Computer programs to estimate overoptimism in measures of discrimination for predicting the risk of cardiovascular diseases

Haider R. Mannan PhD MSc MA1 and John J. McNeil MBBS MSc PhD FRACP FAFPHM2

1Research Fellow, 2Professor, Department of Epidemiology & Preventive Medicine, Monash University, Melbourne, Victoria, Australia

Abstract

Background Development of chronic disease risk prediction models has become a growing area of research in recent years. The internal validity of such models is sometimes lower than estimated from the development sample. Overfitting or overoptimism of the developed model and/or differences between the samples are likely causes for this. For modelling of an uncommon outcome, bootstrapping for overoptimism is the preferred method for afterwards shrinking of regression coefficients and the model’s discrimination and calibration for overoptimism. However, computer programs for different types of bootstrap validation are not readily available. We developed two SAS macro programs – one for the simple bootstrap that compares the discriminatory performance of the Cox proportional hazards model from the original sample in bootstrap samples; and another (which is more efficient), known as stepwise bootstrap validation, that makes the same comparison but from models developed by variable selection from bootstrap samples in the original sample. These are illustrated through an example from cardiovascular disease (CVD) risk prediction.

Methods Two SAS macro programs for Cox proportional hazards model using Proc PHREG were developed for estimating overoptimism in Harrell’s C and Somers’ D statistics. The computer programs were applied to data on CVD incidence for a Framingham cohort that combined both the original and offspring exams. The risk factors considered were current smoking, diabetes, age, sex, systolic blood pressure, diastolic blood pressure, total cholesterol, high-density lipoprotein cholesterol, triglycerides and body mass index.

Results The degree of overoptimism in both Harrell’s C and Somers’ D statistics were low. Both these statistics were corrected for overoptimism by subtracting overoptimism from their observed values. Between the two bootstrap validation algorithms, the degree of overoptimism was estimated to be higher for stepwise bootstrap validation.

Conclusion The programs are very useful for evaluating the ‘overoptimism corrected’ predictive performance of Cox proportional hazards model.

Introduction

In risk prediction modelling, the reported results on the performance of new algorithms are often known to be overoptimistic. The first and perhaps most obvious reason for this is that researchers sometimes randomly search for a specific dataset such that their new method works better than existing approaches, yielding a so-called ‘dataset bias’. While a method cannot reasonably be expected to yield ‘universally better’ results in all datasets, it would be wrong to report only favourable datasets without mentioning and/or discussing the other results. This strategy induces an optimistic bias. The other source of overoptimism is model uncertainty caused by pre-specification of the structure of the model which leads to selection bias caused by selection of predictors from a larger set of potential predictors.

In chronic disease risk prediction modelling, researchers often use the development sample to optimize their new algorithms. It is a trial-and-error process which constitutes an important component of data analysis research. However, this is an unpredictable search process. The problem is that this search process leads to an artificial optimization of the method’s characteristics to the considered datasets. Hence, the superiority of the new method over an
existing method is sometimes considerably overestimated. In a concrete medical prediction study, fitting a prediction model and estimating its error rate using the same development sample yields a downwardly biased error estimate commonly termed as apparent error. Developing a new algorithm (i.e. selecting one of many variants) and evaluating it by comparison with existing methods using the same dataset may lead to optimistically biased results in the sense that the new algorithm’s characteristics overfit the used dataset. To overcome these methodological limitations of developing a new algorithm based on the same dataset, bootstrapping validation is commonly used to quantify the degree of optimism in the new algorithm. The other advantage of bootstrapping validation is that it can appropriately reflect most sources of model uncertainty, especially, variable selection methods. This feature of bootstrapping validation is not shared by any other model validation method such as split-sample validation or cross-validation which use different datasets for developing and validating a model. These methods are generally less stable [1], and particularly for a small sample or for an uncommon outcome of interest, will be less efficient than bootstrapping validation. Also, these methods do not provide external validation.

