NSC linked to environmental exposure data

Abbreviation: NHIS-NSC, National Health Insurance-National Sample Cohort

The descriptive statistics of the final study population derived from the NHIS-NSC linked to environmental exposure data used for statistics-based variable selection is shown in Table 7.

Table 8. Descriptive statistics of the final study population derived from the NHIS-NSC linked to the data on environmental exposure

Category	N (%) or mean (±SD)
Sociodemographic factors	
Age	
<65 years	118,768 (86.5)
≥65 years	18,481 (13.5)
Sex	
Male	66,905 (51.2)
Female	70,344 (48.8)
Income status	
High income	49,995 (36.4)
Low income	87,254 (63.6)
Clinical laboratory test and measurement	

Hypertension[a]	21,975 (16.0)
Type 2 diabetes[b]	9,620 (7.0)
Hyperlipidemia[c]	32,880 (23.9)
GGT, U/L, median (IQR)	23 (16-39)
Body mass index, kg/m^2	23.8±3.00
Lifestyle behavior	
Cigarette smoking	51,668 (37.7)
Alcohol consumption	61,983 (45.2)
Physically inactive	92,672 (67.5)
Family history	
Family history of CVD	14,219 (10.4)
Underlying medical conditions	
Atrial fibrillation	15 (0.01)
Peripheral artery disease	3,056 (2.2)
Retinal vein occlusion	548 (0.4)
Retinal artery occlusion	55 (0.04)
Anemia	16,545 (12.1)
NAFLD	8,392 (6.1)
CKD	138 (0.1)
Migraine	13,508
Parkinson's disease	131 (0.1)
Severe mental illness	4,717 (3.4)
Systemic lupus erythematosus	235 (0.2)
Rheumatic arthritis	612 (0.5)
Dental (oral) health	
Chronic periodontitis	63,382 (46.2)
Dental caries	51,795 (37.7)
No. of comorbid conditions[d]	
0	37,161 (27.0)
1	51,628 (37.6)
2	36,182 (26.4)
≥3	12,278 (9.0)
Medication	
Aspirin	13,984 (10.2)
Corticosteroid	3,451 (2.5)
Antipsychotics	293 (0.2)
Environmental exposure[e]	
High cumulative exposure to PM 10	40,473 (29.5)
Low urban green space coverage	32,125 (23.4)

NOTE: Data above presented as n (%) or mean ±SD, unless otherwise specified

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)

[b]Defined as fasting serum glucose ≥ 126 mg/dL or with antidiabetic drug prescription (more than 30 days)

[c]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)

[d]Cumulative number of underlying conditions in each cateogry

[e]Cumulative exposure to particulate matter (PM 10) was computed by taking the annual average of

PM 10. High cumulative exposure to PM 10 indicates highest quartile.

Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile. Environmental data were merged with residential area code in the NHIS-NSC with AirKorea database (PM 10) and Ministry of Land, Infrastructure and Transport database (urban green space)

Abbreviations: ; NHIS-NSC; National Health Insurance Service-National Sample Cohort; SD, standard deviation; GGT, gamma-glutamyl transpeptidase; CVD, cardiovascular disease; N/C, non-calculable; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

After adjusting for all of the variables included as the comprehensive risk factors for CVD, CKD (HR=2.897; 95% CI: 1.538-5.391), more than 65 years of age (HR=2.863; 95% CI: 2.598-3.155 vs. less than 65 years of age), and Parkinson's disease (HR=2.831; 95% CI: 1.511-5.304) were some of the most notable risk factors associated with incident CVD that showed statistical significance. Due to the relatively small number of events for atrial fibrillation, the association between atrial fibrillation and incident CVD could not be calculated (Table 8).

Table 9. Multivariable analysis of all variables for the association of CVD risk factors and incident CVD in the NHIS-NSC linked to the data on environmental exposure

Category	HR (95% CI)	*p*-value
Sociodemographic factors		
Age (≥65 vs. <65)	2.863 (2.598-3.155)	<.0001
Male (vs. female)	1.348 (1.191-1.525)	<.0001
Low income (vs. high income)	1.122 (1.028-1.225)	0.01
Clinical laboratory test and measurement		
Hypertension[a] (yes vs. no)	1.460 (1.323-1.612)	<.0001
Type 2 diabetes[b] (yes vs. no)	1.499 (1.320-1.702)	<.0001
Hyperlipidemia[c] (yes vs. no)	1.219 (1.108-1.341)	<.0001
GGT (per unit increase in log scale)	1.115 (1.044-1.191)	0.0012
Body mass index (per unit increase)	1.027 (1.012-1.041)	0.0003
Lifestyle behavior		
Cigarette smoking (yes vs. no)	1.458 (1.295-1.641)	<.0001
Alcohol consumption (yes vs. no)	0.816 (0.738-0.903)	<.0001
Physically inactive (yes vs. no)	1.296 (1.176-1.428)	<.0001
Family history		
Family history of CVD (yes vs. no)	1.254 (1.098-1.433)	0.0009
Underlying medical conditions		
Atrial fibrillation (yes vs. no)	N/C	-

Peripheral artery disease (yes vs. no)	1.029 (0.821-1.290)	0.805
Retinal vein occlusion (yes vs. no)	1.283 (0.793-2.075)	0.3093
Retinal artery occlusion (yes vs. no)	0.645 (0.090-4.623)	0.6627
Anemia (yes vs. no)	1.051 (0.918-1.202)	0.4724
NAFLD (yes vs. no)	1.227 (1.054-1.427)	0.0081
CKD (yes vs. no)	2.879 (1.538-5.391)	0.001
Migraine (yes vs. no)	1.304 (1.141-1.489)	<.0001
Parkinson's disease (yes vs. no)	2.831 (1.511-5.304)	0.0012
Severe mental illness (yes vs. no)	1.162 (0.938-1.439)	0.1687
Systemic lupus erythematosus (yes vs. no)	1.679 (0.745-3.780)	0.2112
Rheumatic arthritis (yes vs. no)	1.742 (1.078-2.814)	0.0234
Dental (oral) health		
Chronic periodontitis (yes vs. no)	1.097 (1.006-1.197)	0.0372
Dental caries (yes vs. no)	1.032 (0.945-1.128)	0.4783
No. of comorbid conditions[d]		
1 (vs. 0)	1.075 (0.957-1.207)	0.2228
2 (vs. 0)	1.225 (1.086-1.381)	0.0009
≥3 (vs. 0)	1.438 (1.234-1.674)	<.0001
Medication		
Aspirin (yes vs. no)	1.396 (1.245-1.564)	<.0001
Corticosteroid (yes vs. no)	1.338 (1.086-1.647)	0.0061
Antipsychotics (ye vs. no)	1.127 (0.527-2.411)	0.7581
Environmental exposure[e]		
High cumulative exposure to PM 10 (vs. low)	1.080 (0.982-1.187)	0.1123
Low urban green space coverage (vs. high)	1.167 (1.058-1.287)	0.002

NOTE: HR (95% CI) presented above were adjusted for all other variables presented in the table except for the index score of comorbid conditions (underlying conditions omitted in the adjustment due to collinearity).

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as fasting serum glucose ≥ 126 mg/dL or with antidiabetic drug prescription (more than 30 days)
[c]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)
[d]Cumulative number of all underlying health conditions.
[e]Cumulative exposure to particulate matter (PM 10) was computed by taking the annual average of PM 10. High cumulative exposure to PM 10 indicates highest quartile.
Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile. Environmental data were merged with residential area code in the NHIS-NSC with AirKorea database (PM 10) and Ministry of Land, Infrastructure and Transport database (urban green space)

Abbreviations: CVD, cardiovascular disease; NHIS-NSC; National Health Insurance Service-National Sample Cohort; GGT, gamma-glutamyl transpeptidase; N/C, non-calculable; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

The categories (variables) selected after removing those that did not show statistical significance were as follows: sociodemographic factors (age, male, low income), clinical laboratory test and measurement (hypertension, type 2 diabetes, hyperlipidemia, GGT, and body mass index), lifestyle behavior (cigarette smoking, alcohol consumption, physically inactive), family history (family history of CVD), underlying medical conditions (NAFLD, CKD, migraine, Parkinson's disease, rheumatic arthritis), dental health (chronic periodontitis), medication (aspirin, corticosteroid), and environmental exposure (low urban green space coverage). The strength of the association of statistically significant variables with incident CVD is shown in Table 9.

Table 10. Multivariable analysis of statistically significant variables for the association of CVD risk factors and incident CVD in the NHIS-NSC linked to the data on environmental exposure

Category	HR (95% CI)	*p*-value
Sociodemographic factors		
Age (≥65 vs. <65)	2.885 (2.619-3.177)	<.0001
Male (vs. female)	1.334 (1.180-1.508)	<.0001
Low income (vs. high income)	1.120 (1.026-1.223)	0.0114
Clinical laboratory test and measurement		
Hypertension[a] (yes vs. no)	1.458 (1.321-1.609)	<.0001
Type 2 diabetes[b] (yes vs. no)	1.498 (1.319-1.700)	<.0001
Hyperlipidemia[c] (yes vs. no)	1.220 (1.109-1.342)	<.0001
GGT (per unit increase in log scale)	1.001 (1.000-1.002)	0.0015
Body mass index (per unit increase)	1.026 (1.012-1.040)	<.0001
Lifestyle behavior		
Cigarette smoking (yes vs. no)	1.457 (1.294-1.641)	<.0001
Alcohol consumption (yes vs. no)	0.813 (0.734-0.899)	<.0001
Physically inactive (yes vs. no)	1.292 (1.172-1.424)	<.0001
Family history		
Family history of CVD (yes vs. no)	1.257 (1.100-1.436)	0.0008
Underlying medical conditions		
NAFLD (yes vs. no)	1.235 (1.062-1.436)	0.0062
CKD (yes vs. no)	2.995 (1.607-5.582)	0.0006
Migraine (yes vs. no)	1.311 (1.148-1.497)	<.0001
Parkinson's disease (yes vs. no)	2.949 (1.582-5.498)	0.0007
Rheumatic arthritis (yes vs. no)	1.820 (1.132-2.926)	0.0134
Dental (oral) health		

Chronic periodontitis (yes vs. no)	1.105 (1.014-1.204)	0.0225
Medication		
Aspirin (yes vs. no)	1.403 (1.253-1.569)	<.0001
Corticosteroid (yes vs. no)	1.344 (1.092-1.655)	0.0053
Environmental exposure[e]		
Low urban green space coverage (vs. high)	1.156 (1.049-1.275)	0.0034

NOTE: HR (95% CI) presented above were adjusted for all other variables presented in the table except for the index score of comorbid conditions (underlying conditions omitted in the adjustment due to collinearity). Statistical significance was set to $p<0.05$ when selecting the variables.

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as fasting serum glucose ≥ 126 mg/dL or with antidiabetic drug prescription (more than 30 days)
[c]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)
[d]Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile. Environmental data were merged with residential area code in the NHIS-NSC with Ministry of Land, Infrastructure and Transport database (urban green space)

Abbreviations: CVD, cardiovascular disease; NHIS-NSC; National Health Insurance Service-National Sample Cohort; GGT, gamma-glutamyl transpeptidase; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

The subset of best (minimum) AIC was found in step 23, which includes the variables from sociodemographic factors (age, male, low income) to underlying medication conditions (atrial fibrillation, peripheral artery disease, retinal vein/artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, and systemic lupus erythematosus (Figure 11).

