Journal of Economic Theory and Econometrics, Vol. 34, No. 4, Dec. 2023, 86–114

Long-term Predictors of Cardiovascular Disease: A Machine Learning Approach *

Young-Joo Kim †

Abstract This study investigates long-term cardiovascular disease (CVD) risk predictors for middle-aged and older adults in Korea. Using the Least Absolute Shrinkage and Selection Operator (Lasso) and the double-selection Lasso, this study provides novel evidence that Body Mass Index (BMI) is a single risk factor with long-term predictability for CVD odds ratio, selected apart from age, which is non-modifiable. The lasting effect of BMI on CVD risk remains robust and consistent across different methods and specifications that account for variable selection errors in high-dimensional logit regression and BMI's time trends. In addition to the long-term predictive role of BMI in CVD risk, the disease burden associated with increased BMI is quantified by comparing the marginal effects of BMI to those of age across various groups. The marginal effect of elevated BMI is more pronounced in men than women and among the employed compared to the non-employed. Leading a healthy lifestyle through the control of BMI is a critical element for preventing CVD based on the empirical findings of the current study.

Keywords BMI, CVD, Lasso, odds ratio, marginal effect

JEL Classification I12, C55

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea [NRF- 2020S1A5A2A03046422].

†Department of Economics, Hongik University, Wausanro 94, Mapogu, Seoul, Korea, Korea; y.j.kim@hongik.ac.kr

Submitted November 2, 2023, Accepted December 18, 2023

INTRODUCTION

Cardiovascular disease (CVD) is a chronic disease associated with blood vessels and the heart, and it is a leading cause of disability and death around the globe (WHO, 2021). CVD includes coronary heart disease, heart failure, stroke, and peripheral arterial disease, and it has a long-term effect on health and quality of life once the onset of the disease (NHS, 2022).

With the growing proportion of the aged population over the last few decades, the prevalence of CVD has dramatically increased. For example, the number of significant heart-related disease incidents in Korea has increased from 1.38 million to 1.62 million between 2016 and 2020 (Health Insurance Review and Assessment Service, 2021). The consequent increases in healthcare costs and lost productivity due to CVD an individual and society has to bear has increased accordingly. In Korea, more than 0.1% of the GDP was spent on treating and managing heart-related diseases and cerebrovascular disease in 2018 (RIHP, 2020; Bank of Korea, 2020).

One way of reducing the disease burden is to classify high-risk group of individuals likely to develop the disease early and to provide appropriate interventions. We are able to reach this goal first by identifying risk factors that determine the onset of illness among the potential factors that describe an individual's health status, behavior, and other demographic and socioeconomic characteristics associated with CVD. Since CVD has a long incubation period, it is critical to establish risk factors, in particular, that have long-term predictability before diagnosing the disease.

Numerous studies have made constant efforts to disclose what contributes to CVD. See, for example, Dahlöf (2010), Forouhi and Sattar (2006), Kim (2021), Ouyang *et al.* (2022), and Qian *et al.* (2022). Among various factors, obesity is presented as one of the risk factors (Bays *et al.*, 2021; Kim, 2021; Lavie *et al.*, 2009). Since obesity is measured by body mass index (BMI), defined as weight in kilograms divided by height in meters squared (kg/m2), the relationship between BMI and CVD risk has been explored by many studies. However, the evidence on the long-term predictability of BMI on CVD incidence is scarce as few studies follow up over ten years, focusing on Western Europe and North America (Berrington *et al.*, 2010; Fernstrom *et al.*, 2020; PSC, 2009). Several studies that examined Korean data from hospital-recruited or nationwide cohort studies also followed only short-term or less than ten years (Kim *et al.*, 2008; Kim *et al.*, 2017; Lee and Lee, 2018; Yeoum, 2003).

In exploring risk factors of CVD, the analyses are often restricted to anthropometric and biochemical factors or health-related behaviors, leaving other

aspects of life unexplained for predicting the disease. With accumulated large-scale data that covers individual's behavioral, biological, environmental, and socioeconomic factors, we are able to explore this issue, but there arise technical problems. To select only relevant determinants out of a vast set of data in predicting disease incidence, we first need to consider the possibility that the number of individual-specific variables is larger compared to the sample size. Having many candidate predictors is especially the case for the survey data, where detailed responses are recorded with categorical variables from which many binary variables are generated. The second issue is how to derive valid inferences on the selected variable. For constructing confidence regions of a regression parameter of the selected variable, we need to take into account potential variable selection errors in a high dimensional non-linear setting of logit regression.

The present study examines potential CVD risk factors, addressing the challenging issues described above. Using the panel data of middle-aged and older Koreans that traces over twelve years, we investigate what aspects of individual characteristics are associated with CVD incidence and examine the long-term predictability of the selected variables. Incorporating group-level heterogeneity, we also examine by sex and employment status. We then estimate the marginal effects of selected predictors, controlling for biases arising from variable selection errors. In answering these questions, we use the Least Absolute Shrinkage and Selection Operator (Lasso) estimation method proposed by Tibshirani (1996) and the double-selection Lasso developed for the non-linear binary model proposed by Belloni *et al.* (2016).

DATA

KLOSA

The data for this study comes from the Korean Longitudinal Study of Aging (KLoSA). The KLoSA is a nationally representative longitudinal survey of Koreans aged 45 years or older and is available through public repositories (https://survey.keis.or.kr/eng/klosa/klosa01.jsp). The survey participants were randomly selected nationwide except from Jeju Island from each of five divided geographic regions, proportional to the local population. By following up with respondents every two years over time, the KLoSA aims to collect micro-level panel data for the study of health, social and economic conditions, and activities of older adults, which help to develop socioeconomic policies that address comprehensive life-cycle characteristics of older adults in preparation for an aging society (KEIS, 2022). Computer-assisted personal interviews have collected in-

dividual and household-level information on sociodemographic characteristics, physical and mental health, labor market outcomes, and economic status.