The current article aims at illustrating SAS computer programs for estimating optimism in measures of discrimination using two bootstrap validation algorithms through a concrete example from an active research field of chronic disease risk prediction-validation of a cardiovascular disease (CVD) risk prediction model. We used the Cox proportional hazards model as the platform for such development as it is the most commonly used survival method for risk modelling of a chronic disease. The most commonly used measures of model discrimination are used for assessing optimism in predictive performance of this model. These include Harrell’s C and Somers’ D statistics. We can apply the bootstrap method to any performance measure including the $R^2$ and calibration measures. This will be the subject of another paper. Furthermore, methodological comparison between the two bootstrap algorithms that we have employed is beyond the scope of this paper.

**Methods**

Two bootstrap validation methods are used for estimating overoptimism in Harrell’s C and Somers’ D statistics for Cox proportional hazards model. The bootstrap methods and the two measures of discrimination are discussed next.

**Bootstrap validation**

In bootstrapping method, bootstrap samples of the same size as the original sample are drawn with replacement from the original sample, reflecting the drawing of samples from an underlying population. With a simple bootstrap validation [2,3], first a model is repeatedly fitted in bootstrap samples and its performance is evaluated in the original sample. The decrease between model performance in the bootstrap sample and performance in the original sample is used to obtain the optimism in model performance in the original estimate. This optimism is subsequently subtracted from the original estimate to obtain optimism-corrected performance estimate. A more efficient approach is to test models from the bootstrap samples in the original sample to obtain an estimate of the model’s overoptimism [1,4]. In this approach, first a model is constructed in the original sample by selecting predictors from a larger set of candidate predictors using a variable selection method. The apparent performance of the model is determined using a model performance indicator. Then, the model is constructed using each bootstrap sample replaying every step that was done in the original sample and the bootstrap performance as the apparent performance of the model on bootstrap samples is determined. Then the bootstrap model is applied to the original sample to determine its predictive performance. The decrease between model performance in each bootstrap sample and from the bootstrap samples in the original sample is used to obtain the optimism in model performance in the original estimate. This optimism estimate is subtracted from the apparent performance to obtain the optimism-corrected performance estimate. To distinguish this method from simple bootstrapping, we call this method stepwise bootstrapping throughout this paper.

**Measures of discrimination**

Harrell’s C statistic for survival data is defined as probability of concordance given comparability [5]. Any two subjects are comparable if for any two subjects i and j, $T_i > T_j$ or $T_i < T_j$ where $T$ denotes survival time. Any two subjects are concordant if for $T_i > T_j$, $Z_i < Z_j$ or for $T_i < T_j$, $Z_i > Z_j$. Thus, mathematically, Harrell’s C statistic can be defined as:

$$C = \Pr(Z_i > Z_j | T_i < T_j) = \Pr(Z_i < Z_j | T_i > T_j)$$

The Harrell’s C is the probability that the survivor has the lower hazard ratio plus half the (possibly negligible) probability that the two subjects have equal hazard ratios.

