Computers in Biology and Medicine 91 (2017) 159–167

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Comparison of variable selection methods for high-dimensional survival data with competing events

Julia Gilhodes [a], Christophe Zemmour [b], Soufiane Ajana [a], Alejandra Martinez [c], Jean-Pierre Delord [d], Eve Leconte [e], Jean-Marie Boher [b], Thomas Filleron [a,*]

[a] Department of Biostatistics, Institut Claudius Regaud, IUCT-O, Toulouse, France
[b] Department of Clinical Research and Investigation, Biostatistics and Methodology Unit, Institut Paoli-Calmettes, Aix Marseille University, INSERM, IRD, SESSTIM, Marseille, France
[c] Department of Surgery, Institut Claudius Regaud, IUCT-O, Toulouse, France
[d] Department of Medical Oncology, Institut Claudius Regaud, IUCT-O, Toulouse, France
[e] TSE-R, Université de Toulouse, France

ARTICLE INFO

Keywords:
Competing risks
High-dimensional data
Random survival forest
Boosting
Variable selection
Stability

ABSTRACT

Background: In the era of personalized medicine, it's primordial to identify gene signatures for each event type in the context of competing risks in order to improve risk stratification and treatment strategy. Until recently, little attention was paid to the performance of high-dimensional selection in deriving molecular signatures in this context. In this paper, we investigate the performance of two selection methods developed in the framework of high-dimensional data and competing risks: Random survival forest and a boosting approach for fitting proportional subdistribution hazards models.

Methods: Using data from bladder cancer patients (GSE5479) and simulated datasets, stability and prognosis performance of the two methods were evaluated using a resampling strategy. For each sample, the data set was split into 100 training and validation sets. Molecular signatures were developed in the training sets by the two selection methods and then applied on the corresponding validation sets.

Results: Random survival forest and boosting approach have comparable performance for the prediction of survival data, with few selected genes in common. Nevertheless, many different sets of genes are identified by the resampling approach, with a very small frequency of genes occurrence among the signatures. Also, the smaller the training sample size, the lower is the stability of the signatures.

Conclusion: Random survival forest and boosting approach give good predictive performance but gene signatures are very unstable. Further works are needed to propose adequate strategies for the analysis of high-dimensional data in the context of competing risks.

1. Introduction

Over the last decade, gene signatures based on micro-array data are on the rise in oncology [1,2]. The main objective of gene signatures is to improve the management of cancer patients by prognostication and treatment prediction [3]. Different studies demonstrated that gene signatures were not unique and strongly dependent on both the patients' selection and the regression models used [4–6]. Gene signatures are generally developed and validated using time-to-event endpoints such as metastasis free survival, disease free survival, or overall survival. As several event types are included in their definition, these endpoints can be considered as composite [7]. In order to improve risk stratification and treatment strategy, it will be interesting to identify gene signatures for each event type in the context of competing risks [8,9]. For example, loco-regional recurrence is becoming less common in breast cancer. To better guide optimal loco-regional treatment, it is important to identify gene signatures which specifically predict the risks of loco regional recurrence. Breast cancer patients are also at risk of other event types, such as distant metastasis and death, which can preclude the occurrence of loco-regional recurrence. Other various cancers can be greatly impacted by the development of genes signatures for a given event type.

Recently, several regression methods for handling high-dimensional data have been extended to the competing risk data setting. Until recently, little attention was paid to the performance of such methods in

* Corresponding author. Institut Claudius Regaud, IUCT-Oncopole, Bureau des Essais Cliniques, 1 avenue Irène Joliot Curie, 31059 Toulouse, France.
E-mail address: filleron.thomas@iuct-oncopole.fr (T. Filleron).

https://doi.org/10.1016/j.compbiomed.2017.10.021
Received 5 May 2017; Received in revised form 19 October 2017; Accepted 19 October 2017

deriving molecular signatures for predicting cumulative incidence in competing risk settings. One popular approach in the context of competing risks with high-dimensional data is to use cause specific hazard modeling. Cox proportional hazard is fitted using a penalized regression model for the event of interest and by considering individuals who fail from competing events as censored observations [10]. But a covariate that reduces the cause specific hazard of a competing risk can indirectly increase the cumulative incidence of the event of interest [11]. In fact, cumulative incidence represents the probability of disease in presence of a competing risk. For low-dimensional data, the Fine & Gray model, which is an extension of the Cox model, has been proposed to model the subdistribution hazard [12]. In high-dimensional data (number of covariates >> number of observations), the Fine and Gray model cannot be fitted to identify most predictive genes and less traditional approaches are required. Methods based on random forests have recently been adapted for survival analysis in presence of competing risks [13], with a modified weighted log-rank splitting rule modeled according to the Gray's test [14]. On the other hand, Binder et al. [15] have proposed a gene selection method based on the Fine and Gray model with a boosting approach. These different methods, now implemented in statistical packages, become increasingly popular for the analysis of competing risks data. But, to our knowledge and contrary to classical survival methods, there is no previous work which has compared these two methods on different criteria such as stability and prognostic ability.