This study uses the second and eighth follow-up surveys collected in 2008 and 2020. The first survey was not used as the monthly allowance and household assets information was not collected until the second survey. The year 2008 will be referred to as the baseline throughout this paper. The baseline survey has a rich data collection on health and economic status, and the follow-up survey in 2020 has the most recent information of the respondents. The final sample included 3,104 adults who were aged 49 years and older as of 2008, excluding the respondents who had missing values for health and other characteristics.

MEASURES

The primary outcome of interest is cardiovascular disease (CVD) incidence observed in 2020. In this study, CVD is defined to cover heart diseases, hypertension, diabetes, and cerebrovascular disease based on diagnosed chronic disease information in KLoSA. To predict CVD incidence observed in 2020, we examine individuals' health and other characteristics collected in the second wave, 2008. Our final sample of 3,104 adults is those who have not developed CVD in the second wave as of 2008 to exclude incumbent CVD patients and explore potential risk factors for the future development of the disease.

For CVD predictors, individual and household-level characteristics are considered. For an anthropometric measure, body mass index (BMI), a primary measure of obesity, is used. BMI is measured as weight in kilograms divided by height in meters squared. A BMI between 18.5 and 23 indicates healthy weight, a BMI between 23 and 25 for overweight, and a BMI greater than or equal to 25 indicates obesity.

Demographic and socioeconomic statuses have been measured by education level, household size, housing type (apartment or detached house), household asset and income, employment status, occupation details, and work hours from the primary and supplemental jobs. Health status and health-related behaviors are measured by self-rated health status, indicators of alcohol drinking, smoking, and doing regular exercise, and the device-rated handgrip strength. For cognitive ability, Mini-Mental State Examination (MMSE) scores were used. Other health-related measures are types of national health insurance, private health insurance participation, and completion of national health examinations over the past two years. The parental living statuses of the father and mother are used as a family health history indicator. For a robustness check of CVD predictors, we also classified individuals who were not diagnosed with mental health disorders

or other chronic diseases related to lung, liver, prostate, rheumatoid arthritis, and all types of cancers.

For social activity, religion and frequency of meeting close friends, all measured as categorical variables, are used. The life satisfaction over the overall quality, health status, and economic status are measured with life satisfaction scores that range between 0 and 100, with higher values indicating higher satisfaction levels. Regions of residence and regional sizes are also included in the study. The final sample contains 106 variables that capture these detailed characteristics.

METHODS

Our study employs the Least Absolute Shrinkage and Selection Operator (Lasso) estimation method to select the most influential factors in predicting CVD incidence and conduct inference. The Lasso is a commonly employed machine learning technique for distilling vital variables from a vast array of potential candidates, particularly in the context of big data. The data that contains numerous categorical variables, often prevalent in survey data, including the KLoSA, present a specific challenge in this process. With such big data, the Lasso has emerged as an effective screening tool for assessing CVD risk predictors in recent academic literature (Ouyang *et al.*, 2022; Qian *et al.*, 2022; Shen *et al.*, 2023).

When predicting CVD incidence based on a large set of covariates, only a few predictors may be valuable, rendering the majority irrelevant. Such a predictive model, in which coefficients are non-zero for only a small number of predictors, is referred to as a sparse model. The ordinary least squares (OLS) estimator faces limitations when selecting predictors from an extensive pool of candidates as the number of predictors is larger relative to the sample size. For sparse models, model simplicity and prediction accuracy can be achieved by imposing the penalty to shrink some estimated coefficients to be exactly 0.

The Lasso estimator minimizes the penalized sum of the negative log-likelihood function, which can be computed rapidly thanks to its convex nature. The penalty term of the Lasso estimator is the sum of the absolute values of the coefficients, and the penalty level is chosen by cross-validation.

That is, for $p(X,b) = 1/(1+\exp(-b_0 - b_1 X_{1i} - \ldots - b_k X_{ki}))$,

$$S^{Lasso}(b;\lambda_{Lasso}) = -\sum_{i=1}^{n} Y_i ln(p(X_i,b)) - \sum_{i=1}^{n} (1-Y_i) ln(1-p(X_i,b)) + \lambda_{Lasso} \sum_{j=1}^{k} \left|b_j\right|,$$

where λ_{Lasso} is the Lasso shrinkage parameter, Y_i is CVD incidence for individual i observed at the eighth period, and $X_i = \{X_{ji},\ j = 1,\ldots,k\}$ is the collection of all individual characteristics observed at the second period.

Our study employs the Lasso method to select variables and construct confidence intervals. Additionally, we introduce a robust post-selection inference procedure for the coefficients of primary interest. This approach, suggested by Belloni *et al.* (2016), is designed to mitigate potential biases that may arise from variable selection errors in the high-dimensional logistic regression, albeit it might not enhance efficiency. For this approach, an optimal value of the penalty parameter is chosen by their plugin iterative formula.

The robust procedure is based on the double-selection principle, initially proposed by Belloni *et al.* (2014) for high-dimensional linear regression to estimate treatment effects. This principle has been extended to various contexts since then, as reviewed by Chernozhukov *et al.* (2018). The double-selection principle is recognized for reducing biases in post-selection inference when machine learning methods are used for econometric inference.

We next apply the preceding approach with an extended predictor set to encompass variables observed in the second and fourth periods for a robustness check of findings from the prior approach. Controlling for these factors observed during the interim period in the long-term predictive regression allows us to examine the potential channels under which the selected predictor in the baseline year affects the CVD incidence in later years.

RESULTS

DESCRIPTIVE CHARACTERISTICS OF THE RESPONDENTS

Some of the candidate predictive variables are presented in Table 1. The final sample includes 1,336 men and 1,768 women, whose average age in the baseline year is 58. The average BMI of the sample is 23.031, which is close to the overweight threshold.