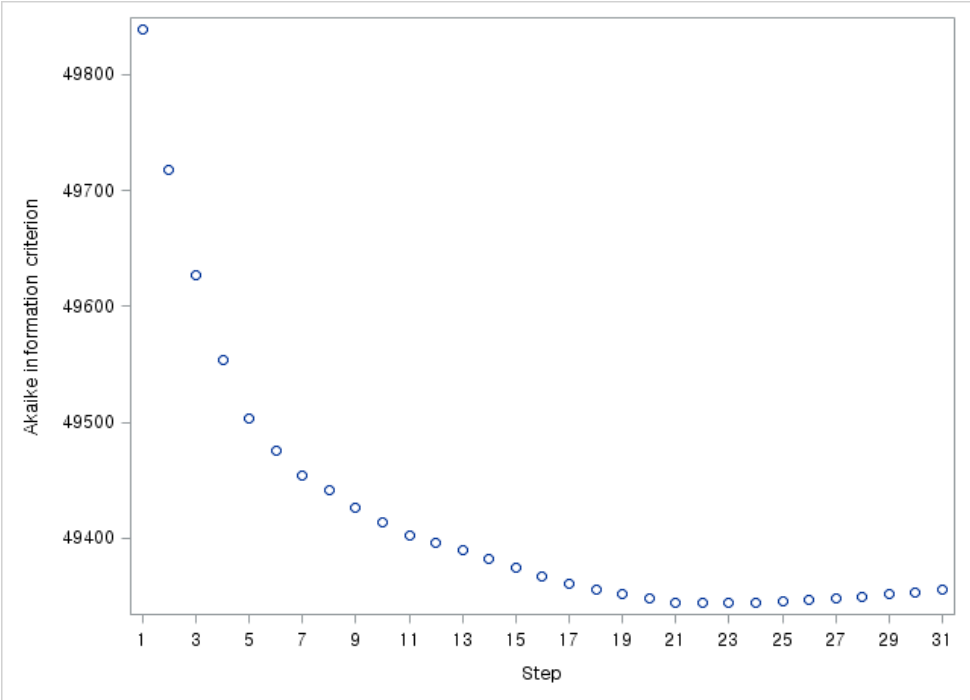

Figure 11. Step vs. Akaike information criterion plot in the stepwise selection fashion for selecting the subset of the variables with the best (minimum) Akaike information criterion

The magnitude of the associations of subset of variables with best (minimum) AIC with incident CVD is shown in Table 10.

Table 11. Multivariable analysis of the variable subset with best (minimum) Akaike Information Criteria for the association of CVD risk factors and incident CVD in the NHIS-NSC linked to the data on environmental exposure

Category	HR (95% CI)	*p*-value
Sociodemographic factors		
Age (≥65 vs. <65)	3.096 (2.818-3.401)	<.0001
Male (vs. female)	1.355 (1.198-1.533)	<.0001
Low income (vs. high income)	1.115 (1.021-1.217)	0.0153
Clinical laboratory test and measurement		
Hypertension[a] (yes vs. no)	1.483 (1.344-1.637)	<.0001
Type 2 diabetes[b] (yes vs. no)	1.567 (1.382-1.777)	<.0001
Hyperlipidemia[c] (yes vs. no)	1.280 (1.166-1.406)	<.0001
GGT (per unit increase in log scale)	1.001 (1.000-1.002)	0.0022
Body mass index (per unit increase)	1.030 (1.016-1.045)	<.0001
Lifestyle behavior		
Cigarette smoking (yes vs. no)	1.456 (1.293-1.639)	<.0001
Alcohol consumption (yes vs. no)	0.811 (0.733-0.897)	<.0001
Physically inactive (yes vs. no)	1.289 (1.170-1.421)	<.0001

Family history		
Family history of CVD (yes vs. no)	1.266 (1.108-1.446)	0.0005
Underlying medical conditions		
Atrial fibrillation (yes vs. no)	N/C	-
Peripheral artery disease (yes vs. no)	1.146 (0.917-1.432)	0.231
Retinal vein occlusion (yes vs. no)	1.330 (0.823-2.150)	0.2444
Retinal artery occlusion (yes vs. no)	0.616 (0.086-4.413)	0.6294
Anemia (yes vs. no)	1.062 (0.928-1.216)	0.38
NAFLD (yes vs. no)	1.247 (1.072-1.450)	0.0043
CKD (yes vs. no)	2.958 (1.580-5.538)	0.0007
Migraine (yes vs. no)	1.322 (1.158-1.510)	<.0001
Parkinson's disease (yes vs. no)	2.894 (1.549-5.407)	0.0009
Severe mental illness (yes vs. no)	1.183 (0.959-1.459)	0.1168
Systemic lupus erythematosus (yes vs. no)	2.034 (0.911-4.539)	0.083

NOTE: HR (95% CI) presented above were adjusted for all other variables presented in the table except for the index score of comorbid conditions (underlying conditions omitted in the adjustment due to collinearity).

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as fasting serum glucose ≥ 126 mg/dL or with antidiabetic drug prescription (more than 30 days)
[c]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; GGT, gamma-glutamyl transpeptidase; N/C, non-calculable; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

After fitting the Cox proportional hazards model regularized by elastic net penalty (Figure 12 and 13), the following variables displayed non-zero coefficients and therefore retained in the model: sociodemographic factors (age, low income), clinical laboratory test and measurement (hypertension, hyperlipidemia), lifestyle behavior (alcohol consumption, physically inactive), family history (family history of CVD), underlying medical conditions (atrial fibrillation, peripheral artery disease, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus, rheumatic arthritis), dental health (chronic periodontitis, dental caries), medication (aspirin, corticosteroid, antipsychotics), and environmental exposure (high cumulative exposure to PM 10, low urban green space coverage).

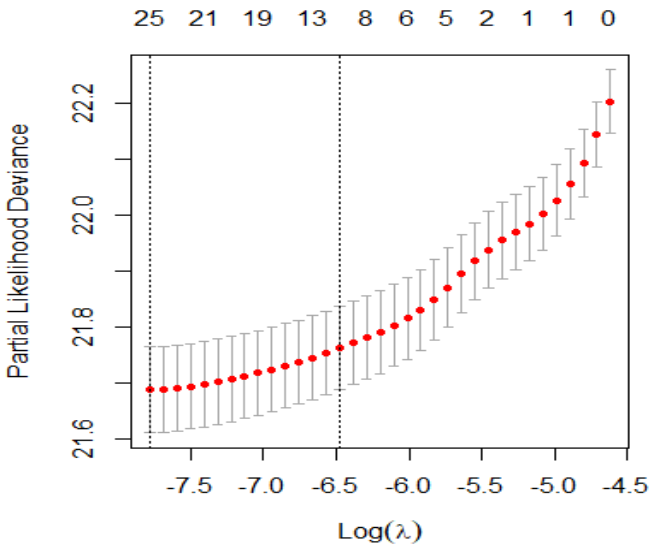

Figure 12. Log lambda vs. partial likelihood deviance plot in the regularized Cox proportional hazards model with elastic net penalty

The left vertical line represents the point with minimum log lambda and the right vertical line represents the point within the 1 standard error of the minimum log lambda

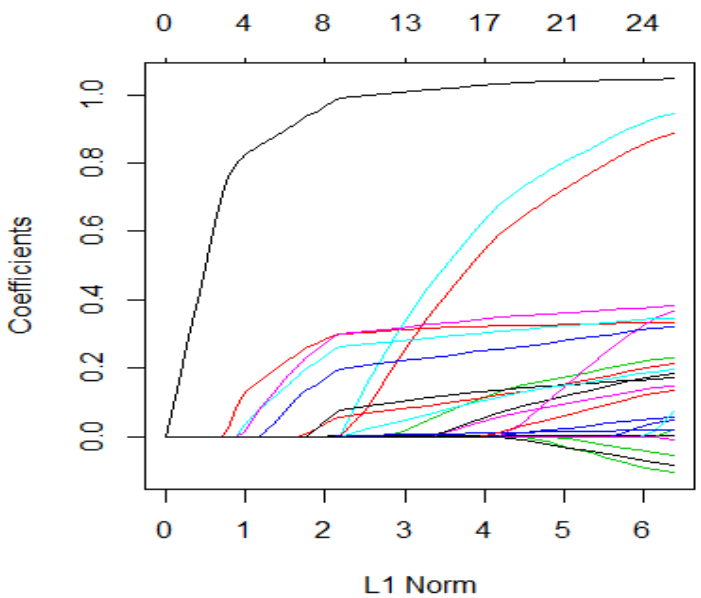

Figure 13. Regularization path for Cox proportional hazards model with elastic net penalty with each line representing the change of coefficient values for each variable

The strength of association between variables selected from the regularized Cox proportional hazards regression with elastic net penalty and incident CVD is shown in Table 11.

Table 12. Multivariable analysis of the variables selected from the Cox regression model regularized by an elastic net penalty for the association of CVD risk factors and incident CVD in the NHIS-NSC linked to the data on environmental exposure

Category	HR (95% CI)	*p*-value
Sociodemographic factors		
Age (≥65 vs. <65)	2.879 (2.613-3.172)	<.0001
Low income (vs. high income)	1.101 (1.009-1.202)	0.0308
Clinical laboratory test and measurement		
Hypertension[a] (yes vs. no)	1.566 (1.420-1.727)	<.0001
Hyperlipidemia[b] (yes vs. no)	1.258 (1.146-1.382)	<.0001
Lifestyle behavior		
Alcohol consumption (yes vs. no)	1.130 (1.033-1.236)	0.0074
Physically inactive (yes vs. no)	1.240 (1.125-1.366)	<.0001
Family history		
Family history of CVD (yes vs. no)	1.238 (1.084-1.414)	0.0017
Underlying medical conditions		
Atrial fibrillation (yes vs. no)	N/C	
Peripheral artery disease (yes vs. no)	1.012 (0.807-1.269)	0.918
Retinal artery occlusion (yes vs. no)	0.703 (0.099-4.994)	0.7243
Anemia (yes vs. no)	0.930 (0.815-1.062)	0.2841
NAFLD (yes vs. no)	1.298 (1.117-1.510)	0.0007
CKD (yes vs. no)	3.392 (1.814-6.342)	0.0001
Migraine (yes vs. no)	1.180 (1.034-1.347)	0.0138
Parkinson's disease (yes vs. no)	2.740 (1.464-5.128)	0.0016
Severe mental illness (yes vs. no)	1.096 (0.885-1.356)	0.4018
Systemic lupus erythematosus (yes vs. no)	1.554 (0.692-3.493)	0.2858
Rheumatic arthritis (yes vs. no)	1.555 (0.964-2.510)	0.0704
Dental (oral) health		
Chronic periodontitis (yes vs. no)	1.151 (1.055-1.256)	0.0015
Dental caries (yes vs. no)	1.023 (0.936-1.117)	0.6191
Medication		
Aspirin (yes vs. no)	1.500 (1.340-1.680)	<.0001
Corticosteroid (yes vs. no)	1.309 (1.063-1.612)	0.0111
Antipsychotics (ye vs. no)	1.197 (0.560-2.559)	0.6423
Environmental exposure[c]		
High cumulative exposure to PM 10 (vs. low)	1.068 (0.972-1.175)	0.1722
Low urban green space coverage (vs. high)	1.172 (1.062-1.292)	0.0015

NOTE: HR (95% CI) presented above were adjusted for all other variables presented in the table except for the index score of comorbid conditions (underlying conditions omitted in the adjustment due to collinearity).

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)
[c]Cumulative exposure to particulate matter (PM 10) was computed by taking the annual average of PM 10. High cumulative exposure to PM 10 indicates highest quartile.
Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile. Environmental data were merged with residential area code in the NHIS-NSC with AirKorea database (PM 10) and Ministry of Land, Infrastructure and Transport database (urban green space)

Abbreviations: CVD, cardiovascular disease; NHIS-NSC; National Health Insurance Service-National Sample Cohort; GGT, gamma-glutamyl transpeptidase; N/C, non-calculable; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

The categories (variables) meeting all of the three statistics-based criteria for variable selection methods (statistical significance, best AIC, and elastic net penalty) were sociodemographic factors (age, low income), clinical laboratory test and measurement (hypertension, hyperlipidemia), lifestyle behavior (alcohol consumption, physically inactive), family history (family history of CVD), and underlying conditions (NAFLD, CKD, migraine, and Parkinson's disease). The association of the variables meeting all of the three statistics-based criteria were similar to the associations found in each criteria (Table 12).

Table 13. Multivariable analysis of the variables meeting the three criteria (statistical significance, best AIC, and elastic net) for the association of CVD risk factors and incident CVD in the NHIS-NSC linked to the data on environmental exposure from particulate matter and urban green space

Category	HR (95% CI)	*p*-value
Sociodemographic factors		
Age (≥65 vs. <65)	3.174 (2.892-3.484)	<.0001
Low income (vs. high income)	1.091 (0.999-1.190)	0.0518
Clinical laboratory test and measurement		
Hypertension[a] (yes vs. no)	1.616 (1.466-1.780)	<.0001
Hyperlipidemia[b] (yes vs. no)	1.359 (1.240-1.489)	<.0001
Lifestyle behavior		
Alcohol consumption (yes vs. no)	1.129 (1.033-1.234)	0.0075
Physically inactive (yes vs. no)	1.229 (1.115-1.354)	<.0001
Family history		
Family history of CVD (yes vs. no)	1.250 (1.095-1.428)	0.001
Underlying medical conditions		
NAFLD (yes vs. no)	1.337 (1.150-1.555)	0.0002
CKD (yes vs. no)	3.499 (1.878-6.518)	<.0001

Migraine (yes vs. no)	1.204 (1.056-1.373)	0.0056
Parkinson's disease (yes vs. no)	2.884 (1.547-5.376)	0.0009

NOTE: HR (95% CI) presented above were adjusted for all other variables presented in the table except for the index score of comorbid conditions (underlying conditions omitted in the adjustment due to collinearity).