The main objective of this publication is to compare different selection methods for high-dimensional time-to-event data in the context of competing risks using a published data set on bladder cancer and simulated datasets. After presenting an example of the application of these methods on the former, a resampling strategy was performed to evaluate both gene selection and predictive accuracy and to explore the effect of the training set sample size on the performance.

2. Patients and methods

2.1. General principles: competing risks setting

Fundamentals of competing risks have been extensively reviewed in the literature [11,16,17]. In a competing risks setting, patients are at risk for different event types (for example k). We only observed the pair of variables (Y, Δ) where Y corresponds to the time to first event (or last follow-up news) and Δ the type of first event:

$$\Delta = \begin{cases} 0, & censored \\ 1, & \text{event of type1} \\ \dots & \\ k, & \text{event of type k} \end{cases}$$

One quantity of interest is the cumulative incidence of event k, denoted $\mathrm{F_k(t)}$, which corresponds to the probability of event k before time t in the presence of competing events [18]. The corresponding mathematical expression is:

$$\mathrm{F_k(t)} = \Pr[Y \leq t, \Delta = k] = \int_0^t S(t)\lambda_k(t)dt$$

with $S(t)$ the probability of not having failed from some event estimated by Kaplan Meier and $\lambda_k(t)$ the cause specific hazard of event k. In order to compare cumulative incidence associated with each event type, a k-sample test is proposed by Gray [14].

To evaluate influence of covariate, Fine & gray proposed to model the subdistribution hazard of event k by Ref. [15]:

$$h_k(t|x_i) = h_{k,0}(t)\exp\left(x_i'\beta\right)$$

with $\mathrm{h_{k,0}}(t)$ an unspecified baseline subdistribution hazard of event k, x_i the vector of covariates and β the vector of the regression models coefficients.

2.2. Bladder cancer dataset

The public data set GSE5479 has been downloaded from the platform GEO, corresponding to 1381 preprocessed custom microarrays from 404 non-muscle invasive bladder cancer samples used by Dyrskjøt et al. [19]. Only patients for whom original progression classifier and clinical covariates were available (age, sex, stage, grade and treatment) were included in our study (n = 301). The event of interest, which is progression or death due to bladder cancer, occurred for 84 patients (11 events after 5 years). The competing event, death due to another cause or unknown cause, was observed in 33 patients and 184 patients were censored at the last follow-up time. Cumulative incidences obtained using Prentice estimators were respectively 26.8% and 11.2% at 5 years for event of interest and competing event.

2.3. Simulated datasets

Using the algorithm employed by Binder et al. and Tapak et al. [10,15], four datasets have been simulated with different sample sizes. For each dataset, we have generated two competing events and 1500 covariates normally distributed. Among the 1500 covariates, sixteen of them were informative and selected from three blocks of correlated covariates:

- four have an increasing effect on both type 1 and 2 hazards (block 1: correlated genes with correlation in block equals to 0.5)
- four have an increasing effect on the type 1 hazard and a decreasing effect on the type 2 hazard (block 2: correlated genes with correlation in block equals to 0.35)
- four have a decreasing effect on the type 1 hazard only and four others have an increasing effect on type 2 only (block 3: correlation equals to 0.05)

Among these sixteen covariates, the true vectors of coefficients β_1 for event 1 and β_2 for event 2 take the following values:

$$\beta_1 = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, -0.5, -0.5, -0.5, -0.5, 0, 0, 0, 0)$$

$$\beta_2 = (0.5, 0.5, 0.5, 0.5, -0.5, -0.5, -0.5, -0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5)$$

The remaining covariates have no direct effect on both hazards, also true coefficients were set to 0. For each event, survival times were generated using cause specific exponential model. Censoring times follow a uniform distribution U[0;9], resulting in a censoring rate of ≈35%. Parameters used to simulate each dataset are summarized in Supp Table 1.

2.4. Learning methods for classification

This section briefly describes the two selection methods investigated in this publication. A more detailed mathematical presentation can be found in previous works.

2.5. Random Survival forests

Random Survival forests (RSF) is a non-parametric method of variable selection for right-censored survival data, which was introduced by Ishwaran et al., in 2008 [20], and then adapted for competing risks by the same authors in 2014 [13]. Main steps of the algorithm are presented Fig. 1A. Random Forests are induced from bootstrap samples of the training set with modified weighted log-rank splitting rule according to Gray's test(14). About 37% of the data are excluded in each bootstrap sample; these are the so-called “out-of-bag data”. The remaining data are used to build each tree. Variable hunting is used for the random forest variable selection, which is a combination of “variable importance”