Identifying the best set of cardiovascular disease risk predictors with automated variable selection methods in the Northern Swedish population.

Ka Chun Tsang

3 Jul, 2017

# Abstract

### Background

Predictors used in cardiovascular disease (CVD) risk prediction algorithms are commonly chosen based on expert knowledge. Such “hand-picked” method however is difficult to replicate across various settings, in which data available to the experts can be different.

### Methods

In this study, easily implemented automated variable selection methods are used to select potent CVD risk predictors from a health questionnaire deployed in Västerbotten Intervention Programme. The questionnaire contains more than 100 questions on psychosocial health, anthropometric measurements, and daily habits. The methods are backward selection in Cox model, LASSO, and survival boosted regression tree. Predictor sets found are compared to the predictor set used in non-laboratory Framingham risk algorithm. Temporal validation is applied to examine the prediction performance.

### Results

Age and gender are the only predictors that are chosen by all methods. Predictors representing socioeconomic status, predisposing risk, physical activity, job satisfaction, alchohol consumption are chosen by the automated methods, leading to different combinations of predictors. Although the sets of predictors are different, all predictor sets show similar prediction performance (x test:mean,CI,p-value). Yet, predictor sets from FHS shows the best performance (AUC:).

### Conclusion

The results suggest that automated variable selection methods can be adopted to select predictors for accurate prediction and expert knowledge may further enhance the prediction performance. The authors therefore suggest that automated variable selection methods and expert knowledge can collaborate to identify the best set of predictors for improved prediction of CVD.

### key words

**variable selection**, **risk prediction**, **cardiovascular disease**

**Key message**  
*Automated variable selection methods can be implemented easily on large amount of available predictors.*  
Predictor set chosen by expert knowledge performs better than sets selected by automated methods, yet the difference is not significant.  
\*Expert knowledge and automated variable selection methods can work together for better prediction.

# Introduction

Predictors used in cardiovascular disease (CVD) risk prediction algorithms are commonly chosen based on expert knowledge, for example, in the Framingham Heart Study (FHS) the set of predictors are then first recruited by expert knowledge, then passed to statistical models. After that predictors which are statisticall significant (p-value <0.05) are selected to the multivariate risk prediction model1.

This “hand-picked” method by the expert has the advantage of identifying predictors amid the available variables in various contexts. However it is difficult to replicate and compare across various settings, in which data available to the experts can be different. Therefore few easily implemented automated variable selection methods are proposed to assist the predictor recruitment process across different settings.

## CVD in Northern Sweden

CVD has been the most burdensome disease in Northern Sweden. A community-based intervention has been initiated to tackle the situation - the Västerbotten Intervention Programme (VIP), in which all citizens in the Västerbotten county were invited to health examination and health counselling at the age of 40, 50 and 602.

A comprehensive health questionnaire designed by expert committee is given to participants to capture several dimensions of health related CVD, such as, anthropometric measurements3, job satisfation4, and physical activity5.

However, there are more than 100 individual questions in the health questionnaire. There is a need of locatoing the best predictors for quick CVD risk assessment in clinical setting. In this study, automated variable selection methods are used to look for the best set of predictors in the context of Northern Sweden.

## Automated variable selection methods in Cox model

Cox model has been the most popular statistical model in CVD risk prediction6. The automated variables selection methods should be thus able to be implemented in the Cox model and easily implemented. The R code is attached in the appendix for reference.

These methods are Cox backward selection, LASSO, and survival boosted regression tree (sBRT).

### Cox backward selection

The Cox backward selection method is based on Lawless and Signal7 and implemented in R through the work of Harrell Jr8. This method removes variables that demands for more degree of freedom. The process is stopped by Akaike’s information criteria (AIC)9.

### LASSO

LASSO stands for least absolute shrinkage and selection operator. It is a shrinkage method which can shrink coefficients of variables to zero and therefore perform variable selection10. LASSO has been extended to the Cox proportional hazards model to shrink the log partial likelihood11.