[a]Defined as systolic blood pressure≥140 mmHg or diastolic blood pressure ≥90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as fasting serum glucose≥126 mg/dL or with antidiabetic drug prescription (more than 30 days)

Abbreviations: AIC, Akaike Information Criteria; CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; GGT, gamma-glutamyl transpeptidase; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

3. Model performance evaluation

The baseline characteristics of training cohort used for model development and test cohort used for performance evaluation of the DeepSurv and Cox proportional hazards models are shown in Table 13. There was no statistically significant difference between training cohort and test cohort for each of the variable used as input features.

Table 14. Baseline characteristics of training and test cohort derived from the NHIS-NSC linked to the data on environmental exposure from particulate matter and urban green space used for model development and evaluation

Category	Training cohort (*N*=109,799)	Test cohort (*N*=27,450)	*p*-value
Sociodemographic factors			
Age			
<65 years	95,035 (86.5)	23,733 (86.5)	0.981
≥65 years	14,764(13.5)	3,717 (13.5)	
Sex			
Male	53,558 (48.8)	13,347 (48.6)	0.645
Female	56,241 (51.2)	14,103 (51.4)	
Income status			
High income	39,978 (36.4)	10,017 (36.5)	0.802
Low income	69,821 (63.6)	17,433 (63.5)	
Clinical laboratory test and measurement			
Hypertension[a]	17,484 (15.9)	4,491 (16.4)	0.077
Type 2 diabetes[b]	7,623 (6.9)	1,997 (24.1)	0.054
Hyperlipidemia[c]	26,261 (23.9)	6,619 (24.1)	0.497
GGT, U/L	23 (16-39)	23 (16-39)	

Body mass index, kg/m^2	23.8 (3.0)	23.8 (3.0)	0.296
Lifestyle behavior			
Cigarette smoking	41,466 (37.7)	10,202 (37.1)	0.067
Alcohol consumption	49,665 (45.2)	12,318 (44.9)	0.286
Physically inactive	74,274 (67.6)	18,398 (67.0)	0.059
Family history			
Family history of CVD	11,361 (10.3)	18,398 (10.4)	0.754
Underlying medical conditions			
Atrial fibrillation	12 (0.01)	3 (0.01)	0.250
Peripheral artery disease	2,434 (2.2)	622 (2.3)	0.622
Retinal vein occlusion	417 (0.4)	131 (0.5)	0.220
Retinal artery occlusion	47 (0.04)	8 (0.03)	0.312
Anemia	13,220 (12.0)	3,325 (12.1)	0.741
NAFLD	6,760 (6.2)	1,632 (5.9)	0.191
CKD	108 (0.1)	30 (0.1)	0.609
Migraine	10,800 (9.8)	2,708 (9.8)	0.885
Parkinson's disease	110 (0.1)	21 (0.08)	0.256
Severe mental illness	3,819 (3.5)	898 (3.3)	0.093
Systemic lupus erythematosus	187 (0.17)	48 (0.2)	0.870
Rheumatic arthritis	488 (0.44)	124 (0.5)	0.871
Dental (oral) health			
Chronic periodontitis	50,699 (46.2)	12,683 (46.2)	0.929
Dental caries	41,365 (37.7)	10,430 (38.0)	0.324
Medication			
Aspirin	11,200 (10.2)	2,784 (10.1)	0.775
Corticosteroid	2,720 (2.5)	731 (2.6)	0.078
Antipsychotics	237 (0.2)	56 (0.2)	0.703
Environmental exposure[e]			
High cumulative exposure to PM 10	32,391 (29.5)	8,082 (29.4)	0.851
Low urban green space coverage	25,712 (23.4)	6413 (23.4)	0.847

NOTE: Data above presented as n (%) or mean ±SD, unless otherwise specified. *p*-value calculated from chi-square test for categorical variables (Fisher's exact test for variables containing categories with observations less than 5) and t-test for continuous variable.

[a]Defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or with antihypertensive prescription (more than 30 days)
[b]Defined as fasting serum glucose ≥ 126 mg/dL or with antidiabetic drug prescription (more than 30 days)
[c]Defined as total cholesterol ≥ 240 mg/dL or with statin prescription (more than 30 days)
[d]Cumulative number of underlying conditions in each cateogry
[e]Cumulative exposure to particulate matter (PM 10) was computed by taking the annual average of PM 10. High cumulative exposure to PM 10 indicates highest quartile.

Urban green space coverage was calculated by the area of parks and artificially designed green space divided by the area of residential districts. Low urban green space indicates lowest quartile. Environmental data were merged with residential area code in the NHIS-NSC with AirKorea database

(PM 10) and Ministry of Land, Infrastructure and Transport database (urban green space)

Abbreviations: NHIS-NSC; National Health Insurance Service-National Sample Cohort; SD, standard deviation; GGT, gamma-glutamyl transpeptidase; CVD, cardiovascular disease; N/C, non-calculable; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease; PM, particulate matter

Among the Uno's C-index for the models constructed with DeepSurv, the hybrid approach using all of the variables that showed statistical significance from the Cox proportional hazards model (Model 2) showed the best performance (Uno's C-index 0.7069; change in C-index +0.045, percent change of +6.73 %, *p*-value for difference in the C-index <0.0001 compared to the model with basic clinical factors). The worst performance of the DeepSurv model was found in the model constructed with common variables included in Model 2-4 (meeting all three of the statistics-based criteria) (Uno's C-index 0.6630; change in C-index +0.001; percent change of +0.11 %, *p*-value for difference in the C-index 0.7231 compared to the model with basic clinical factors). Also, the DeepSurv model with subset of variables with best AIC (Model 3) and variables selected from regularized Cox proportional hazards model (Model 4) showed poor performance with no statistically significant improvement from the model with basic clinical factors for Model 3 and marginal improvement for Model 4, respectively. Comparison of the predictive performance DeepSurv models with input features derived from basic clinical factors, ESC SCORE factors, and multivariable factors with hybrid approach are shown in Table 14.

Table 15. Comparison of the predictive performance of the models for CVD risk with Cox proportional hazards deep neural network (DeepSurv) model with all variables (Model 1) and hybrid approaches (Model 2-5) in the NHIS-NSC linked to the data on environmental exposure

DeepSurv Model	Uno's C-index	Change in C-index	Percent Change	*p*-value for difference
Basic clinical factors	0.6623	- (ref)	- (ref)	- (ref)
ESC SCORE factors	0.6835	+0.021	+3.20 %	<0.0001
Multivariable factors				
Model 1 (All variables)	0.6983	+0.036	+5.44 %	<0.0001
Hybrid approaches with Cox PHM				
Model 2 (Statistically significant variables from Cox PHM)	0.7069	+0.045	+6.73 %	<0.0001
Model 3 (Subset of variables with best AIC from Cox PHM)	0.6782	+0.016	+2.40 %	0.2287
Model 4 (Variable selected from regularized Cox PHM by elastic net penalty)	0.6840	+0.022	+3.28 %	<0.0001
Model 5 (Common variables included in Model 2-4)	0.6630	+0.001	+0.11 %	0.7231

NOTE: Variables included in each model is listed below.
Basic clinical factors: age, sex, BMI

ESC SCORE: age, sex, systolic blood pressure, total cholesterol, and cigarette smoking

Multivariable factors:
Model 1 includes the following variables: age, sex, income status, hypertension, type 2 diabetes, hyperlipidemia, GGT, BMI, cigarette smoking, alcohol consumption, physically inactive, family history of CVD, atrial fibrillation, peripheral artery disease, retinal vein occlusion, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus, rheumatic arthritis
is, chronic periodontitis, dental caries, aspirin, corticosteroid, antipsychotics, high particulate matter (PM10), and low urban green space

Model 2 includes statistically significant variables from the Cox proportional hazards model: age, sex, income status, hypertension, type 2 diabetes, hyperlipidemia, GTP, BMI, cigarette smoking, alcohol consumption, physically inactive, family history of CVD, NAFLD, CKD, migraine, Parkinson's disease, rheumatic arthritis, chronic periodontitis, aspirin, corticosteroid, and low urban green space

Model 3 includes variables with best (minimum) AIC: age, sex, income status, hypertension, type 2 diabetes, hyperlipidemia, GGT, BMI, cigarette smoking, alcohol consumption, physically inactive, family history of CVD, atrial fibrillation, peripheral artery disease, retinal vein occlusion, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus

Model 4 includes variables selected from the regularized Cox proportional hazards regression model with elastic net penalty: age , income status, hypertension, hyperlipidemia, alcohol consumption, physically inactive, family history of CVD, atrial fibrillation, peripheral artery disease, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus, rheumatic arthritis, chronic periodontitis, dental caries, aspirin, corticosteroid, antipsychotics, high particulate matter (PM10), and low urban green space

Model 5 includes: age, income status, hypertension, hyperlipidemia, alcohol consumption, physically inactive, family history of CVD, NAFLD, CKD, migraine, Parkinson's disease

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; ESC SCORE, European Society of Cardiology Systematic Coronary Risk Evaluation; Cox PHM, Cox proportional hazards model; AIC, Akaike Information Criteria; GGT, gamma-glutamyl transpeptidase; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

*Difference in Uno's C-index (concordance statistic) in each model compared to the model with basic clinical factors

In the Cox proportional hazards models, the best performance was observed in the model with all variables without any statistics-based variable selection methods (Model 1) (Uno's C-index 0.7041; change in C-index +0.041, percent change of +6.17 %, *p*-value for difference in the C-index <0.0001 compared to the model with basic clinical factors). With the exception of the Cox proportional hazards regression model built with common variables meeting all three of the statistics-based criteria for statistical significance, best AIC, and elastic net penalty (Model 5), other models showed statistically significant improvement in predictive performance (Table 14). The overall performance benefit of using DeepSurv was observed in the hybrid approach of Model 2 with hybrid approach using statistically significant variables selected from the Cox proportional hazards model (Uno's C-index: 0.7069) and was the highest performance observed in DeepSurv and Cox proportional hazards models using variable sets selected from basic clinical factors, ESC SCORE factors, and multivariable factors with statistics-based approach for variable selection (Model 1-5). Due to the poor performance of the models constructed with variable sets from elastic net penalty and common variables meeting all of the statistics-based criteria, these two approaches were neglected in the models developed with progressively adding input features in each data category from sociodemographic factors to environmental exposure. Therefore, the progressive approach was limited to all variables and statistically significant variables in each data category.