When the sample is stratified by sex, we see that men and women are similar in various aspects except for the following variables. These are education level, self-assessed health status, life satisfaction scores, handgrip strength, and health-related behaviors of drinking, smoking, and exercising. On average, men tend to have higher education levels than women, reported higher life satisfaction scores, and have higher handgrip strength. On the other hand, men are more likely to drink alcohol, smoke, and engage in regular exercise than women. All individuals were free of CVD in the baseline year.

SELECTED RISK FACTORS OF CVD

We obtain predictors of CVD risk from the Lasso and present the selected predictors' marginal effects on the log odds ratios from the logit regression in Table 2. As shown in column 1, thirteen variables out of 106 are selected from the full sample. These are age, BMI, education level, household size, indicators of how often to meet close friends, having a mother alive, life satisfaction scores on health and overall life, and a subset of regional indicators. For sensitivity analysis, we next restrict the sample to those who were not diagnosed with any chronic diseases, including both CVD and non-CVD or mental disorders in the baseline year. This restriction confines the sample to a healthy group of 2,447 individuals. A similar set of variables is selected for this group, but health-related variables, including indicators of no alcohol drinking and self-rated health status, are newly added.

We next focus on individuals aged 75 or younger and then aged 65 or younger subsequently. The respondents' ages in the entire sample in the baseline year ranged from 47 to 86, and these age restrictions excluded 95 and 726 individuals from the sample, respectively. Focusing on relatively younger groups of individuals yields similar and smaller sets of predictors that are primarily comprised of age, BMI, education level, life satisfaction scores, having a mother alive, health status, and regional dummies.

In Table 3, we explore heterogeneity in the set of risk predictors by sex and employment status. For men, seven variables, including age, BMI, university education level, household income, handgrip strength, having a mother alive, and a regional dummy, are selected as shown in column 1. For women, on the other hand, household size, indicators of meeting friends, life satisfaction scores, no smoking, and regular exercise are further selected in addition to age, BMI, and education levels as listed in column 2. For the following two columns 3 and 4, we examine how risk factors differ depending on baseline labor market activity. For the employed, age, BMI, education levels, household income, having a mother alive, and region of residence matter, but life satisfaction score or social interactions with friends are not selected anymore. For the non-employed, life satisfaction scores and social interactions within and outside the household, measured by household size and how frequently they meet friends, are estimated to be important predictors in addition to age, BMI, and education levels.

One of the critical findings in Table 3, where we explore heterogeneity across different groups of individuals by sex and employment status, is that only age and BMI are estimated to be significant predictors of CVD for all groups. In particular, BMI is the single risk factor other than age with long-term CVD incidence

predictability.

Focusing on BMI, we now compare the BMI effect across the subsamples. For men, the estimated marginal effect of BMI on the log odds ratio of CVD is 0.180. This BMI effect is six times the size of the age effect for this group. That is, the marginal effect of BMI associated with a one-unit increase in BMI corresponds to a six-year age effect. For women, the marginal effect of BMI on the log odds ratio is 0.125, about four times the age effect. The marginal effects of BMI for the employed and the non-employed are 0.171 and 0.114, respectively, and the magnitudes of these BMI effects are about four and three times the age effect for each group.

MARGINAL EFFECTS OF BMI FROM DOUBLE-SELECTION LASSO

As we found that BMI is the single risk factor with long-term CVD predictability, we now want to draw inferences on the marginal effect of BMI on the log odds ratio of CVD. For this step, we select another set of variables using Lasso, but this time, we focus on potential controls that are associated with BMI. By including the selected controls in the model through the double-selection procedure proposed by Belloni *et al.* (2016), we consider bias arising from the correlation between BMI and individual and household characteristics that are excluded by the first step Lasso. The estimated marginal effects of BMI from the double-selection Lasso logit regression are presented in Table 4 for the full and subsamples. First, we examine the full sample in column 1 for which the selected controls are age and handgrip strength. As shown in column 1, the estimated marginal effect of BMI is 0.151, which is slightly bigger than 0.141, the marginal effect of BMI from the logit regression with predictors in Table 2. We can interpret that a one-unit increase in BMI is associated with an increase in the log odds ratio of CVD incidence by 0.151. Converting this marginal effect into exponential function specification, we can also interpret that a one-unit increase in BMI leads to an increase in the CVD odds ratio by 1.163 in twelve years. In columns 2 to 4 of Table 4, we present the marginal effects of BMI for the restricted samples, and the estimates range between 0.163 and 0.158. The double-selection method estimates are slightly larger than those in Table 2, and they are all statistically significant.

Following the sample splitting specifications in Table 3, we also estimate the marginal effect of BMI from the double-selection procedure for each subsample and present the results in columns 5 through 8 in the lower panel of Table 4. We find that the marginal effect of BMI ranges from 0.194 to 0.119 across the subsamples of men, women, the employed, and the non-employed, and all of

the estimates are statistically significant as in the full sample. The estimated marginal effects of BMI are again slightly larger than those in Table 3, indicating that the bias from errors in variable selection is small and negative, if any.

ROBUSTNESS CHECK OF THE BMI EFFECTS

Based on the finding that BMI is a significant long-term predictor of CVD incidence, we now include variables from the fourth survey and select predictors from the combined second and fourth surveys. The fourth survey point will be referred to as an interim period. By looking at the significance of the baseline BMI in this extended long-term predictive regression, we can examine potential channels under which the baseline BMI affects CVD incidence in later years. For instance, if the baseline BMI effect on later CVD incidence is explained away by the factors observed during the interim period, only short-term prior BMI predicts CVD risk. On the other hand, if the baseline BMI is still a significant predictor even after controlling for the interim BMI, the effect of the baseline higher BMI is not limited to the channel that raises future BMI but has a permanent effect on CVD risk.