### Survival boosted regression tree

Survival boosted regression tree (sBRT) is an application of boosted regression tree under the Cox model. Boosted regression tree is a product of boosting and classification tree, which forms a combination of many classification trees10. The variable selection process in sBRT is based on relative importance, which is computed based on the frequency of using a particular variable for creating split and the improvement to the prediction model brought by such split12.

# Objective

The objective of this study is to examine if the set of predictors identified by automated variable selection methods can produce comparable predictive performance to the set of predictors identified by expert knowledge.

# Method

## Event free samples

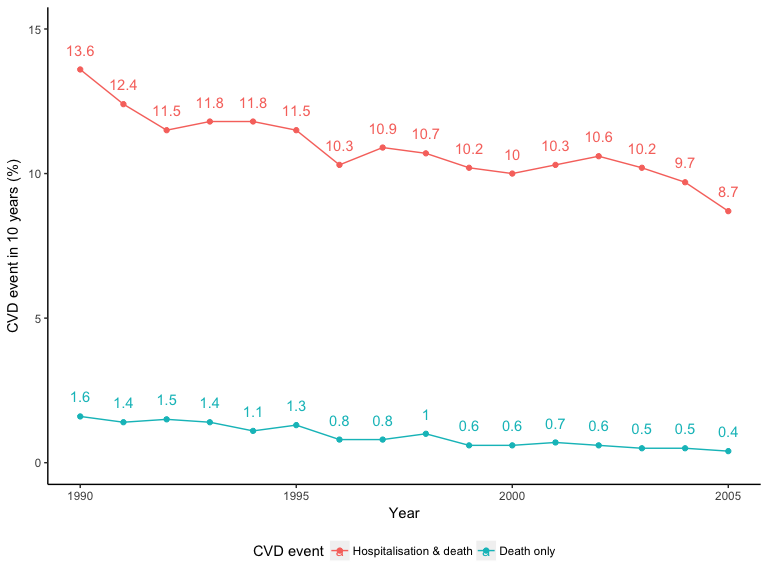
First, observations participated health examination in the VIP between 1990 and 2005 are extracted. Then, only observations without CVD history are selected. Therefore, observations are excluded if they had reported of having CVD event or angina treatment in the health questionnaire, or, had been admitted to the hospital due to CVD before the examination date based on the hospitalisation record.

## Outcome

The outcome of interest is gathered from the hospitalisation record and death registry from 1990 to 2015. The outcome is selected to be the first event of either hospitalisation due to CVD or CVD death. Definition of CVD event is based on ICD-9 and ICD-10. The ICD codes can be found in the appendix.

The primary reason for choosing both CVD morbidity and mortality is due to the diminishing mortality rate coupled with a constant morbidity rate as shown in figure 1. This suggests that advancement in medical technology and service improves the survival chance when one experiences a CVD event. The use of both CVD hospitalisation and death also put emphasis on primary CVD prevention.

Survival time is obtained by calculating the difference between the first event as recorded in the hospitalisation record and death registry.

 Figure 1. Hospitalisation

## Predictors

Predictors are extracted from the health questionnaire answered during health examination. Before passing all predictors to the automated variables selection methods, management on missing data, multicollinearity, and outliers is performed.

### Missing data

Since the questionnaire has been being reviewed and updated, some predictors appeared and disappeared along time. Predictors are excluded when the overall missing rate between 1990 and 2005 is larger than 30%.

Missing data among categorical variables are coded into category “Missing” to seek for potential missing-not-at-random. Missing data in numeric variables are imputed by multivariate imputation by chained equations13.

### Multicollinearity

Correlation among all predictors are examined to unveil multicollinearity. Correlation among categorical variables is examined with Goodman and Kruskal’s tau (36), in which numeric variables are divided into 5 quantiles to become categorical variables. No correlation is greater than 0.5 or less than -0.5.

Correlation between numeric predictors is examined by Pearson’s correlation coefficient. The year of birth is strongly correlated to age, while systolic blood pressure is highly correlated with diastolic blood pressure. Age and systolic blood pressure are kept for analysis.