Table 16. Comparison of the predictive performance of the models for CVD risk with Cox proportional hazards model in the NHIS-NSC linked to the data on environmental exposure

Cox PH Model	Uno's C-index	Change in C-index	Percent Change	*p*-value for difference*
Basic clinical factors	0.6642	- (ref)	- (ref)	- (ref)
ESC SCORE factors	0.6838	+0.020	+2.95 %	<0.0001
Multivariable factors				
Model 1 (All variables)	0.7052	+0.041	+6.17 %	<0.0001
Model 2 (Statistically significant variables from Cox PHM)	0.7041	+0.040	+6.01 %	<0.0001
Model 3 (Subset of variables with best AIC from Cox PHM)	0.6988	+0.035	+5.21 %	<0.0001
Model 4 (Variable selected from regularized Cox PHM by elastic net penalty)	0.6873	+0.023	+3.48 %	0.0025
Model 5 (Common variables included in Model 2-4)	0.6706	+0.006	+0.96 %	0.3362

NOTE: Variables included in each model is listed below.
Basic clinical factors: age, sex, BMI

ESC SCORE: age, sex, systolic blood pressure, total cholesterol, and cigarette smoking

Multivariable factors:
Model 1 includes the following variables: age, sex, income status, hypertension, type 2 diabetes, hyperlipidemia, GGT, BMI, cigarette smoking, alcohol consumption, physically inactive, family history of CVD, atrial fibrillation, peripheral artery disease, retinal vein occlusion, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus, rheumatic arthritis
is, chronic periodontitis, dental caries, aspirin, corticosteroid, antipsychotics, high particulate matter (PM10), and low urban green space

Model 2 includes statistically significant variables from the Cox proportional hazards model: age, sex, income status, hypertension, type 2 diabetes, hyperlipidemia, GTP, BMI, cigarette smoking, alcohol consumption, physically inactive, family history of CVD, NAFLD, CKD, migraine, Parkinson's disease, rheumatic arthritis, chronic periodontitis, aspirin, corticosteroid, and low urban green space

Model 3 includes variables with best (minimum) AIC: age, sex, income status, hypertension, type 2 diabetes, hyperlipidemia, GGT, BMI, cigarette smoking, alcohol consumption, physically inactive, family history of CVD, atrial fibrillation, peripheral artery disease, retinal vein occlusion, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus

Model 4 includes variables selected from the regularized Cox proportional hazards regression model with elastic net penalty: age , income status, hypertension, hyperlipidemia, alcohol consumption, physically inactive, family history of CVD, atrial fibrillation, peripheral artery disease, retinal artery occlusion, anemia, NAFLD, CKD, migraine, Parkinson's disease, severe mental illness, systemic lupus erythematosus, rheumatic arthritis, chronic periodontitis, dental caries, aspirin, corticosteroid, antipsychotics, high particulate matter (PM10), and low urban green space

Model 5 includes common variables included in model 2,3, and 4 (meeting all of the criteria): age, income status, hypertension, hyperlipidemia, alcohol consumption, physically inactive, family history of CVD, NAFLD, CKD, migraine, Parkinson's disease

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; ESC SCORE, European Society of Cardiology Systematic Coronary Risk Evaluation; AIC, Akaike Information Criteria; GGT, gamma-glutamyl transpeptidase; NAFLD, non-alcoholic fatty liver disease; CKD, chronic kidney disease

*Difference in Uno's C-index (concordance statistic) in each model compared to the model with basic clinical factors

When the input features in each data category containing all the variables ranging from sociodemographic factors to environmental exposure were progressively added for DeepSurv and Cox proportional hazards, the performance of the Cox proportional hazards model steadily improved with the highest performance observed in step 7 (all of the variables included in sociodemographic, clinical laboratory test and measurement, lifestyle behavior, family history, medical conditions, dental health, and medication). Extending the data category to environmental exposure did not offer improved performance compared to the model constructed in the previous step. The DeepSurv model showed the best performance in step 3 (all of the variables included in sociodemographic, clinical laboratory test and measurement, and lifestyle behavior). For DeepSurv models, using additional data categories beyond step 3 did not show performance benefit. Although addition of environmental exposure showed a minimal improvement in the DeepSurv model (step 8 compared to step 7), this change was not statistically significant (*p*-value for difference in Uno's C-index). Comparison of the performance between DeepSurv and Cox proportional hazards regression models by progressively adding data categories is depicted in Figure 14.

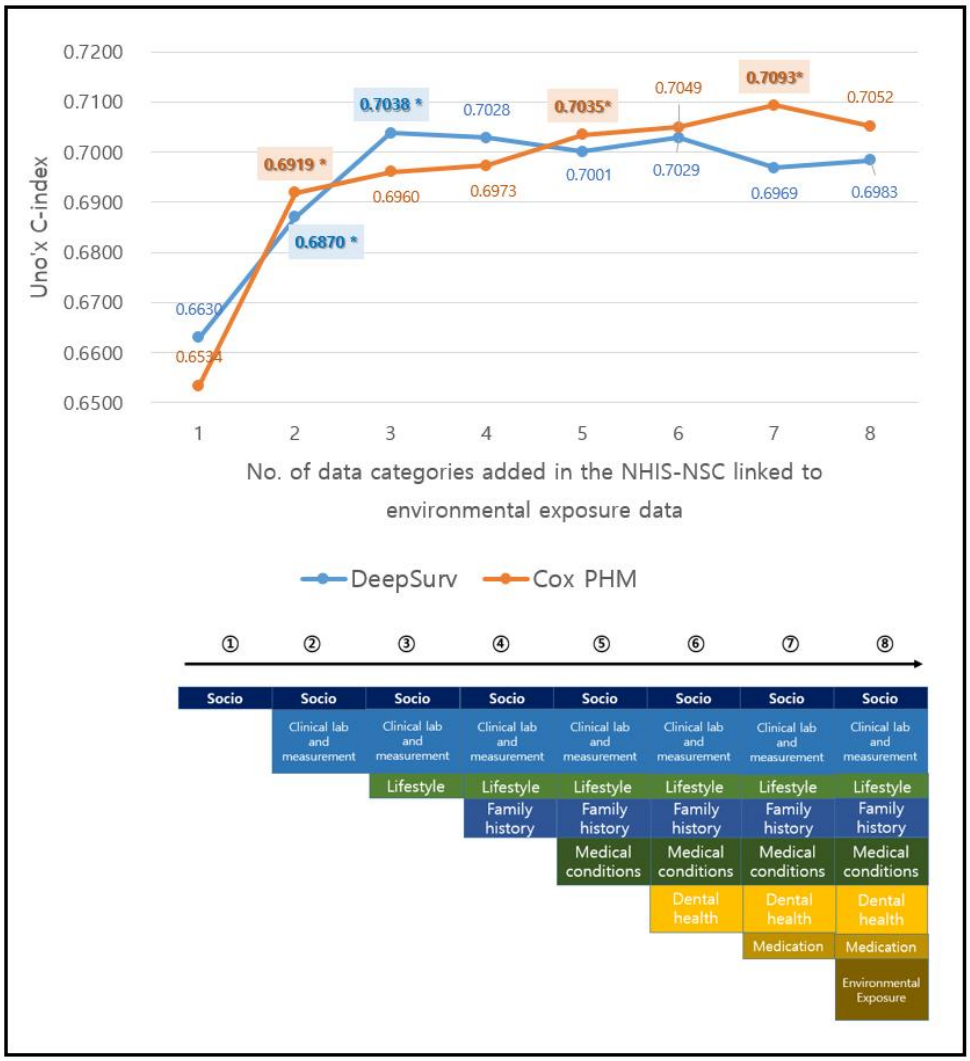

Figure 14. Performance evaluation of the DeepSurv and Cox proportional hazards model for CVD risk by progressively adding variables from accessible data categories in the NHIS-NSC linked to environmental exposure data

NOTE: See Table 9 for the list of variables included in each data category from 1 to 8

*Denotes the statistical significance for difference in Uno's concordance statistic (C-index) compared to the previous model (i.e. 2. vs. 3 and 3 vs. 4). Refer to Table 8 for the list of variables included in each data category.

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; Cox PHM, Cox proportional hazards model; Socio, sociodemographic factors

As the data category and the relevant variables that showed statistical significance in the Cox proportional hazards regression analysis for incident CVD event was progressively added to the DeepSurv and Cox proportional hazards model, performance for both of the models showed steady improvement. However, DeepSurv model constructed in step 3 (all of the statistically significant variables included in sociodemographic, clinical laboratory test and measurement, and lifestyle behavior, which is same as all variables since all of the variables in these data categories showed statistical significance) outperformed the Cox proportional hazards model. Progressively adding data categories containing statistically significant variables chosen from Cox proportional hazards regression for DeepSurv (hybrid approach) models showed superior performance to the Cox proportional hazards models across all of the data categories. Although addition of the environmental exposure data (low urban green space coverage) offered marginal improvement in the performance for both DeepSurv and Cox proportional hazards models, the difference in Uno's C-index was not statistically significant compared to the model constructed in the step before (difference in Uno's C-index in step 8 compared to step 7).

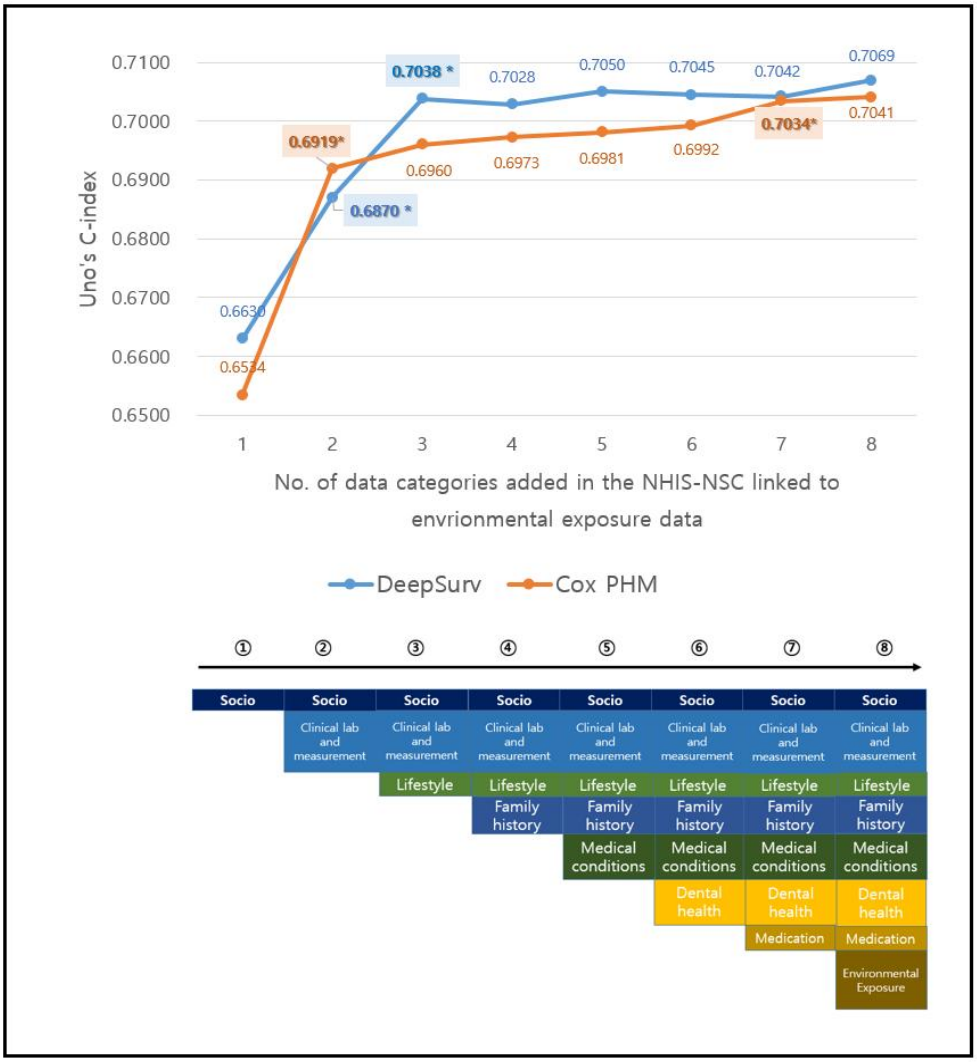

Figure 15. Performance evaluation of the DeepSurv with a hybrid approach and Cox proportional hazards model for CVD risk by progressively adding statistically significant variables from accessible data categories in the NHIS-NSC linked to environmental exposure data

NOTE: See Table 10 for the list of variables included in each data category from 1 to 8

*Denotes the statistical significance for difference in Uno's concordance statistic (C-index) compared to the previous model (i.e. 2. vs. 3 and 3 vs. 4). Refer to Table 9 for the list of variables included in each data category.

Abbreviations: CVD, cardiovascular disease; NHIS-NSC, National Health Insurance Service-National Sample Cohort; Cox PHM, Cox proportional hazards model; Socio, sociodemographic factors

IV. DISCUSSION

1. Key findings and contributions

In this large, population-based data derived from a nationally representative cohort linked to environmental exposure data, the best performance of the DeepSurv model was found in a hybrid approach when a set of statistically significant variables from the Cox proportional hazards regression. Unlike DeepSurv, the best performance for Cox proportional hazards regression was observed when all variables were used as input features. With progressively adding all and statistically significant variables in each category ranging from sociodemographic factors to environmental exposure, the performance of Cox proportional hazards regression steadily increased for when all and statistically significant variables were used as input features. Meanwhile, incremental improvement in the performance was notable in the DeepSurv models when only statistically significant variables were progressively added. Also, input features from simple assessment of sociodemographic factors, clinical laboratory test and measurement, and lifestyle behavior for DeepSurv robustly outperformed Cox proportional hazards model with addition of statistically significant variables up to data categories pertaining to lifestyle behavior, family history, medical condition, dental health, and medication. For both DeepSurv and Cox proportional hazards models, extending the data category to environmental exposure did not offer significant performance benefit.