The selected predictors out of 207 variables from the second and fourth surveys are age, BMI, education level, regional dummies, indicators of meeting friends, self-rated health, and household income, consistent with the main results in Table 2. Among the selected, the BMI measures from both the baseline year and the interim year are selected and statistically significant. As we split the sample by sex and employment status, we again find that age and BMI are the only risk predictors selected across all groups: men, women, employed, and non-employed. For comparison with prior results, we report the estimated marginal effects of BMI from the baseline and interim years from the double-selection Lasso logit regression in Table 5. Due to missing observations from the combined set of two survey data, the sample size becomes smaller for each subsample in Table 5.

The results from the full sample are presented in column 1. The marginal effect of baseline BMI on the log odds of the CVD incidence is 0.063, about 42 percent of the marginal effect from the model without the interim period BMI in column 1 of Table 4. The marginal effect of the interim period BMI is 0.119 and statistically significant. Thus, the combined marginal effects of BMI in the baseline and interim periods are summed to 0.1819, greater than the marginal effect of a single BMI from the baseline year. Overall, with the extended data, the estimated marginal effect of baseline BMI becomes reduced, and BMI in the interim year becomes a predominant factor, but the combined effect of BMI

from the baseline and the interim year is larger than the baseline BMI effect from Table 4.

Our results indicate that the interim year's BMI explains part of the baseline BMI effect, but the baseline BMI has lasting effects on the CVD odds ratio. When we regress the interim period's BMI on the baseline year's BMI, the autoregression coefficient is estimated to be 0.829 with a standard error of 0.011. We can see a strong correlation and persistence of BMI over time, but the initial BMI has a significant and lasting effect on CVD risk over twelve years.

DISCUSSION

This study aims to identify potential long-term predictors of CVD risk in the big data setup. By looking at an extensive set of candidate risk factors over twelve years of follow-up, this study presents novel evidence that BMI is the single risk factors, other than age, that is selected to have long-term predictability of CVD incidence. We confirm the lasting effect of BMI on CVD risk across specifications and methods with which we take account of a strong autocorrelation of BMI over the years and biases arising from variable selection errors in high dimensional logit regression. Considering that age is a non-modifiable factor, our findings indicate that BMI is an important and the only element among observable and measurable variables in this study that an individual's behavior can change to prevent the disease.

Previous studies examined predictors of CVD risk for adults in Korea and other parts of the world. The conventional short-term predictors discussed in the literature are biomarkers such as elevated triglyceride levels, higher fasting insulin and glucose levels, high systolic blood pressure, and lower high-density lipoprotein-cholesterol (Cho, 2020; Fernstrom *et al.*, 2020; Forouhi and Sattar, 2006). The long-term predictors, on the other hand, are behavioral factors that affect the conventional biomarkers such as a sedentary lifestyle, lack of physical activity, insufficient sleep duration, increased calorie intake, and obesity (Fernstrom *et al.*, 2020; Forouhi and Sattar, 2006; Lavie *et al.*, 2009; McKeigue *et al.*, 1991). The present study demonstrates that the long-term predictive factors can be summarized into BMI, a primary measure of obesity. It would be an interesting future research topic to investigate if BMI is singled out as a representative measure among behavioral, lifestyle, and socioeconomic factors that increase the risk of CVD.

The strength of the present study also lies in quantifying the disease burden associated with an increased BMI. By comparing the marginal effects of BMI

relative to the age effect for each group of interest, we show heterogeneity in the BMI effect across the groups. The marginal effect of an increased BMI is higher for men than women and higher for the employed than the non-employed such that a one-unit increase in BMI corresponds to a six-year age effect for men, a four-year age effect for women and the employed, and a three-year age effect for the non-employed. Therefore, we can identify the at-risk group with a higher disease burden associated with an increased BMI.

Nevertheless, our study has some limitations. As we follow up with adults who were free of CVD in baseline and have survived over long years, we exclude those who may have passed away from CVD. Second, as we focus on predictors of disease risk, not causal factors, the interpretation of the estimates is limited to predictive factors, although we used long lags between BMI and CVD incidence to exclude the possibility of reverse causality. Despite these limited aspects, we contribute to the literature by presenting solid evidence on the most important risk factors that have long-term predictability of CVD, resolving many-predictor problems typical in the conventional survey data with many categorical and binary variables. Further, our findings are applicable to the general population in Korea as our analyses are drawn from a nationally representative set of middle-aged and older Koreans.

CONCLUSION

An increased proportion of the older adult population, in combination with the obesity epidemic, raises the burden of CVD that our society has to bear in the current and future generations. As a precautionary step, developing a less invasive and less expensive health assessment tool that can be easily practiced early before disease onset is desirable to prevent CVD effectively. This study provides new evidence on the long-term predictability of CVD by showing that BMI is a reliable early indicator to detect middle-aged and older adults with an increased risk of developing CVD later in life. The findings of this study have policy implications that health interventions encouraging middle-aged and older adults to maintain a healthy body weight range through close monitoring of BMI over a long horizon can reduce CVD prevalence.

REFERENCES

Myerson, R. (1981). "Optimal auction design," *Mathematics of Operations Research*, 6, 58–73.

Myerson, R. and M. Satterthwaite (1983). "Efficient mechanisms for bilateral trading," *Journal of Economic Theory*, 29, 265–281.

Rustichini, A., Satterthwaite, M., and S. Williams (1994). "Convergence to efficiency in a simple market with incomplete information," *Econometrica*, 62, 1041–1063.

Bank of Korea (2020). "Optimal auction design," *National Accounts 2018 Final and 2019 Preliminary*, June 24.

Bays, H. E., Taub, P. R., Epstein, E., Michos, E. D., Ferraro, R. A., Bailey, A. L., ..., and P.P. Toth (2021). "Ten things to know about ten cardiovascular disease risk factors," *American Journal of Preventive Cardiology*, 5, 100149.