Somers’ D is a rank statistic that measures how concordant predicted hazards are with observed failure times [2]. For survival data, D ranges between 0 and 1, although the lower confidence limit can be negative. For example, a Somers’ D value of 0.8 implies that when one of two subjects is observed to survive another, it is 80% more likely that the survivor has the lower of the two hazard ratios, predicted by the model, than that the survivor has the higher of the two predicted hazard ratios. It has been proposed that concordance measures of predictive accuracy in survival models are natural extensions to the proportion of variation measures [6]. Thus, the square of D, like R-squared, ranges between 0 and 1. Harrell’s C statistic can be transformed to estimate Somers’ D statistic for survival data using the relationship, $D = 2C - 1$.
The design, selection criteria and examination procedures of FHS have previously been elaborated in detail [7–11]. The outcome of interest is time (in years) until the occurrence of first CVD which includes stroke, myocardial infarction, angina pectoris, coronary insufficiency and sudden death. The study cohort consisted of people from examination 1 (1971–1975) of the offspring cohort and from examinations 10 and 11 (1968–1971) of the original cohort for whom high-density lipoprotein (HDL) cholesterol levels were measured for the first time. For the original cohort, in most cases (81.3%), HDL was measured for the first time at examination 11, while for some cohort members, it was examination 10. Follow-up was performed through the 22nd examination cycle, a span of approximately 24 years. For the offspring cohort, risk factor measurements were from the first examination cycle (1971–1975), whereas follow-up was performed through the sixth examination cycle, approximately 24 years later. Participants were considered eligible if at the baseline they were aged 30–49 years, were free of CVD (CHD and stroke), cancer and chronic kidney disease (888 cases were excluded), and had complete information on covariates. After exclusions, the study included 1303 persons (1062 members, it was examination 10. Follow-up was performed at examination 11, while for some cohort members, it was examination 10. Follow-up was performed through the 22nd examination cycle, a span of approximately 24 years. For the offspring cohort, risk factor measurements were from the first examination cycle (1971–1975), whereas follow-up was performed through the sixth examination cycle, approximately 24 years later. Participants were considered eligible if at the baseline they were aged 30–49 years, were free of CVD (CHD and stroke), cancer and chronic kidney disease (888 cases were excluded), and had complete information on covariates. After exclusions, the study included 1303 persons (1062 events, 241 censored). The exclusions for stroke, cancer and chronic kidney disease were done to reduce the effects of reverse causation.

### Selection of risk factors

Both systolic and diastolic blood pressures (SBP and DBP) were considered for assessing the overall risk associated with hypertension as both the Framingham and other investigators have observed that both these variables correlate with major chronic diseases [8,12]. In case of cholesterol, we considered both total and HDL cholesterol levels because the latter is also considered to be important [8]. In case of cigarette smoking, we considered a simple binary current smoker/non-smoker measure. The other risk factors examined were age, sex, diabetes status, triglycerides and body mass index (BMI).

### Measurement of risk factors

We took the average of two physician-obtained measures to compute the examination blood pressure. To determine serum total and HDL cholesterol levels, we used standardized enzymatic methods. In the original Framingham cohort, diabetes was diagnosed if a casual whole blood glucose measurement was 150 mg dL$^{-1}$ or above, or the individual was being treated with insulin or oral hypoglycaemics. In the offspring Framingham cohort, we used a more recent definition of diabetes as of requiring treatment or a fasting glucose level of 140 mg dL$^{-1}$ or above from plasma measurement. Cigarette smoking status was obtained by self-report.

### Development and assessment of predictive models

We used Cox’s proportional-hazards model [13] to relate risk factors to the risk of an incident CVD event during a median follow-up of 21 years. All continuous covariates were logarithmically transformed to improve discrimination and calibration of the models and to minimize the influence of extreme observations. We examined mortality increase for each unit increase on the logarithmic scale in each continuous variable and compared categories for other variables – for example, current smoking or past smoking compared with never smoking. All covariates were tested for the assumption of proportional hazards. For a nominal covariate, the proportional hazards were tested by examining whether there is a non-significant interaction between that covariate and log of survival time [14] and for a continuous covariate by plotting Schoenfeld residuals against log(survival time). The discriminative ability of a model was assessed by Harrell’s C and Somers’ D statistics while internal validation was assessed by bootstrapping method, and for comparison purposes, we used the same model structure for estimating overoptimism by the two bootstrapping methods.