This study makes two main contributions in the data-driven cardiovascular health research: (1) the evaluation of the predictive performance of DeepSurv and Cox proportional hazards regression with input features selected from extensive literature review and statistics-based variable selection methods and

(2) their relative contribution to the performance of the models progressively extending the data category from sociodemographic to environmental exposure data. Overall, this study provided evidence on the best method for selecting input features for DeepSurv and Cox proportional hazards model and explored the potential benefit of expanding the data categories as input features, especially with data linkage to environmental exposure.

2. Comparison to other studies

It is well-known that excessive data collection could be costly and maybe unnecessary if the collected data do not extensively contribute to the predictive performance of the models for CVD outcome[140]. In general, most of the studies on CVD risk assessment did not provide further evidence on the possibility of change in model performance if more data on CVD risk factors from different data categories were added to the evaluation of the CVD risk. Also, most of the studies have examined the CVD risk as a binary outcome rather than the aspect of survival analysis with deep neural networks.

With regards to the degree of adding variable from multiple data categories to the predictive model with DeepSurv and Cox proportional hazards regression, significant improvement was observed when clinical laboratory test and measurement and lifestyle behavior data were added to sociodemographic factors. However, only marginal improvement in the performance was found in DeepSurv and Cox proportional hazards regression when environmental exposure data were added in the hybrid approach.

This finding suggests, in relation with previous studies, that data from multiple dimensions should be considered for the optimal performance of the

predictive model. In the COronary CT Angiography EvaluatioN For Clinical Outcomes: An InteRnational Multicenter (CONFIRM) registry study conducted with patients suspected of coronary artery disease (CAD), the investigators used the information abstracted from the computed tomographic angiography and clinical variables to build a machine learning (ML) based model for predicting all-cause mortality[141]. The CONFIRM registry study found that the ensemble ML technique outperformed the well-established models including Framingham Risk Score[142] (FRS), segment stenosis score[143] (SSS), segment involvement score[144] (SIS), and modified Duke index[145] (DI). While the performance of the ML technique for predicting the all-cause mortality as a binary outcome was notable in this study, the metrics that were compared (FRS, SSS, SIS, and DI) to the ML technique are used for assessing cardiovascular risk, not the risk of all-cause mortality, While the CONFIRM registry study used a wide range of data dimensions from clinical imaging and data, the input features were selected based on the information gain and the contribution of each data type were not examined. Also, the patients in the CONFIRM registry comprised of those suspected with CAD whose coronary angiography data were available. In the analysis with multiple data categories derived from the NHIS-NSC, the subjects were free of CVD at baseline without any information on clinical imaging, and the survival analysis models were implemented rather than binary classification models for the cardiovascular outcome.

A previous study using ML algorithms among elderly patients referred to the Sandwell and West Birmingham Hospitals National Health Service (NHS) Trust in the United Kingdom to identify atrial fibrillation (AF) found that adding cardiovascular biomarker data (brain natriuretic peptide [BNF], fibroblast growing

factor [FGF-23], tumor necrosis factor-related apoptosis-induced ligand receptor 2 [TRAIL-2]) into the clinical risk factors (age, sex, and body mass index) improved the predictive performance of the model[13]. While this study shows the value of quantifying cardiovascular biomarkers as additional input features for improving the performance of the model for identifying AF, other common risk factors such as hypertension or diabetes were not considered in the models. Although the extensive collection of the biomarker data associated with AF has shown to be valuable in this study, the contribution of the common and rather easily collectable risk factors[146,147] (e.g. hypertension and diabetes) was not extensively evaluated. Since information on hypertension or diabetes can be more easily assessed with clinical measurement and added to the ML-based model compared to the biomarker data, finding the input features that could maximize the performance of the model should be implemented considering both cost and effectiveness of collecting such data. In the NHIS-NSC study with data linkage to the environmental exposure data with DeepSurv and Cox proportional hazards regression, the input features selected from the statistics-based models (hybrid approach for DeepSurv) were also incrementally added. With this attempt in the NHIS-NSC study, the potential contribution of each data dimension were comprehensively examined. Also, the time element was considered as a part of the survival analysis model in the NHIS-NSC study compared to the report from the Sandwell and West Birmingham Hospitals NHS Trust.

While DeepSurv successfully models increasingly complex relationships between a patient's covariates and their risk of failure in other studies, especially in real survival data experiments such as Worcester Heart Attack Study (WHAS) with 1638 observations and 5 features (age, sex, BMI, left heart failure complications,

and order of MI, which showed approximately 0.05 increase in C-index compared to cox proportional hazards model. In the NHIS-NSC study, significant performance benefit of DeepSurv was observed when clinical laboratory test and measurement data and lifestyle behavior data were added to the sociodemographic data. When implementing predictive modeling for future CVD risk using deep neural network and regression-based survival analysis from the NHIS-NSC, input features from sociodemographic, clinical laboratory test and measurement, and lifestyle behavior should be primarily considered before collecting risk factors from other data categories such as medical/dental claims, medication use, and data linkage to environmental exposure.

Recent evidence suggests that learning-based algorithms do not always outperform traditional statistical methods for the prediction of clinical outcomes. A recent systematic review comparing the performance of ML algorithms (e.g. random forest, artificial neural networks, and support vector machines) and logistic regression for binary classification of study outcomes based on 71 studies showed that the difference in area under the curve was negligible among the comparisons with low risk of bias (difference in AUC: 0.00; 95% CI: -0.18 to 0.18 for 145 comparisons at low risk of bias)[148]. In addition, data from a recent study on Medicare patients concluded that, ML algorithms showed only marginal improvement over logistic regression for predicting hospitalization with heart failure[149].

Although the outcome and evaluation method for comparison of DeepSurv with a hybrid approach was based on predicted log-hazard and time-dependent C-index rather than binary classification of the outcomes, the findings from the recent meta-analysis and Medicare study support the evidence that

learning-based algorithms do not always show superior performance to the traditional statistical methods.

Also, while the exact factors that lead to the difference in performance was not fully known in this study due to the lack of explainable artificial intelligence (XAI) technique for DeepSurv, more studies should be conducted with different number of risk factors, follow-up duration, and number of events. Although Cox proportional hazards model has been the standard method for clinical risk prediction, use of deep learning based survival analysis such as DeepSurv should be considered in clinical risk modeling with patient-level survival data o find the best performance of the model when multiple source and data likage are available.

3. Strengths and limitations

In contrast to the most recent predictive modeling studies for cardiovascular event as a binary outcome, this study with a large population-based data implemented deep learning based survival analysis combined with Cox proportional hazards regression (i.e. hybrid approach) with sufficient patient-level data linked to environmental exposure. Thus, this study allowed the evaluation of model performance from statistics-based variable selection methods and progressively adding variables from different data categories.

Potential limitations of this study should also be noted. This analysis addressing the survival model does not clearly provide etiological background or proposed mechanism. Furthermore, larger studies with more data dimensions (e.g. biomarkers, clinical imaging and measurement data, etc) and longer follow-up duration in diverse populations are needed to assess the effectiveness of extensive

data collection of cardiovascular risk factors for predictive performance of the deep learning based survival models. The current analysis has a limited generalizability because the data was derived from a single ethnic-group with a relatively short follow-up duration for the outcome. Also, data on environmental exposure to particulate matter and urban green space coverage was linked to the NHIS-NSC using residential area code as the common key. This area-level exposure potentially owes to the large variations among individuals residing in the same administrative area with different daily exposure to particulate matter and urban green space. Thus, future study should collect individualized measurement for environmental exposure (i.e. data from portable particulate monitor or smartphone geolocation) and utilize them to test if addition of individually measured environmental data could significantly improve the performance of deep learning based survival analysis model.

4. Implications

In the real-world settings, especially for the enrollees of the NHIS who undergo national health screening, their past history of medical/dental claims and drug prescription are routinely collected and easily traceable through the NHIS system. Therefore, policymakers for public health could consider the cost associated with collecting the information before implementing predictive modeling approach of any type. While the NHIS currently provides data-driven service on computing the health risk for few diseases based on the integrated information derived from the medical claims, climate, and social network service, whether the predicted health outcome could be improved when other data dimensions are added is not clearly determined. Because the NHIS database is

being managed at the national level, marginal improvement with the multivariable model in the predictive models could inform a large number of the individuals of their future health risk. Also, the performance of Cox proportional hazards regression steadily increased as more number of variables in data categories were added up from sociodemographic factors to environmental exposure data. Nonetheless, the DeepSurv model with variables included in sociodemographic factors, clinical laboratory and measurement, and health behavior could offer the best predictive performance with minimal number of input features, and given that these are relatively easily accessible from the insurance eligibility and health screening data in the clinical settings or the NHIS system, the effectiveness of expanding the data on CVD risk factors up to environmental exposure data for the entire population should be carefully reviewed before implementation of prediction models for CVD risk assessment.

5. Future perspectives

The current DeepSurv model for CVD risk assessment lacks explainability and automated variable selection technique considering survival (i.e. right-censored) data. Along with the ongoing expansion of the available data source, advanced deep learning based survival analysis models in the future should be implemented with explainability and automated variable selection techniques for comprehensive cardiovascular risk assessment and evaluating the relative contribution of each data type for predictive performance (Figure 16). Also, data on environmental exposure should be measured for each individual for more precise assessment for cardiovascular health risk.

While this study could only integrate information on sociodemographic,

medical/dental claims, lifestyle behavior, medication, some of the clinical laboratory test and measurement, and environmental exposure, future study should consider expanding the data collection to multiple sources such as dietary factors, web search quires, financial data, genomics, gut microbiome, and medical images for potential applications in personalized healthcare, public health policy, and data-driven health research.

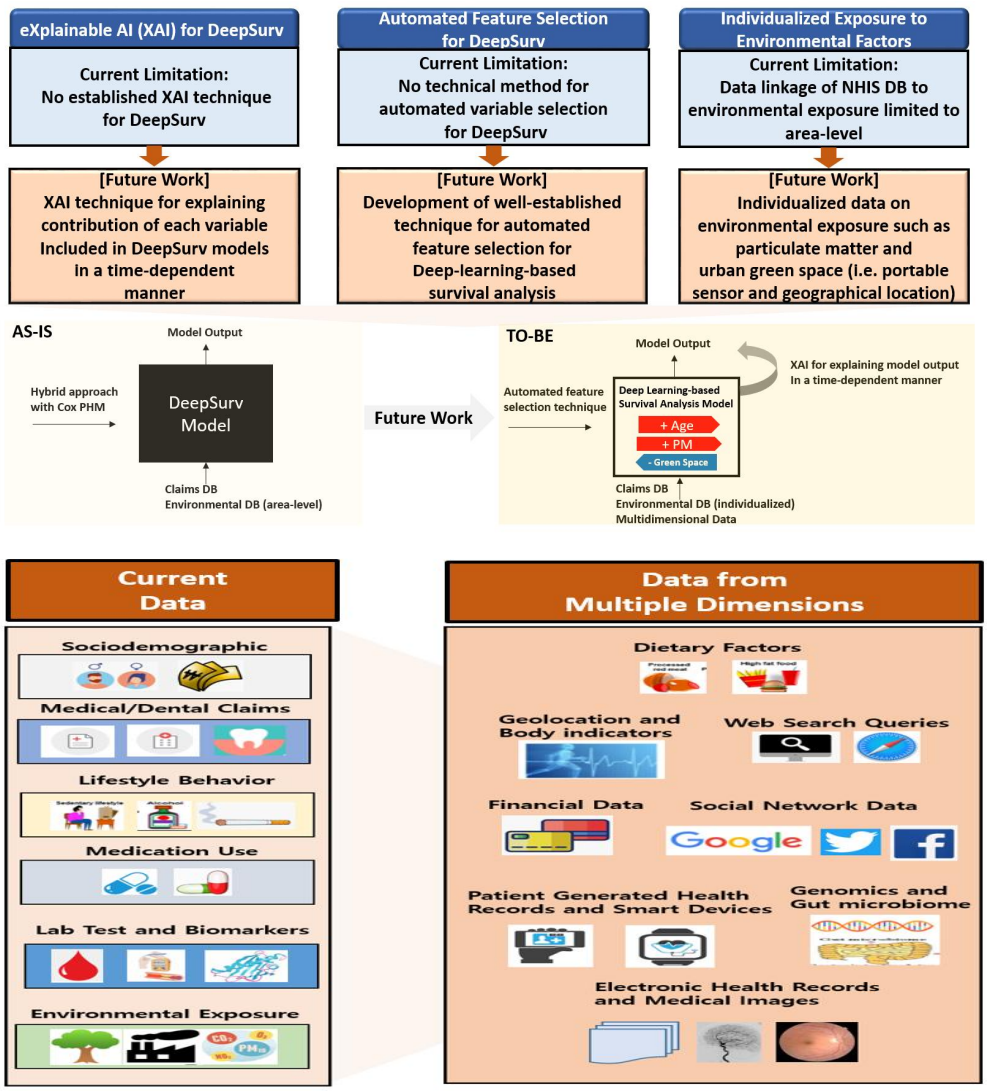

Figure 16. Future perspectives of the data-driven cardiovascular research using integrated data from multiple dimensions for advanced deep-learning based survival analysis models