Belloni, A., Chernozhukov, V., and C. Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 81(2), 608–650.

Belloni, A., Chernozhukov, V., and Y. Wei (2016). "Post-selection inference for generalized linear models with many controls," *Journal of Business & Economic Statistics*, 34(4), 606–619.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and J. Robins (2018). "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

Cho, S. O. (2020). "Assessment of the contribution of risk factors that cause cardiovascular disease in Koreans," *Journal of the Korea Academia-Industrial cooperation Society*, 21(6), 592–602.

Dahlöf, B. (2010). "Cardiovascular disease risk factors: epidemiology and risk assessment," *The American Journal of Cardiology*, 105(1), 3A–9A.

Berrington de Gonzalez, A., Hartge, P., Cerhan, J. R., Flint, A. J., Hannan, L., MacInnis, R. J., ..., and M. J. Thun (2010). "Body-mass index and mortality among 1.46 million white adults," *New England Journal of Medicine*, 363(23), 2211–2219.

Fernstrom, M., Fernberg, U., and A. Hurtig-Wennlof (2020). "The importance of cardiorespiratory fitness and sleep duration in early CVD prevention: BMI, resting heart rate and questions about sleep patterns are suggested in risk assessment of young adults, 18–25 years," *BMC Public Health*, 20(1), 1–11.

Forouhi, N. G. and N. Sattar (2006). "CVD risk factors and ethnicity—a homogeneous relationship?," *Atherosclerosis Supplements*, 7(1), 11–19.

Health Insurance Review and Assesment Service (2021). "Health Insurance Review and Assessment Service," *Statistics on Heart Diseases*, September, 2021

Kim, C. G., LEE, S. H., and S. K. Cha (2017). "Influencing factors on cardiocerebrovascular disease risk factors in young men: focusing on obesity indices," *Journal of Korean Biological Nursing Science*, 1–10.

Kim, H. C. (2021). "Epidemiology of cardiovascular disease and its risk factors in Korea," *Global Health & Medicine*, 3(3), 134–141.

Kim, J. H., Choi, S. R., Lee, J. R., Shin, J. H., Lee, S. J., Han, M. A., ..., and S. Y. Kim (2008). "Association of hemoglobin A1c with cardiovascular disease risk factors and metabolic syndrome in nondiabetic adults," *Korean Diabetes Journal*, 32(5), 435–444.

Korea Employment Information Service (2022). "Health Insurance Review and Assessment Service," *Korea Longitudinal Study of Aging User Guide for first through eight surveys*, 1-110.

Lavie, C. J., Milani, R. V., and H. O. Ventura (2009). "Obesity and cardiovascular disease: risk factor, paradox, and impact of weight loss, 18–25 years," *Journal of the American college of cardiology*, 53(21), 1925–1932.

Lee, J. B. and D. H. Choi (2021). "Statistics on Heart Diseases in Korea," *Health Insurance Review and Assessment Service*, 1–11.

Lee, K. H. and S. B. Lee (2018). "Effect of lifestyle on cardiovascular risk in 10 years according to Framingham risk score of middle-aged women—The based on 2016 Korea National Health and Nutritional Examination Survey," *Korea Soc. Wellnes*, 1, 77288.

McKeigue, P. M., Shah, B., and M. G. Marmot (1991). "Relation ofcentral obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians," *The Lancet*, 337(8738), 382–386.

National Health Service (2022). "Cardiovascular disease, Health A to Z".

Ouyang, N., Li, G., Wang, C., and Y. Sun (2022). "Construction of a risk assessment model of cardiovascular disease in a rural Chinese hypertensive population based on lasso-Cox analysis," *The Journal of Clinical Hypertension*, 24(1), 38–46.

Prospective Studies Collaboration (2009). "Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies," *The Lancet*, 373(9669), 1083–1096.

Qian, X., Li, Y., Zhang, X., Guo, H., He, J., Wang, X., Yan, Y., Ma, J., Ma, R., and S. Guo (2022). "A cardiovascular disease prediction model based on routine physical examination indicators using machine learning methods: a cohort study," *Frontiers in Cardiovascular Medicine*, 9, 854287.

Korea Employment Information Service (2020) "Table 3-13".

Shen, X., He, W., Sun, J., Zhang, Z., Li, Q., Zhang, H., and M. Long (2023). "Development and Validation of a Nomogram to Predict the Future Risk of Cardiovascular Disease," *Reviews in Cardiovascular Medicine*, 24(2), 35.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.

World Health Organization (2021) "Cardiovascular diseases (CVDs)".

Yeoum, S. G. (2003). "The Investigation of the Risk Factors of Cardiovascular Disease for Postmenopausal Women Over 50 Years,," *Journal of Menopausal Medicine*,9(3).

Table 1: Descriptive Characteristics of the KLoSA Participants in the baseline year 2008

Variables	All N=3,104 M (SD)	Men N=1,336 M (SD)	Women N=1,768 M (SD)
Age	58.511	59.070	58.088
	(8.431)	(8.368)	(8.456)
BMI	23.031	23.089	22.987
	(2.437)	(2.230)	(2.581)
Education level			
Elementary school or	0.351	0.246	0.430
below	(0.477)	(0.431)	(0.495)
Middle school	0.183	0.166	0.196
	(0.387)	(0.372)	(0.397)
High school	0.346	0.395	0.309
	(0.476)	(0.489)	(0.462)
College or higher	0.119	0.192	0.064
	(0.324)	(0.394)	(0.245)
Household size	2.984	3.057	2.929
	(1.259)	(1.208)	(1.294)
Household income†	2952.3	3048.0	2880.0
	(2913.4)	(2765.6)	(3019.0)
Handgrip strength	25.996	32.928	20.758
	(7.857)	(5.922)	(4.260)
Life scores-overall	64.275	65.494	63.354
	(17.11)	(16.87)	(17.23)
Life scores-health	62.355	65.614	59.893
	(19.03)	(17.59)	(19.71)
Life scores-economic	54.436	55.921	53.314
	(20.80)	(20.58)	(20.90)
MMSE Score	26.979	27.582	26.523
	(3.413)	(2.748)	(3.778)
Meeting close friends			
Two to three times a	0.137	0.127	0.144
week	(0.344)	(0.333)	(0.351)
Two times a month	0.064	0.076	0.055
	(0.246)	(0.266)	(0.229)