Outliner in numeric variables are removed with the Tukey’s method, in which data points are classified as outliners when they are either above or below 1.5 times of the inter-quantile range. Box plots and histograms before and after the outlier removal can be found in the appendix.

## Statistical analysis

### Data partition

The whole dataset is divided into two sets based on the year of health examination. Observations from 1990 to 1999 are placed into *training set*, which is used to construct statistical models. Observations from 2000 to 2005 are placed into *test set*, which is not involved in modelling and is used only for examining the prediction performance. This practice is regarded as temporal validation. Temporal validation simulates the real-life scenerio that risk prediction models are built from historical data and applied to patients in future. The use of temporal validation also helps reduce overfitting.

### Procedure

All predictors were passed into the automated variable selection methods. Each set of predictors are then fitted into Cox proportional-hazard regression respectively. A predictor set containing all selected predictors by any method is also fitted. All models were compared based on the predictive performance - area under the receiver operative curve (AUC) [pencina2004overall].

## Ethical approval

Consent of the participants was obtained during the health examination by trained healthcare professional. Data extraction of this study has been approved by the ethical committee in Umeå University and Västerbotten county council.

# Results

|  |  |  |
| --- | --- | --- |
|  | Methods |  |
| Predictors | Expert | Backward Cox | LASSO | sBRT | Full |
| :——– | :——-: | :——–: | :——–: | :——–: | :——–: |
| Reported diabetes status | X |  |  |  |  |
| Taking anti-hypertensive drugs |  |
| Education level |  |
| Ever have sick leave longer than 6 months |  |
| Talk to colleague during break |  |
| Berry picking at free time |  |
| Last visit from colleagues |  |
| Consume alcohol |  |
| Smoking habit |  |
| Sex | XXX |  |
| Age | XXX | 1 |
| Body mass index | X |  |
| Total cholesterol | X10 |
| Systolic blood pressure | XXX4 |

# Appendix

* R code
* ICD code of CVD events
* Box plots and histograms before and after the outlier removal
* Full list of predictors (compared between 4 groups: train+CVD, test+CVD, train+noCVD, test+noCVD)

# Require effect estimates with confidence intervals and exact P values)

pandoc -S -o try.docx –filter pandoc-citeproc try.mdown This is then the next paragraph. [^1] [^1]: just a footnote… This is the second sub section. Please refer to [here](www.facebook.com).

1. D’Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care. *Circulation*. Am Heart Assoc; 2008;**117**(6):743–753.

2. Norberg M, Wall S, Boman K, Weinehall L. The västerbotten intervention programme: Background, design and implications. *Global health action*. Taylor & Francis; 2010;**3**(1):4643.

3. Dawber TR, Meadors GF, Moore Jr FE. Epidemiological approaches to heart disease: The framingham study. *American Journal of Public Health and the Nations Health*. American Public Health Association; 1951;**41**(3):279–286.

4. Li J, Loerbroks A, Bosma H, Angerer P. Work stress and cardiovascular disease: A life course perspective. *Journal of occupational health*. Japan Society for Occupational Health; 2016;**58**(2):216–219.

5. Lapidus L, Bengtsson C. Socioeconomic factors and physical activity in relation to cardiovascular disease and death. a 12 year follow up of participants in a population study of women in gothenburg, sweden. *Heart*. BMJ Publishing Group Ltd; 1986;**55**(3):295–301.

6. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* [Internet]. 2016 May [cited 2017 Apr 20];**353**:i2416. Available from: <http://www.bmj.com/content/353/bmj.i2416>

7. Lawless J, Singhal K. Efficient screening of nonnormal regression models. *Biometrics*. JSTOR; 1978;318–327.

8. Harrell F. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015.

9. Akaike H. Likelihood of a model and information criteria. *Journal of econometrics*. Elsevier; 1981;**16**(1):3–14.

10. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics New York; 2001.

11. Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in medicine*. Wiley Online Library; 1997;**16**(4):385–395.

12. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

13. Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*. American Statistical Association; 2011;**45**(3).