Abbreviation: XAI, eXplainable Artificial Intelligence

V. CONCLUSION

In summary, abstracting information on multivariable factors offered improvement in the predictive performance of DeepSurv model for CVD risk compared to the basic clinical factors comprised of age, sex, and body mass index, especially with a hybrid approach when statistically significant variables from Cox proportional hazards model were selected as input feature set and progressively added. Information on sociodemographic factors, clinical laboratory test and measurement, and lifestyle behavior enriched the performance benefit of DeepSurv model for CVD risk assessment that was superior to the Cox proportional hazards models with statistically significant variables added up to medication use. To attain the best performance of the predictive modeling for CVD risk with DeepSurv using the minimum number of data categories, sociodemographic factors, clinical laboratory test and measurement, and lifestyle behavior data abstracted from the NHIS-NSC should be primarily considered. Also, extensive data linkage for input features should be carefully determined prior to the model development with DeepSurv as expanding the data categories up to environmental exposure data from the NHIS-NSC only offered marginal improvement in predictive performance for CVD risk. Future studies with deep learning based survival analysis for CVD risk assessment should be implemented with explainable artificial intelligence technique, automated variable selection methods, and individualized data on environmental exposure with other data sources derived from multiple dimensions.

REFERENCES

1. McAloon CJ, Boylan LM, Hamborg T, et al. The changing face of cardiovascular disease 2000–2012: An analysis of the world health organisation global health estimates data. *International journal of cardiology.* 2016;224:256-264.
2. Santulli G. Epidemiology of cardiovascular disease in the 21st century: updated numbers and updated facts. *J Cardiovasc Dis.* 2013;1(1):1-2.
3. Organization WH. *WHO global coordination mechanism on the prevention and control of noncommunicable diseases: final report: WHO GCM.* World Health Organization;2018.
4. Tarride J-E, Lim M, DesMeules M, et al. A review of the cost of cardiovascular disease. *Canadian Journal of Cardiology.* 2009;25(6):e195-e202.
5. Mensah GA, Brown DW. An overview of cardiovascular disease burden in the United States. *Health affairs.* 2007;26(1):38-48.
6. Le C, Fang Y, Linxiong W, Shulan Z, Golden A. Economic burden and cost determinants of coronary heart disease in rural southwest China: a multilevel analysis. *Public health.* 2015;129(1):68-73.
7. Roth GA, Johnson C, Abajobir A, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American College of Cardiology.* 2017;70(1):1-25.
8. Mendis S. The contribution of the Framingham Heart Study to the prevention of cardiovascular disease: a global perspective. *Progress in cardiovascular diseases.* 2010;53(1):10-14.
9. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *Bmj.* 2012;344:e3318.
10. Ridker PM. Clinical application of C-reactive protein for cardiovascular disease detection and prevention. *Circulation.* 2003;107(3):363-369.
11. de Oliveira C, Watt R, Hamer M. Toothbrushing, inflammation, and risk of cardiovascular disease: results from Scottish Health Survey. *Bmj.* 2010;340:c2451.
12. Brook RD, Franklin B, Cascio W, et al. Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation.* 2004;109(21):2655-2671.
13. Chua W, Purmah Y, Cardoso VR, et al. Data-driven discovery and validation of circulating blood-based biomarkers associated with prevalent atrial fibrillation. *European heart journal.* 2019;40(16):1268-1276.
14. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal.* 2017;38(23):1805-1814.
15. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circulation research.* 2017;121(9):1092-1101.
16. Collaboration ABI. Ankle brachial index combined with Framingham Risk

Score to predict cardiovascular events and mortality: a meta-analysis. *JAMA: the journal of the American Medical Association.* 2008;300(2):197.
17. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *Journal of the American Heart Association.* 2018;7(22):e009476.
18. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one.* 2017;12(4).
19. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology.* 2018;18(1):24.
20. Faraggi D, Simon R. A neural network model for survival data. *Statistics in medicine.* 1995;14(1):73-82.
21. Xiang A, Lapuerta P, Ryutov A, Buckley J, Azen S. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational statistics & data analysis.* 2000;34(2):243-257.
22. Mariani L, Coradini D, Biganzoli E, et al. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast cancer research and treatment.* 1997;44(2):167-178.
23. Lee C, Luo Z, Ngiam KY, et al. Big healthcare data analytics: Challenges and applications. In: *Handbook of Large-Scale Distributed Computing in Smart Healthcare.* Springer; 2017:11-41.
24. Islam MM, Razzaque MA, Hassan MM, Ismail WN, Song B. Mobile cloud-based big healthcare data processing in smart cities. *IEEE Access.* 2017;5:11887-11899.
25. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS medicine.* 2014;11(10).
26. Taiyari K. *How much data are required to develop and validate a risk prediction model?* , UCL (University College London); 2017.
27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software.* 2011;39(5):1.
28. Control D, Trial C, Interventions EoD, Group CR. Risk factors for cardiovascular disease in type 1 diabetes. *Diabetes.* 2016;65(5):1370-1379.
29. Kim K, Choi S, Hwang SE, et al. Changes in exercise frequency and cardiovascular outcomes in older adults. *European heart journal.* 2019.
30. Kim K, Park SM, Lee K. Weight gain after smoking cessation does not modify its protective effect on myocardial infarction and stroke: evidence from a cohort study of men. *European heart journal.* 2018;39(17):1523-1531.
31. Son JS, Choi S, Kim K, et al. Association of blood pressure classification in Korean young adults according to the 2017 American College of Cardiology/American Heart Association guidelines with subsequent cardiovascular disease events. *Jama.* 2018;320(17):1783-1792.
32. Cox DR. Regression models and life-tables. *Journal of the Royal*

Statistical Society: Series B (Methodological). 1972;34(2):187-202.
33. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research.* 2012;13(1):281-305.
34. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine.* 2011;30(10):1105-1117.
35. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama.* 1982;247(18):2543-2546.
36. Harrell Jr FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine.* 1984;3(2):143-152.
37. McInnes MD, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *Jama.* 2018;319(4):388-396.
38. Alzamora MT, Forés R, Pera G, et al. Ankle-brachial index and the incidence of cardiovascular events in the Mediterranean low cardiovascular risk population ARTPER cohort. *BMC Cardiovascular Disorders.* 2013;13(1):119.
39. Collaboration ERF. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *The Lancet.* 2010;375(9709):132-140.
40. Blaha MJ, Budoff MJ, DeFilippis AP, et al. Associations between C-reactive protein, coronary artery calcium, and cardiovascular events: implications for the JUPITER population from MESA, a population-based cohort study. *The Lancet.* 2011;378(9792):684-692.
41. Pletcher MJ, Tice JA, Pignone M, Browner WS. Using the coronary artery calcium score to predict coronary heart disease events: a systematic review and meta-analysis. *Archives of internal medicine.* 2004;164(12):1285-1292.
42. Sniderman AD, Williams K, Contois JH, et al. A meta-analysis of low-density lipoprotein cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B as markers of cardiovascular risk. *Circulation: Cardiovascular Quality and Outcomes.* 2011;4(3):337-345.
43. Hwang Y-C, Ahn H-Y, Han KH, Park S-W, Park C-Y. Prediction of future cardiovascular disease with an equation to estimate apolipoprotein B in patients with high cardiovascular risk: an analysis from the TNT and IDEAL study. *Lipids in Health and Disease.* 2017;16(1):158.
44. Du G, Song Z, Zhang Q. Gamma-glutamyltransferase is associated with cardiovascular and all-cause mortality: a meta-analysis of prospective cohort studies. *Preventive medicine.* 2013;57(1):31-37.
45. Yang W, Kim CK, Kim DY, Jeong HG, Lee SH. Gamma-glutamyl transferase predicts future stroke: A Korean nationwide study. *Annals of Neurology.* 2018;83(2):375-386.
46. Wahid A, Manek N, Nichols M, et al. Quantifying the association between physical activity and cardiovascular disease and diabetes: a systematic review and meta-analysis. *Journal of the American Heart Association.* 2016;5(9):e002495.
47. Kim K, Choi S, Hwang SE, et al. Changes in exercise frequency and cardiovascular outcomes in older adults. *European Heart Journal.* 2020;41(15):1490-1499.

48. Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *European heart journal.* 2011;32(12):1484-1492.
49. Chandola T, Ferrie JE, Perski A, Akbaraly T, Marmot MG. The effect of short sleep duration on coronary heart disease risk is greatest among those with sleep disturbance: a prospective study from the Whitehall II cohort. *Sleep.* 2010;33(6):739-744.
50. Vyas MV, Garg AX, Iansavichus AV, et al. Shift work and vascular events: systematic review and meta-analysis. *Bmj.* 2012;345:e4800.
51. Hublin C, Partinen M, Koskenvuo K, Silventoinen K, Koskenvuo M, Kaprio J. Shift-work and cardiovascular disease: a population-based 22-year follow-up study. *European journal of epidemiology.* 2010;25(5):315-323.
52. Kodama S, Saito K, Tanaka S, et al. Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. *Jama.* 2009;301(19):2024-2035.
53. Laukkanen JA, Kurl S, Salonen R, Rauramaa R, Salonen JT. The predictive value of cardiorespiratory fitness for cardiovascular events in men with various risk profiles: a prospective population-based cohort study. *European heart journal.* 2004;25(16):1428-1437.
54. Chainani V, Shaharyar S, Dave K, et al. Objective measures of the frailty syndrome (hand grip strength and gait speed) and cardiovascular mortality: A systematic review. *International journal of cardiology.* 2016;215:487-493.
55. Celis-Morales CA, Welsh P, Lyall DM, et al. Associations of grip strength with cardiovascular, respiratory, and cancer outcomes and all cause mortality: prospective cohort study of half a million UK Biobank participants. *Bmj.* 2018;361.
56. Yang J, Christophi CA, Farioli A, et al. Association between push-up exercise capacity and future cardiovascular events among active adult men. *JAMA network open.* 2019;2(2):e188341-e188341.
57. Khan Z, Almeida DR, Rahim K, Belliveau MJ, Bona M, Gale J. 10-Year Framingham risk in patients with retinal vein occlusion: a systematic review and meta-analysis. *Canadian Journal of Ophthalmology.* 2013;48(1):40-45. e41.
58. Rim TH, Kim DW, Han JS, Chung EJ. Retinal vein occlusion and the risk of stroke development: a 9-year nationwide population-based study. *Ophthalmology.* 2015;122(6):1187-1194.
59. Zhou Y, Zhu W, Wang C. Relationship between retinal vascular occlusions and incident cerebrovascular diseases: A systematic review and meta-analysis. *Medicine.* 2016;95(26).
60. Rim TH, Han J, Choi YS, et al. Retinal artery occlusion and the risk of stroke development: twelve-year nationwide cohort study. *Stroke.* 2016;47(2):376-382.
61. Palmer SC, Hayen A, Macaskill P, et al. Serum levels of phosphorus, parathyroid hormone, and calcium and risks of death and cardiovascular disease in individuals with chronic kidney disease: a systematic review and meta-analysis. *Jama.* 2011;305(11):1119-1127.
62. Di Angelantonio E, Chowdhury R, Sarwar N, Aspelund T, Danesh J,