Table 1: Descriptive Characteristics of the KLoSA Participants in the baseline year 2008

Variables	All N=3,104 M (SD)	Men N=1,336 M (SD)	Women N=1,768 M (SD)
Three to four times a year	0.033 (0.177)	0.031 (0.175)	0.033 (0.180)
Married	0.874 (0.332)	0.944 (0.230)	0.822 (0.383)
No Alcohol drink	0.519 (0.500)	0.210 (0.408)	0.753 (0.431)
No Smoke	0.709 (0.454)	0.359 (0.480)	0.973 (0.163)
Regular exercise	0.383 (0.486)	0.400 (0.490)	0.369 (0.483)
Self-rated health status			
Fair	0.268 (0.443)	0.227 (0.419)	0.299 (0.458)
Bad	0.091 (0.288)	0.052 (0.223)	0.120 (0.326)
Parental living status			
Mother alive	0.281 (0.450)	0.293 (0.455)	0.271 (0.445)
No parent alive	0.597 (0.491)	0.577 (0.494)	0.613 (0.487)
Region of residence			
Seoul	0.138 (0.345)	0.130 (0.337)	0.144 (0.351)
Daegu	0.066 (0.248)	0.063 (0.243)	0.068 (0.252)
Gyunggi	0.135 (0.341)	0.134 (0.341)	0.135 (0.342)
Gangwon	0.042 (0.200)	0.045 (0.207)	0.040 (0.195)
Chungbuk	0.043 (0.202)	0.046 (0.210)	0.040 (0.195)

Table 1: Descriptive Characteristics of the KLoSA Participants in the baseline year 2008

Variables	All N=3,104 M (SD)	Men N=1,336 M (SD)	Women N=1,768 M (SD)
Chungnam	0.073	0.073	0.074
	(0.261)	(0.260)	(0.262)
Regional type			
Large city	0.442	0.439	0.445
	(0.497)	(0.496)	(0.497)
Small/Medium city	0.303	0.309	0.299
	(0.460)	(0.462)	(0.458)
Rural area	0.254	0.252	0.256
	(0.435)	(0.434)	(0.436)

Notes. M(SD): Mean (standard deviation). †Annual Household income in 10,000 won. Three life scores represent life satisfaction scores for overhall life, health status, and economic status, respectively.

Table 2: Marginal Effects of Selected Risk Factors on the Log Odds Ratio of CVD

VARIABLES	(1) All	(2) No Chronic Disease	(3) Age≤75	(4) Age≤65
Age	0.0357***	0.0330***	0.0390***	0.0386***
	(0.00579)	(0.00628)	(0.00587)	(0.00934)
BMI	0.141***	0.161***	0.142***	0.142***
	(0.0168)	(0.0195)	(0.0174)	(0.0195)
Elementary school	0.207**		0.197**	0.247**
	(0.0942)		(0.0954)	(0.113)
College or higher	-0.295**	-0.422***	-0.301**	-0.352**
	(0.143)	(0.145)	(0.132)	(0.153)
Household size	-0.0806**	-0.0623	-0.0670**	
	(0.0328)	(0.0420)	(0.0326)	
Household income		-4.43e-05**		
		(2.21e-05)		
Life scores-overall	-0.00524*		-0.00540*	-0.00360
	(0.00278)		(0.00287)	(0.00324)
Life scores-health	-0.00172	-0.00309	-0.00189	-0.00299
	(0.00246)	(0.00228)	(0.00238)	(0.00275)
Meet friend 3/week	-0.320***	-0.454***	-0.341***	
	(0.120)	(0.141)	(0.124)	
Mother alive	-0.155*	-0.182*	-0.154*	-0.176*
	(0.0916)	(0.105)	(0.0929)	(0.102)
No Alcohol		-0.300***		
		(0.0937)		
Self-rated health-Fair		0.195*		0.167
		(0.108)		(0.107)
Seoul	0.359***	0.577***	0.297***	0.398***
	(0.116)	(0.129)	(0.113)	(0.131)
Gyunggi	0.315***	0.487***		
	(0.117)	(0.131)		

Table 2: Marginal Effects of Selected Risk Factors on the Log Odds Ratio of CVD

VARIABLES	(1) All	(2) No Chronic Disease	(3) Age≤75	(4) Age≤65
Chungbuk	-0.805***	-0.693***	-0.886***	
	(0.253)	(0.260)	(0.245)	
Chungnam	-0.468***		-0.540***	-0.509**
	(0.170)		(0.171)	(0.209)
Observations	3,104	2,447	3,009	2,378

Notes. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Marginal effects of selected risk predictors on the log odds ratio of CVD from logit regression are presented. Columns (2) to (4) are for subsamples each with the restrictions of No Chronic Disease, Age≤75, and Age≤65 in the baseline year 2008.

