Introduction

Collection of large number of clinical and genomic data for individuals are initiated with the hope of finding clinically significant diagnostic and prognostic factors for diseases. In studies involving progression in joint damage in psoriatic arthritis, for example, interests may lie in detecting HLA alleles associated with different disease courses [1]. Along with the genomic data in clinical cohort studies, it is common to have time-to-event end points resulting in right-censored data [2]. One of the objectives in these studies is to predict the future course of the disease at population or patient level based on covariate information.

Numerous parametric, semi-parametric and non-parametric methods have been developed in the literature. Cox proportional hazards model is well-establish and popular among those approaches for its flexibility and simplicity [3]. Several extensions have been implemented such as the penalized Cox and boosting in Cox regression to accommodate variable selection or prediction [2,4]. Random survival forest (RSF) is a non-parametric ensemble tree method which extends Breiman’s random forest to right-censored data [5]. RSF serves as an attractive alternative in handling data consist of more covariates than subjects, complex and nonlinear relationships between response and covariates and non-proportionality of hazards [6,7]. In this note we give an overview on RSF in a nontechnical manner. Specifically, we will discuss the three main aspects in tree-based estimation procedures in Section 2 and end with a brief discussion in Section 3.

Tree-Based Methods

A tree is composed of numerous nodes as illustrated in (Figure 1). Tree estimation is in general based on recursively performing binary partitioning on the covariate space using some pre-defined splitting rule. The result is a collection of candidates ‘nodes’, starting from the top (single-node) of the tree to a number of terminal nodes or leafs [8]. RF is an ensemble of trees, hence the term forest, where the final prediction is averaged over all trees in the forest. Each tree is non-deterministic as the tree is grown on a subspace of individuals who were picked from bootstrapping the whole dataset [9]. Growing a single tree is well known to exhibit high variances in predicted outcomes. By combining the trees, however, variance as well as bias in prediction can be substantially decreased [10].
There are three main aspects in tree-based estimation procedures, namely (1) the node splitting rule to partition the covariate space, (2) selection on the size of the tree to avoid overfitting and (3) estimation in the terminal nodes once the tree has grown to full size. Splitting is usually done via evaluation of the loss function to accommodate (1) and pruning and cross-validation is typically done for (2) [6]. In the setting with time-to-event data there are two splitting rules that are mostly adopted in growing survival trees: the log-rank splitting and log-rank score splitting. Precise details on the two splitting rules can be found in [6,11]. The growing process is terminated when the tree is grown to full size, i.e., when the terminating threshold has been met. The terminating threshold could be defined as, for example, a user pre-specified minimum observations per terminal node or when further splitting to daughter nodes is no better than mother node. When the tree is grown to full size, each terminal node within the survival tree is made up of a set of individuals that were ‘dropped down the tree’ via the same route and cumulative hazard estimates by Nelson-Aalen estimator may be obtained from each terminal nodes [12]. The final estimate of the cumulative hazard (CH) is thus an average measure of CH from all survival trees grown in the RSF.

Since the mechanisms with regard to the construction of the RSF is unknown, interpretability is thus burdensome. However, the RSF does output the variable importance measure (VIMP) which measures the increase (or decrease) in prediction error for the forest ensemble [5]. While the idea of VIMP has good empirical performance, it is informally defined and remained somewhat ad hoc. More recently, [6] proposed the minimal depth (MD) measure which formalized the process of variable selection. Splitting and variable selection is traditionally done simultaneously, and it has been noted that such an approach is prone to overfitting and selection bias towards covariates with many possible splits [13]. Considering selection bias is important in survival tree construction because the biased selection would lead to the wrong summary measures such as VIMP and MD. [13] proposed a conditional Inference framework where he separated variable selection and splitting into a two-stage procedure and has shown that such an approach has a better performance, especially when the data consists of covariates with many split-points. More recently, [14] had proposed using maximally selected rank statistics in order to have unbiased variable splitting. Intensive simulation studies on this issue has also been demonstrated by [14,15].

**Discussion**

RSF is an attractive method when the goal is to do prediction. Its advantage is more apparent when the dimension of covariates is large, relationship between response and covariates are complex or when the proportional hazard assumption is at risk [8]. Although RSF acts as a great alternative in analyzing time-to-event data, care must be taken when choosing and tuning the trees based on the form of available data.

**References**