Gudnason V. Chronic kidney disease and risk of major cardiovascular disease and non-vascular mortality: prospective population based cohort study. *Bmj.* 2010;341:c4986.
63. Targher G, Byrne CD, Lonardo A, Zoppini G, Barbui C. Non-alcoholic fatty liver disease and risk of incident cardiovascular disease: a meta-analysis. *Journal of hepatology.* 2016;65(3):589-600.
64. Zeb I, Li D, Budoff MJ, et al. Nonalcoholic fatty liver disease and incident cardiac events: the multi-ethnic study of atherosclerosis. *Journal of the American College of Cardiology.* 2016;67(16):1965-1966.
65. Zakai NA, Katz R, Hirsch C, et al. A prospective study of anemia status, hemoglobin concentration, and mortality in an elderly cohort: the Cardiovascular Health Study. *Archives of internal medicine.* 2005;165(19):2214-2220.
66. Alves M, Caldeira D, Ferro JM, Ferreira JJ. Does Parkinson's disease increase the risk of cardiovascular events? A systematic review and meta-analysis. *European Journal of Neurology.* 2020;27(2):288-296.
67. Huang Y-P, Chen L-S, Yen M-F, et al. Parkinson's disease is related to an increased risk of ischemic stroke—a population-based propensity score-matched follow-up study. *PLoS One.* 2013;8(9).
68. Lafon A, Pereira B, Dufour T, et al. Periodontal disease and stroke: a meta-analysis of cohort studies. *European journal of neurology.* 2014;21(9):1155-e1167.
69. Hansen GM, Egeberg A, Holmstrup P, Hansen PR. Relation of periodontitis to risk of cardiovascular and all-cause mortality (from a Danish nationwide cohort study). *The American journal of cardiology.* 2016;118(4):489-493.
70. Park S-Y, Kim S-H, Kang S-H, et al. Improved oral hygiene care attenuates the cardiovascular risk of oral health disease: a population-based study from Korea. *European heart journal.* 2019;40(14):1138-1145.
71. Kim K, Hyeon J, Lee SA, et al. Role of Total, Red, Processed, and White Meat Consumption in Stroke Incidence and Mortality: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. *Journal of the American Heart Association.* 2017;6(9):e005983.
72. Larsson SC, Virtamo J, Wolk A. Red meat consumption and risk of stroke in Swedish men. *The American journal of clinical nutrition.* 2011;94(2):417-421.
73. Bernstein AM, Pan A, Rexrode KM, et al. Dietary protein sources and the risk of stroke in men and women. *Stroke.* 2012;43(3):637-644.
74. Haring B, Misialek JR, Rebholz CM, et al. Association of dietary protein consumption with incident silent cerebral infarcts and stroke: The Atherosclerosis Risk in Communities (ARIC) study. *Stroke.* 2015;46(12):3443-3450.
75. Larsson SC, Orsini N. Fish consumption and the risk of stroke: a dose–response meta-analysis. *Stroke.* 2011;42(12):3621-3623.
76. Mozaffarian D, Longstreth W, Lemaitre RN, et al. Fish consumption and stroke risk in elderly individuals: the cardiovascular health study. *Archives of internal medicine.* 2005;165(2):200-206.
77. Gadiraju TV, Patel Y, Gaziano JM, Djoussé L. Fried food consumption and cardiovascular health: a review of current evidence. *Nutrients.*

2015;7(10):8424-8430.
78. Guallar-Castillón P, Rodríguez-Artalejo F, Lopez-Garcia E, et al. Consumption of fried foods and risk of coronary heart disease: Spanish cohort of the European Prospective Investigation into Cancer and Nutrition study. *Bmj.* 2012;344:e363.
79. Wang X, Ouyang Y, Liu J, et al. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. *Bmj.* 2014;349:g4490.
80. Larsson SC, Virtamo J, Wolk A. Total and specific fruit and vegetable consumption and risk of stroke: a prospective study. *Atherosclerosis.* 2013;227(1):147-152.
81. Narain A, Kwok C, Mamas M. Soft drinks and sweetened beverages and the risk of cardiovascular disease and mortality: a systematic review and meta-analysis. *International journal of clinical practice.* 2016;70(10):791-805.
82. Pase MP, Himali JJ, Beiser AS, et al. Sugar-and artificially sweetened beverages and the risks of incident stroke and dementia: a prospective cohort study. *Stroke.* 2017;48(5):1139-1146.
83. Sofi F, Conti AA, Gori AM, et al. Coffee consumption and risk of coronary heart disease: a meta-analysis. *Nutrition, Metabolism and Cardiovascular Diseases.* 2007;17(3):209-223.
84. Kleemola P, Jousilahti P, Pietinen P, Vartiainen E, Tuomilehto J. Coffee consumption and the risk of coronary heart disease and death. *Archives of internal medicine.* 2000;160(22):3393-3400.
85. Guo J, Astrup A, Lovegrove JA, Gijsbers L, Givens DI, Soedamah-Muthu SS. Milk and dairy consumption and risk of cardiovascular diseases and all-cause mortality: dose–response meta-analysis of prospective cohort studies. In: Springer; 2017.
86. Geiker NRW, Larsen ML, Dyerberg J, Stender S, Astrup A. Egg consumption, cardiovascular diseases and type 2 diabetes. *European Journal of Clinical Nutrition.* 2018;72(1):44-56.
87. Qin C, Lv J, Guo Y, et al. Associations of egg consumption with cardiovascular disease in a cohort study of 0.5 million Chinese adults. *Heart.* 2018;104(21):1756-1763.
88. Bøhn SK, Ward NC, Hodgson JM, Croft KD. Effects of tea and coffee on cardiovascular disease risk. *Food & function.* 2012;3(6):575-591.
89. Tanabe N, Suzuki H, Aizawa Y, Seki N. Consumption of green and roasted teas and the risk of stroke incidence: results from the Tokamachi–Nakasato cohort study in Japan. *International journal of epidemiology.* 2008;37(5):1030-1040.
90. Mayhew AJ, de Souza RJ, Meyre D, Anand SS, Mente A. A systematic review and meta-analysis of nut consumption and incident risk of CVD and all-cause mortality. *British Journal of Nutrition.* 2016;115(2):212-225.
91. Bao Y, Han J, Hu FB, et al. Association of nut consumption with total and cause-specific mortality. *New England Journal of Medicine.* 2013;369(21):2001-2011.
92. Ronksley PE, Brien SE, Turner BJ, Mukamal KJ, Ghali WA. Association of alcohol consumption with selected cardiovascular disease outcomes: a

systematic review and meta-analysis. *Bmj.* 2011;342:d671.
93. Smyth A, Teo KK, Rangarajan S, et al. Alcohol consumption and cardiovascular disease, cancer, injury, admission to hospital, and mortality: a prospective cohort study. *The Lancet.* 2015;386(10007):1945-1954.
94. Threapleton DE, Greenwood DC, Evans CE, et al. Dietary fibre intake and risk of cardiovascular disease: systematic review and meta-analysis. *Bmj.* 2013;347:f6879.
95. Kokubo Y, Iso H, Saito I, et al. Dietary fiber intake and risk of cardiovascular disease in the Japanese population: the Japan Public Health Center-based study cohort. *European journal of clinical nutrition.* 2011;65(11):1233-1241.
96. Zhou Y-H, Tang J-Y, Wu M-J, et al. Effect of folic acid supplementation on cardiovascular outcomes: a systematic review and meta-analysis. *PloS one.* 2011;6(9):e25142.
97. Albert CM, Cook NR, Gaziano JM, et al. Effect of folic acid and B vitamins on risk of cardiovascular events and total mortality among women at high risk for cardiovascular disease: a randomized trial. *Jama.* 2008;299(17):2027-2036.
98. Chen GC, Lu DB, Pang Z, Liu QF. Vitamin C intake, circulating vitamin C and risk of stroke: a meta-analysis of prospective studies. *Journal of the American Heart Association.* 2013;2(6):e000329.
99. Osganian SK, Stampfer MJ, Rimm E, et al. Vitamin C and risk of coronary heart disease in women. *Journal of the American College of Cardiology.* 2003;42(2):246-252.
100. Bolland MJ, Grey A, Avenell A, Gamble GD, Reid IR. Calcium supplements with or without vitamin D and risk of cardiovascular events: reanalysis of the Women's Health Initiative limited access dataset and meta-analysis. *Bmj.* 2011;342:d2040.
101. Messenger W, Nielson C, Li H, et al. Serum and dietary vitamin D and cardiovascular disease risk in elderly men: a prospective cohort study. *Nutrition, Metabolism and Cardiovascular Diseases.* 2012;22(10):856-863.
102. Mozaffarian D, Fahimi S, Singh GM, et al. Global sodium consumption and death from cardiovascular causes. *New England Journal of Medicine.* 2014;371(7):624-634.
103. Cook NR, Cutler JA, Obarzanek E, et al. Long term effects of dietary sodium reduction on cardiovascular disease outcomes: observational follow-up of the trials of hypertension prevention (TOHP). *Bmj.* 2007;334(7599):885.
104. Chung M, Lee J, Terasawa T, Lau J, Trikalinos TA. Vitamin D with or without calcium supplementation for prevention of cancer and fractures: an updated meta-analysis for the US Preventive Services Task Force. *Annals of internal medicine.* 2011;155(12):827-838.
105. Van Hemelrijck M, Michaelsson K, Linseisen J, Rohrmann S. Calcium intake and serum concentration in relation to risk of cardiovascular death in NHANES III. *PLoS One.* 2013;8(4).
106. D'Elia L, Barba G, Cappuccio FP, Strazzullo P. Potassium intake, stroke, and cardiovascular disease: a meta-analysis of prospective studies. *Journal of the American College of Cardiology.* 2011;57(10):1210-1219.
107. Umesawa M, Iso H, Date C, et al. Relations between dietary sodium and

potassium intakes and mortality from cardiovascular disease: the Japan Collaborative Cohort Study for Evaluation of Cancer Risks. *The American journal of clinical nutrition.* 2008;88(1):195-202.
108. Zhang X-W, Hou W-S, Li M, Tang Z-Y. Omega-3 fatty acids and risk of cognitive decline in the elderly: a meta-analysis of randomized controlled trials. *Aging clinical and experimental research.* 2016;28(1):165-166.
109. Amiano P, Machón M, Dorronsoro M, et al. Intake of total omega-3 fatty acids, eicosapentaenoic acid and docosahexaenoic acid and risk of coronary heart disease in the Spanish EPIC cohort study. *Nutrition, Metabolism and Cardiovascular Diseases.* 2014;24(3):321-327.
110. Larsson SC, Wolk A. Urinary cadmium and mortality from all causes, cancer and cardiovascular disease in the general population: systematic review and meta-analysis of cohort studies. *International journal of epidemiology.* 2016;45(3):782-791.
111. Tellez-Plaza M, Guallar E, Howard BV, et al. Cadmium exposure and incident cardiovascular disease. *Epidemiology (Cambridge, Mass).* 2013;24(3):421.
112. Navas-Acien A, Guallar E, Silbergeld EK, Rothenberg SJ. Lead exposure and cardiovascular disease—a systematic review. *Environmental health perspectives.* 2007;115(3):472-482.
113. Lanphear BP, Rauch S, Auinger P, Allen RW, Hornung RW. Low-level lead exposure and mortality in US adults: a population-based cohort study. *The Lancet Public Health.* 2018;3(4):e177-e184.
114. Moon KA, Oberoi S, Barchowsky A, et al. A dose-response meta-analysis of chronic arsenic exposure and incident cardiovascular disease. *International journal of epidemiology.* 2017;46(6):1924-1939.
115. Chen Y, Graziano JH, Parvez F, et al. Arsenic exposure from drinking water and mortality from cardiovascular disease in Bangladesh: prospective cohort study. *Bmj.* 2011;342:d2431.
116. Mitter SS, Vedanthan R, Islami F, et al. Household fuel use and cardiovascular disease mortality: Golestan cohort study. *Circulation.* 2016;133(24):2360-2369.
117. Seo S, Choi S, Kim K, Kim SM, Park SM. Association between urban green space and the risk of cardiovascular disease: A longitudinal study in seven Korean metropolitan areas. *Environment international.* 2019;125:51-57.
118. Lu F, Xu D, Cheng Y, et al. Systematic review and meta-analysis of the adverse health effects of ambient PM2. 5 and PM10 pollution in the Chinese population. *Environmental research.* 2015;136:196-204.
119. Crouse DL, Peters PA, van Donkelaar A, et al. Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: a Canadian national-level cohort study. *Environmental health perspectives.* 2012;120(5):708-714.
120. Hvidtfeldt UA, Sørensen M, Geels C, et al. Long-term residential exposure to PM2. 5, PM10, black carbon, NO2, and ozone and mortality in a Danish cohort. *Environment international.* 2019;123:265-272.
121. Mustafić H, Jabre P, Caussin C, et al. Main air pollutants and myocardial infarction: a systematic review and meta-analysis. *Jama.* 2012;307(7):713-721.