Table 3: Marginal Effects of Selected Risk Factors on the Log Odds Ratio of CVD across Subsamples

VARIABLES	(1) Men	(2) Women	(3) Employed	(4) Non-employed
Age	0.0294***	0.0334***	0.0368***	0.0413***
	(0.00825)	(0.00755)	(0.00828)	(0.00707)
BMI	0.180***	0.125***	0.171***	0.114***
	(0.0281)	(0.0199)	(0.0226)	(0.0237)
Elementary school		0.147	0.276**	
		(0.143)	(0.131)	
High school		-0.151		
		(0.141)		
College or higher	-0.368**		-0.350**	
	(0.167)		(0.172)	
Household size		-0.105**		-0.126***
		(0.0413)		(0.0469)
Household income	-6.40e-05**		-5.24e-05**	
	(2.50e-05)		(2.24e-05)	
Handgrip strength	-0.0225**			
	(0.0114)			
Life scores-overall		-0.00512		-0.00797*
		(0.00373)		(0.00412)
Life scores-econ		-0.00316		-0.00327
		(0.00319)		(0.00347)
Married		-0.0662		
		(0.148)		
Meet friend 3/week		-0.348**		
		(0.160)		
Meet friend 4/year				1.252***
				(0.355)

Table 3: Marginal Effects of Selected Risk Factors on the Log Odds Ratio of CVD across Subsamples

VARIABLES	(1) Men	(2) Women	(3) Employed	(4) Non-employed
Mother alive	-0.249*		-0.219*	
	(0.137)		(0.118)	
No Smoke		-0.607*		
		(0.318)		
Regular exercise		-0.269**		
		(0.117)		
Self-rated health-Bad				0.190
				(0.171)
Seoul	0.426**		0.514***	
	(0.172)		(0.154)	
Gyunggi				0.408*
				(0.210)
Gangwon			0.574**	
			(0.223)	
Chungbuk		-0.980***		-1.023***
		(0.361)		(0.363)
Chungnam				-1.022***
				(0.283)
Observations	1,336	1,768	1,785	1,319

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Marginal effects of selected risk predictors on the log odds ratio of CVD from logit regression are presented. Columns (2) to (4) are for subsamples of Men, Women, Employed, and Non-employed in the baseline year 2008.

Table 4: Marginal Effects of BMI on the Log Odds Ratio of CVD from double-selection Lasso

VARIABLES	(1) All	(2) No Chronic Disease	(3) Age≤75	(4) Age≤65
BMI	0.151*** (0.0173)	0.163*** (0.0204)	0.152*** (0.0176)	0.158*** (0.0204)
Observations	2,788	2,223	2,726	2,201
VARIABLES	(5) Men	(6) Women	(7) Employed	(8) Non-employed
BMI	0.194*** (0.0301)	0.131*** (0.0212)	0.180*** (0.0247)	0.119*** (0.0243)
Observations	1,191	1,597	1,645	1,143

Notes. Marginal effects of BMI on the log odds ratio from logit regression are presented. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5: Marginal Effects of baseline BMI and interim BMI on the Log Odds Ratio of CVD from double-selection Lasso

VARIABLES	(1) All	(2) No Chronic Disease	(3) Age≤75	(4) Age≤65
BMI in the baseline	0.0629**	0.0808**	0.0589*	0.0593*
	(0.0298)	(0.0342)	(0.0302)	(0.0346)
BMI in the interim year	0.119***	0.116***	0.127***	0.136***
	(0.0282)	(0.0323)	(0.0284)	(0.0328)
Observations	2,469	1,995	2,419	1,979
VARIABLES	(5) Men	(6) Women	(7) Employed	(8) Non-employed
BMI in the baseline	0.0892*	0.0485	0.0462	0.0877*
	(0.0494)	(0.0374)	(0.0397)	(0.0452)
BMI in the interim year	0.125***	0.121***	0.178***	0.0577
	(0.0461)	(0.0358)	(0.0381)	(0.0420)
Observations	1,068	1,401	1,478	991

Notes. Marginal effects of BMI on the log odds ratio from logit regression are presented. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

APPENDIX

Table A1: Summary of All Variables

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
Primary outcome					
CVD at the 8th survey	3,104	0.366	0.482	0	1
CVD includes:					
• Hypertension	3,104	0.284	0.451	0	1
• Diabetes	3,104	0.111	0.314	0	1
• Heart diseases	3,104	0.0441	0.205	0	1
• Cerebrovascular diseases	3,094	0.0362	0.187	0	1
Predictors of CVD					
ADL(activities of daily living) §1	3,104	0.00870	0.153	0	6
Age	3,104	58.51	8.431	47	86
3 indicators for alcohol drink					
• Present alcohol drinker	3,104	0.423	0.494	0	1
• Past alcohol drinker	3,104	0.0573	0.233	0	1
• No alcohol drink	3,104	0.519	0.500	0	1
BMI	3,104	23.03	2.437	13.15	37.78
3 indicators for regional size					
• Living in a large city	3,104	0.442	0.497	0	1
• Living in a middle/small city	3,104	0.303	0.460	0	1
• Living in a rural area	3,104	0.254	0.435	0	1
3 indicators for cognition					
• Dementia risk	3,104	0.0219	0.146	0	1

Table A1: Summary of All Variables

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
• Cognitive decline risk	3,104	0.111	0.314	0	1
• Normal cognitive function	3,104	0.868	0.339	0	1
Education level					
• elementary school	3,104	0.351	0.477	0	1
• middle school	3,104	0.183	0.387	0	1
• high school	3,104	0.346	0.476	0	1
• college or higher	3,104	0.119	0.324	0	1
• no education reported	3,104	0.000322	0.0179	0	1
5 indicators for employment					
• Employed	3,104	0.253	0.435	0	1
• Self-employed	3,104	0.250	0.433	0	1
• Unpaid work at family business (18+ hours)	3,104	0.0638	0.244	0	1
• Unpaid work at family business (<18 hours)	3,104	0.00838	0.0912	0	1
• Non-employed	3,104	0.425	0.494	0	1
Regular exercise	3,104	0.383	0.486	0	1
10 indicators for friends meeting					
• Meet friends≥4 a week	3,104	0.330	0.470	0	1
• Meet friends once a week	3,104	0.219	0.414	0	1
• Meet friends 2-3 a week	3,104	0.137	0.344	0	1
• Meet friends once a month	3,104	0.148	0.355	0	1
• Meet friends twice a month	3,104	0.0644	0.246	0	1

