Journal of Biomedical Informatics 44 (2011) 595–606

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An empirical approach to model selection through validation for censored survival data

Ickwon Choi [a,*], Brian J. Wells [b], Changhong Yu [b], Michael W. Kattan [b]

[a] Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106, USA
[b] Department of Quantitative Health Sciences, Cleveland Clinic Foundation, 9500 Euclid Avenue, JJN3, Cleveland, OH 44195, USA

ARTICLE INFO

Article history:
Received 19 June 2010
Available online 16 February 2011

Keywords:
Predictive model
Survival analysis
Variable selection
Cox proportional hazards model
Estimation bias
Selection bias
Censored data
Concordance index

ABSTRACT

Medical prognostic models can be designed to predict the future course or outcome of disease progression after diagnosis or treatment. The existing variable selection methods may be precluded by full model advocates when we build a prediction model owing to their estimation bias and selection bias in right-censored time-to-event data. If our objective is to optimize predictive performance by some criterion, we can often achieve a reduced model that has a little bias with low variance, but whose overall performance is enhanced. To accomplish this goal, we propose a new variable selection approach that combines Stepwise Tuning in the Maximum Concordance Index (STMC) with Forward Nested Subset Selection (FNSS) in two stages. In the first stage, the proposed variable selection is employed to identify the best subset of risk factors optimized with the concordance index using inner cross-validation for optimism correction in the outer loop of cross-validation, yielding potentially different final models for each of the folds. We then feed the intermediate results of the prior stage into another selection method in the second stage to resolve the overfitting problem and to select a final model from the variation of predictors in the selected models. Two case studies on relatively different sized survival data sets as well as a simulation study demonstrate that the proposed approach is able to select an improved and reduced average model under a sufficient sample and event size compared with other selection methods such as stepwise selection using the likelihood ratio test, Akaike Information Criterion (AIC), and lasso. Finally, we achieve better final models in each dataset than their full models by most measures. These results of the model selection models and the final models are assessed in a systematic scheme through validation for the independent performance.

1. Introduction

Medical prognostic models can be designed to predict the future course or outcome of disease progression after diagnosis or treatment. Such models can provide individualized predictions about the characteristics of one single patient. However, there is considerable uncertainty within the statistical modeling community regarding how best to develop an accurate prediction model for censored survival data. Specifically, when it comes to variable selection, some advocate fitting the full model [1] in which predictors are pre-specified with external information from the literature, while variable selection methods remain popular [2,3]. Nonetheless, a full model may be large and complicated to be used as a statistical tool. There is little literature comparing these primary approaches with respect to the predictive accuracy in censored clinical data. Logistic regression models [3–6] have been studies for clinical models. If the goal is to optimize predictive accuracy for finding a set of reduced prognostic factors, a plausible alternative to the full model would be to fit the most accurate, possibly reduced, model. An argument can easily be made for a parsimonious model that is at least as accurate as the full model.

In general, the complexity of a model obtained by a procedure of variable selection is expected to be less than that of the full model, and the variance of the estimated parameters should be lower. Nevertheless, recent studies emphasize the limitations of variable selection, such as bias in the estimates of parameters (*estimation bias*) and the lack of stability in an iterative scheme of variable selection [4]. In a stable algorithm, the effect of computational error during the iteration is no worse than that of a small amount of input data error from multi-collinearity [7,8]. The unstable variable selection algorithm may enlarge initial perturbations after numerous iterations. Furthermore, in variable selection, multi-collinearity between the omitted variables and the selected variables can cause *selection bias*. Dropping influential variables from the effective model results in underfitting to data with increased residuals and biased parameter estimates for

* Corresponding author. Fax: +1 216 368 2800.
E-mail addresses: ickwon.choi@case.edu (I. Choi), WELLSB@ccf.org (B.J. Wells), YuC@ccf.org (C. Yu), kattanm@ccf.org (M.W. Kattan).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.02.005

selected variables (*omission bias*). Adding unimportant variables to the effective model induces overfitting and increases the variance of parameter estimates for correlated predictors [9]. We attempt to reduce the instability and increase the reliability of the selection algorithm using the resampling method of cross-validation [10].

A large sample size is the need for a problem of fitting the full model with numerous and complicated predictors to obtain unbiased estimation in model fitting, and the possibility of overfitting due to model complexity. The sample size problem due to the model complexity can be accounted for by the *curse of dimensionality* [11]. In fact, regression modeling with time-to-event data is much more sensitive to the events per variable (EPV) [12] than the overall sample size. Some researchers carefully guide the EPV ratio to estimate bias and sometimes suggest using shrinkage of the coefficient estimates [4]. However, highly correlated features in this situation may produce high variance, even if there is no estimation bias according to the EPV. Hence, this guidance is crucial to model building at the developing step.