### Results

Based on Table 1, Harrell’s C and Somers’ D statistics for discriminating CVD risk based on the observed sample were estimated to be 0.60109 and 0.20217, respectively. We found that the degrees of overoptimism in these statistics stabilized after 500 bootstrap replications. The degree of overoptimism in these statistics based on simple bootstrapping method with 500 replications is estimated to be 0.00631 and 0.012623, respectively (see Table 1). Thus, the optimism-corrected values for Harrell’s C and Somers’ D statistics are 0.59478 and 0.189547, respectively. For a nominal covariate, the optimistic-corrected values for Harrell’s C and Somers’ D statistics are 0.587537 and 0.175064, respectively (see Table 1). Thus, the optimism-corrected values for Harrell’s C and Somers’ D statistics are 0.59478 and 0.189547, respectively. The discriminative ability of a model was assessed by Harrell’s C and Somers’ D statistics while internal validation was assessed by bootstrapping method, and for comparison purposes, we used the same model structure for estimating overoptimism by the two bootstrapping methods.

### Table 1 Observed values of Harrell’s C and Somers’ D statistics, degrees of overoptimism obtained by simple and stepwise bootstrap methods, and optimism-corrected values of the two statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Simple bootstrap</th>
<th>Stepwise bootstrap</th>
<th>Simple bootstrap</th>
<th>Stepwise bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harrell’s C</td>
<td>0.60109</td>
<td>0.00631</td>
<td>0.013553</td>
<td>0.59478</td>
</tr>
<tr>
<td>Somers’ D</td>
<td>0.20217</td>
<td>0.012623</td>
<td>0.027106</td>
<td>0.189547</td>
</tr>
</tbody>
</table>

© 2012 Blackwell Publishing Ltd
Appendix S1C for estimating the observed and optimism-corrected values of these statistics. Supporting Information Appendix S2A presents the SAS macro for estimating overoptimism in C and D statistics based on stepwise bootstrapping while Supporting Information Appendix S2B shows the SAS codes for selecting the predictors of CVD risk and SAS codes for estimating stepwise bootstrapping-based optimism-corrected values of Harrell’s C and Somers’ D statistics.

Discussion

Overoptimism of performance measures is a common phenomenon in risk prediction modelling. The main causes of overoptimism are data uncertainty and model uncertainty. In this paper we have presented SAS macros for implementing two bootstrap validation approaches for estimating overoptimism in Harrell’s C and Somers’ D statistics in Cox proportional hazards model. The main advantage of using bootstrap validation approaches is that they can appropriately reflect data uncertainty in predictive modelling and most sources of model uncertainty, particularly, variable selection methods. The two model performance measures we have used are the most commonly used measures of discrimination. The illustration we have given here is from CVD risk prediction modelling which is an active area of research in the medical literature. However, the macros can easily be used for any event of interest over any length of follow-up. The SAS macros presented here are easy to follow and modify and can be incorporated quickly by the user for immediate use. Thus, this paper should provide useful analytic tools for researchers interested in developing risk prediction equations for any event of interest over any period of follow-up. These macros would indeed facilitate researchers to quantify overoptimism in the commonly used measures of discrimination while developing risk prediction models and then correctly evaluate their predictive performance through the use of overoptimism-corrected versions of these measures.

Funds

The Framingham Heart Study – Offspring (FHS-O) is conducted and supported by the NHLBI in collaboration with the FHS-O Study Investigators. This manuscript was prepared using a limited access dataset obtained from the NHLBI and does not necessarily reflect the opinions or views of the FHSO or the NHLBI. This research was supported by an NHMRC health services research grant (no. 465130).

Competing interests

The authors declare that they have no competing interests.

References


Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1A Before running the SAS macros, the following SAS Data Step should be run to exclude missing cases in risk factors. This will not be required if there are no missing cases in risk factors.

Supporting information

There are no data to share.
Appendix S1B SAS macros for estimating overoptimism in Harrell’s C and Somers’ D statistics based on simple bootstrapping method.

Appendix S1C SAS macros for estimating the observed and simple bootstrapping-based optimism-corrected values of Harrell’s C and Somers’ D statistics.

Appendix S2A SAS macros for estimating overoptimism in Harrell’s C and Somers’ D statistics based on stepwise bootstrapping method.

Appendix S2B SAS codes for estimating the predictors of CVD risk and for estimating stepwise bootstrapping-based optimism-corrected values of Harrell’s C and Somers’ D statistics.