Table A1: Summary of All Variables

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
• Meet friends 1-2 a year	3,104	0.0261	0.159	0	1
• Meet friends 3-4 a year	3,104	0.0325	0.177	0	1
• Meet friends 5-6 a year	3,104	0.00838	0.0912	0	1
• Meet friends rarely	3,104	0.00322	0.0567	0	1
• No friends to meet	3,104	0.0319	0.176	0	1
5 indicators for health status					
• Health-best	3,104	0.0197	0.139	0	1
• Health-very good	3,104	0.116	0.320	0	1
• Health-good	3,104	0.505	0.500	0	1
• Health-fair	3,104	0.268	0.443	0	1
• Health-bad	3,104	0.0912	0.288	0	1
Household asset*	3,048	24,639	36,470	0	645,800
Household income*	3,104	2,952	2,913	1	60,000
Family size	3,104	2.984	1.259	1	10
Housing type: non-apartment	3,104	0.625	0.484	0	1
Health screening completed	3,104	0.665	0.472	0	1
iADL(instrumental ADL)§2	3,104	0.123	0.674	0	10
Household income squared	3,104	1.720e+07	1.047e+08	1	3.600e+09
Having private health insurance	3,104	0.463	0.499	0	1
12 indicators for job categories					
• do not know	3,104	0.0177	0.132	0	1

Table A1: Summary of All Variables

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
• not reported	3,104	0.00515	0.0716	0	1
• manager	3,104	0.0277	0.164	0	1
• professional	3,104	0.0457	0.209	0	1
• office	3,104	0.0364	0.187	0	1
• service	3,104	0.0677	0.251	0	1
• sales	3,104	0.0667	0.250	0	1
• agriculture	3,104	0.0867	0.281	0	1
• functionary	3,104	0.0480	0.214	0	1
• machine-related	3,104	0.0480	0.214	0	1
• manual	3,104	0.117	0.321	0	1
• Missing/ no job	3,104	0.433	0.496	0	1
Currently work	3,104	0.575	0.494	0	1
Male	3,104	0.430	0.495	0	1
5 indicators for marital status					
• Married	3,104	0.874	0.332	0	1
• Separated	3,104	0.00644	0.0800	0	1
• Divorced	3,104	0.0177	0.132	0	1
• Widowed	3,104	0.0947	0.293	0	1
• Single, never married	3,104	0.00677	0.0820	0	1
Handgrip strength	3,104	26.00	7.857	5	55.75
Cognitive function scores (MMSE) §3	3,104	26.98	3.413	2	30
3 types of national health insurance					
• National health insurance-job	3,104	0.605	0.489	0	1
• National health insurance-region	3,104	0.364	0.481	0	1
• National health insurance-missing	3,104	0.0309	0.173	0	1

Table A1: Summary of All Variables

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
4 indicators for parental living					
• Both parents alive	3,104	0.0944	0.292	0	1
• Father alive	3,104	0.0274	0.163	0	1
• Mother alive	3,104	0.281	0.450	0	1
• Both parents passed away	3,104	0.597	0.491	0	1
6 indicators for religion					
• No religion	3,104	0.477	0.500	0	1
• Christian	3,104	0.195	0.396	0	1
• Catholic	3,104	0.0722	0.259	0	1
• Buddhist	3,104	0.247	0.432	0	1
• Won-Buddhist	3,104	0.000966	0.0311	0	1
• Other religion	3,104	0.00805	0.0894	0	1
Number of siblings alive	2,853	3.746	1.709	1	11
3 indicators for smoking					
• No smoke	3,104	0.709	0.454	0	1
• Past smoker	3,104	0.102	0.303	0	1
• Current smoker	3,104	0.189	0.391	0	1
15 Regions of residence					
• Seoul	3,104	0.138	0.345	0	1
• Pusan	3,104	0.0876	0.283	0	1
• Daegu	3,104	0.0657	0.248	0	1
• Incheon	3,104	0.0316	0.175	0	1
• Gwangju	3,104	0.0474	0.212	0	1
• Daejeon	3,104	0.0506	0.219	0	1
• Ulsan	3,104	0.0370	0.189	0	1
• Gyunggi	3,104	0.135	0.341	0	1
• Gangwon	3,104	0.0419	0.200	0	1
• Chungbuk	3,104	0.0425	0.202	0	1
• Chungnam	3,104	0.0735	0.261	0	1

Table A1: Summary of All Variables

VARIABLES	(1) N	(2) mean	(3) S.D.	(4) min	(5) max
• Jeonbuk	3,104	0.0506	0.219	0	1
• Jeonnam	3,104	0.0499	0.218	0	1
• Gyungbuk	3,104	0.0677	0.251	0	1
• Gyungnam	3,104	0.0815	0.274	0	1
Monthly allowance	3,101	17.48	16.29	0	200
Depression scores[§4]	3,089	3.016	2.686	0	10
Life scores-econ	3,104	54.44	20.80	0	100
Life scores-health	3,104	62.36	19.03	0	100
Life scores-overall	3,104	64.28	17.11	0	100
Weekly work hours (main job)	3,104	27.62	27.04	0	112
Weekly work hours (supplementary jobs)	3,104	0.257	2.987	0	91

Notes: N. number of observations, S.D. standard deviations. ADL[§1] and iADL[§2] are measures of cognitive and physical functioning based on daily activities such as washing, dressing, feeding, and mobility. MMSE[§3] is a mini-mental state examination that evaluates cognitive function. Depression scores[§4] is a summarized measure of CES-D (Center for Epidemiologic Studies Depression) scores that assesses depression status. [*]Household assets and household income are in 10,000 won.