122. Jeong SM, Choi S, Kim K, et al. Effect of change in total cholesterol levels on cardiovascular disease among young adults. *Journal of the American Heart Association.* 2018;7(12):e008819.
123. Choi S, Kim K, Kim SM, et al. Association of obesity or weight change with coronary heart disease among young adults in South Korea. *JAMA internal medicine.* 2018;178(8):1060-1068.
124. Choi S, Kim K, Lee J-K, et al. Association between Change in Alcohol Consumption and Metabolic Syndrome: Analysis from the Health Examinees Study. *Diabetes & metabolism journal.* 2019;43(5):615-626.
125. Choi S, Chang J, Kim K, et al. Association of smoking cessation after atrial fibrillation diagnosis on the risk of cardiovascular disease: a cohort study of South Korean men. *BMC Public Health.* 2020;20(1):168.
126. Oh EH, Ro YS, Kim JE. Epidemiology and cardiovascular comorbidities in patients with psoriasis: A Korean nationwide population-based cohort study. *The Journal of dermatology.* 2017;44(6):621-629.
127. Lee G, Choi S, Kim K, et al. Association of hemoglobin concentration and its change with cardiovascular and all-cause mortality. *Journal of the American Heart Association.* 2018;7(3):e007723.
128. Lee JY, Kang S, Park JS, Jo SJ. Prevalence of psoriasis in Korea: a population-based epidemiological study using the Korean National Health Insurance Database. *Annals of dermatology.* 2017;29(6):761-767.
129. Kim KM, Oh HJ, Choi HY, Lee H, Ryu D-R. Impact of chronic kidney disease on mortality: A nationwide cohort study. *Kidney research and clinical practice.* 2019;38(3):382.
130. Min C, Lim H, Lim J-S, Sim S, Choi HG. Increased risk of migraine in patients with psoriasis: A longitudinal follow up study using a national sample cohort. *Medicine.* 2019;98(17).
131. Choi S, Kim K, Chang J, et al. Association of chronic periodontitis on alzheimer's disease or vascular dementia. *Journal of the American Geriatrics Society.* 2019;67(6):1234-1239.
132. Ko A, Kim K, Son JS, Park HY, Park SM. Association of pre-existing depression with all-cause, cancer-related, and noncancer-related mortality among 5-year cancer survivors: a population-based cohort study. *Scientific Reports.* 2019;9(1):1-9.
133. Bae EH, Lim SY, Han K-D, et al. Trend of prevalence and incidence of systemic lupus erythematosus in South Korea, 2005 to 2015: a nationwide population-based study. *The Korean journal of internal medicine.* 2019.
134. Choi IA, Lee JS, Song YW, Lee EY. Mortality, disability, and healthcare expenditure of patients with seropositive rheumatoid arthritis in Korea: A nationwide population-based study. *PloS one.* 2019;14(1).
135. Kim K, Choi S, Chang J, et al. Severity of dental caries and risk of coronary heart disease in middle-aged men and women: a population-based cohort study of Korean adults, 2002–2013. *Scientific reports.* 2019;9(1):1-7.
136. Hwang IC, Chang J, Kim K, Park SM. Aspirin use and risk of hepatocellular carcinoma in a national cohort study of Korean adults. *Scientific reports.* 2018;8(1):1-9.
137. Rim TH, Kim HS, Kwak J, Lee JS, Kim DW, Kim SS. Association of corticosteroid use with incidence of central serous chorioretinopathy in

South Korea. *JAMA ophthalmology.* 2018;136(10):1164-1169.
138. Leucht S, Corves C, Arbter D, Engel RR, Li C, Davis JM. Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. *The Lancet.* 2009;373(9657):31-41.
139. Choi S, Kim KH, Kim K, et al. Association between Post-Diagnosis Particulate Matter Exposure among 5-Year Cancer Survivors and Cardiovascular Disease Risk in Three Metropolitan Areas from South Korea. *International journal of environmental research and public health.* 2020;17(8):2841.
140. Hogle LF. Data-intensive resourcing in healthcare. *BioSocieties.* 2016;11(3):372-393.
141. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *European heart journal.* 2017;38(7):500-507.
142. Lloyd-Jones DM, Wilson PW, Larson MG, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. *The American journal of cardiology.* 2004;94(1):20-24.
143. Pagali SR, Madaj P, Gupta M, et al. Interobserver variations of plaque severity score and segment stenosis score in coronary arteries using 64 slice multidetector computed tomography: a substudy of the ACCURACY trial. *Journal of cardiovascular computed tomography.* 2010;4(5):312-318.
144. Ayoub C, Erthal F, Abdelsalam MA, et al. Prognostic value of segment involvement score compared to other measures of coronary atherosclerosis by computed tomography: a systematic review and meta-analysis. *Journal of cardiovascular computed tomography.* 2017;11(4):258-267.
145. Lin F, Shaw LJ, Berman DS, et al. Multidetector computed tomography coronary artery plaque predictors of stress-induced myocardial ischemia by SPECT. *Atherosclerosis.* 2008;197(2):700-709.
146. Benjamin EJ, Levy D, Vaziri SM, D'Agostino RB, Belanger AJ, Wolf PA. Independent risk factors for atrial fibrillation in a population-based cohort: the Framingham Heart Study. *Jama.* 1994;271(11):840-844.
147. Psaty BM, Manolio TA, Kuller LH, et al. Incidence of and risk factors for atrial fibrillation in older adults. *Circulation.* 1997;96(7):2455-2461.
148. Jie M, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology.* 2019.
149. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Network Open.* 2020;3(1):e1918962-e1918962.

국 문 초 록

배경 및 목적: 심혈관질환 위험평가 및 예측모델링에서 다양한 심혈관질환 위험인자들의 모델 성능향상에 대한 기여도는 논란의 요지로 보고되어왔다. 또한, 지속적으로 증가하는 활용 가능한 심혈관질환 관련 데이터의 종류와 양에도 불구하고 포괄적인 심혈관질환 위험평가와 최적의 예측 모형 개발을 위해 데이터를 어느 범위와 수준까지 수집해야 하는지에 대한 근거는 부족한 현황이다. 본 연구에서는 콕스 모형과 결합된 딥러닝 기반 생존분석 접근법 및 콕스 모형을 활용한 심혈관질환 위험평가와 예측모델링에서 헬스케어-환경 연계 데이터 활용방법 및 범주에 따른 모델 성능향상에 대한 기여도를 평가하고자 하였다.

연구 방법: 전통적 심혈관질환 위험 평가 도구 및 관찰 연구들에 포함된 심혈관질환 위험요인 관련 변수들을 체계적 문헌고찰 방법론을 활용하여 의학연구 문헌데이터베이스 (PubMed and Embase)에서 포괄적으로 정보를 수집하였다. 미세먼지 누적장기노출 및 도시녹지면적에 대한 환경 노출 데이터와 연계 된 국민건강보험공단 표본코호트, (National Health Insurance Service-National Sample Cohort, NHIS-NSC)에서 각 심혈관질환 위험인자들의 데이터 확보 가능성을 검토하였다. NHIS-NSC를 기준으로 2009년에서 2010년 사이에 국가건강검진을 받은 40세 이상 대상자 중 과거 심혈관질환 병력이 없는 대상자 137,249명의 환자에 대한 정보를 수집하여 2011 년 1 월 1 일부터 2013 년 12 월 31 일까지 신규 발생한 심혈관질환에 대해 시간 경과에 따라 추적 조사하였다. 통계 기반 변수선택 방법은 콕스비례위험모형에서 통계적 유의성, 최소 (최상의) Akaike Information Criteria (AIC)의 하위 집합, elastic net penalty로 정규화 된 콕스비례위험모형에서 선택된 변수 및 위에 언급된 모든 기준을 충족하는 변수 세트로 선정하였다. 위에 명시된 통계적 방법 외 모든 데이터 범주에 속한 변수 및 콕스비례위험모형에서 통계적으로 유의미한 변수 (하이브리드 접근법)를 점진적으로 입력

피쳐로 추가하는 전략으로 딥러닝 기반 생존분석 (Cox proportional hazards deep neural network, DeepSurv) 및 콕스비례위험모형에서 예측 모델들을 훈련 세트 (전체 샘플의 80 %)를 기반으로 개발하였다. DeepSurv 및 콕스비례 위험모형을 활용한 심혈관질환 예측 모델의 성능평가는 생존분석을 활용한 예측 모델링에 가장 적합한 평가지표로 알려진 Uno's concordance statistics (C-index)를 사용하여 테스트 세트 (총 샘플의 20 %)에서 수행하였다.

결과: 체계적 문헌고찰, 데이터 취합 및 추출 가능성 검토 후, 인구사회학적 요인, 건강검진 및 측정 결과, 생활습관, 가족력, 건강상태, 구강건강, 약물 및 환경 노출 데이터 범주에서 총 31 개의 심혈관질환 위험인자가 지역환경 자료와 연계된 NHIS-NSC에서 확인되었다. 통계 기반 변수선택 방법으로 개발한 심혈관질환 예측 모델 중 콕스비례위험모형에서 통계적으로 유의미한 변수를 DeepSurv에 적용한 하이브리드 접근법이 Uno 's C-index 값 0.7069, 모든 변수를 콕스비례위험모형에 적용한 콕스비례위험모형이 Uno 's C-index 값 0.7052로 나타나 기본 임상 요인 (연령, 성별 및 체질량지수)이 포함된 예측 모델과 비교하여 통계적으로 유의미한 모델 예측력 증가를 보였다 (두 모델 모두 Uno's C-index 차이에 대한 p-value : <0.0001). 인구사회학적 특성에서 환경 노출에 이르기까지 각 데이터 범주에서 모두 통계적으로 유의미한 변수들이 심혈관질환 예측 모델링을위한 DeepSurv 및 Cox 비례 위험 회귀에 입력 피쳐로 점진적으로 추가 된 경우, 인구사회학적 요인, 건강검진 및 측정 결과, 생활습관 요인 중 통계적으로 유의미한 변수들로 구성된 DeepSurv 모델이 의약품 사용까지 고려한 Cox 비례 위험 회귀를 기반으로 한 모델 보다 뛰어난 성능을 나타냈다. 미세먼지 및 도시녹지 면적에 대한 환경 노출 데이터를 거주지를 기반으로 NHIS-NSC와 연계 후 점진적으로 입력 피쳐로 추가 시 DeepSurv 및 콕스비례위험모형을 활용한 심혈관질환 예측 모델링 성능을 통계적으로 유의미한 수준으로 개선하지 못했다.

결론: 최소 입력 피쳐를 갖춘 생존 분석 기반 심혈관질환 예측 모델에서 최상의 성능을 얻으려면 인구사회학적, 건강검진 및 측정 결과, 및 생활 습관에 대한 정보를 NHIS-NSC에서 수집하여 DeepSurv의 입력 피쳐로 활용해야한다. 지역환경 자료와 연계된 NHIS-NSC에서 모든 데이터 범주를 사용할 수 있을 때 점진적으로 각 데이터 범주 중 콕스비례위험모형에서 통계적으로 유의미한 심혈관질환 위험인자를 점진적으로 입력 피쳐로 DeepSurv 모델에 추가하는 하이브리드 접근법에서 심혈관질환 예측 모델링 성능이 점차 향상 될 것으로 기대할 수 있다. 주거 지역 코드를 사용한 NHIS-NSC와 환경 노출 데이터 연계는 DeepSurv 및 콕스비례위험모형 모두에서 심혈관질환 예측 모델링 성능이 향상되었지만 통계적으로 유의미한 증가 수준은 아닌 것으로 나타나 환경 노출 데이터 연계 및 적용 시 검토가 필요할 것으로 추정된다.

주요어: 심혈관질환; 헬스케어 데이터; 환경 노출; 딥러닝 기반 생존 분석; 콕스비례위험모형

학번: 2016-21973