The last challenging characteristic of clinical survival data, to tackle in variable selection, is right censoring. There are two types of censoring in classical survival models: (i) Type I: survival until the end of study but whose final event time is unknown; (ii) Type II: lost to follow-up after a certain time. Even though data are incomplete, they contain a certain amount of information to increase the sample size and thus improve performance of the model. However, with the presence of censoring, the behavior of the underlying mechanism produces unclear performance measurements of models and may lead to biased results in variable selection. In survival analysis, Cox regression models are commonly used and one of the major advantages is the ability to utilize censored observations. We use the Cox proportional hazards model [13,14] in this article. In order to consider the censoring in model assessment, many performance measures, which summarize a time dependency using integration [15,16] and are robust to censoring [17], are introduced to quantify the prediction accuracy and the amount of prognostic information represented by the model; Some of these appear in Section 3.2. However, among them, maximizing the C-Index has some patterns to enhance other measures along with it and some merits (see Section 4). As a predictive accuracy, the C-Index is a preferred choice in this study.

Fig. 1 illustrates the optimization path with the initial point of a full model in a variable selection procedure of this study. The selection methods start from the full model, which is a type of single final model, and select the best model, optimized in some criterion. The starting full models can be categorized into three groups depending on the above challenges with the data involving the event size, the model complexity, and the degrees of censoring: (a–c) in Fig 1. The objective of model selection is to achieve the final model with optimal model complexity based on the prediction accuracy while tuning the tradeoff between bias and variance. In theory, the type (a) completes the course at (b), and in the types of (b) and (c), the full model is the final model, in which the difference is that in (c) the full model may suffer from a lack of data, adequately significant predictors, or high rate of right censoring at the initial point.

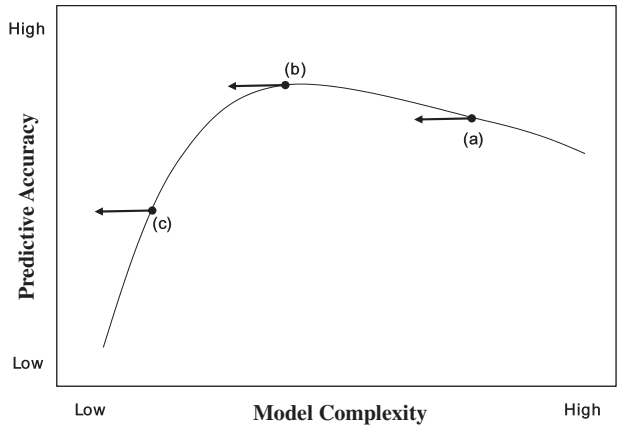

Fig. 1. Types of initial full models in the optimization path to their final models.

The aim of this paper is to propose a novel approach that builds a parsimonious model that is at least as accurate as the full model with respect to the C-Index as an objective criteria. Herein, we propose a new approach to address these problems in two stages: (1) Stepwise Tuning in the Maximum Concordance Index (STMC) as a variable selection process within each set of training folds of outer cross-validation using inner cross-validation for the optimism correction and (2) Forward Nested Subset Selection (FNSS) as overfitting control, which reduces uncertainty and variability in the predictors of chosen models resulting from STMC and builds a single final model. In the new approach, Cox proportional hazards regression models with only main effect terms are used and fitted to two censored clinical data sets in the areas of renal transplantation and prostate cancer. For the comparative study of methods and models, we employ the same scheme as the first stage of our approach to compare our proposed method against the alternatives of the stepwise method that uses the likelihood ratio test and AIC criterion and the lasso using an L_1 absolute value penalty that has two meritorious features of shrinkage and model selection [18,19]. Then, we compare the single final model of a FNSS result with the full model for final model assessment.

Section 2 describes two clinical data sets. In Section 3, we define censored data and performance measures for prediction models and present our new approach for the selection of a reduced and accurate model. In Section 4, our methods are applied to the two data sets and we compare the results. In Section 5, we discuss limitations, further studies, and provide concluding remarks.

2. Data sets

2.1. Prostate cancer data

We procured data from a study that created a post-operative nomogram for predicting the risk of prostate cancer recurrence [20] following institutional Review Board waivers (Cleveland Clinic IRB number: 4270). The cohort consists of a total of 1123 patients (with 167 biochemical recurrences) with clinically localized prostate cancer treated with open radical retropubic prostatectomy between 1987 and 2003. The seven predictors in the full model include the following categorical variables: (1) seminal vesicle involvement (svi), (2) surgical margins (sm), (3) lymph node involvement (lni) and (4) extra-capsular extension (ece), and the continuous variables: (5) prostate specific antigen (psa), (6) experience (surgery experience), and (7) post-operative Gleason sum (pgx) which is treated as an ordinal type variable. In [21], the full model is pre-specified based on medical literature reviews and clinical knowledge of investigators and surgeons prior to an analysis of the data. For the further detail of the description of the data, see [21]. Two missing values in psa are imputed using the R MICE package in the study and other variables are complete. Patients who are lost to follow-up or died from causes other than prostate cancer are right-censored. Table 1 shows the statistical description of the prostate cancer recurrence data in our study, and the estimated coefficients and statistical significance of the predictors in a multivariable Cox proportional hazards regression fitted to the entire data set for the full model and the final model built from the proposed method, which predict the 10-year probability of