Artificial Intelligence in Medicine 49 (2010) 33–42

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A machine learning-based approach to prognostic analysis of thoracic transplantations

Dursun Delen [a,*], Asil Oztekin [b,c], Zhenyu (James) Kong [b]

[a] *Spears School of Business, Oklahoma State University, T-NCB 378, 700 North Greenwood Avenue, Tulsa, OK, 74106, USA*
[b] *School of Industrial Engineering and Management, Oklahoma State University, 322 Engineering North, Stillwater, OK 74078, USA*
[c] *Department of Industrial Engineering, Gediz University, 35230 Cankaya-Izmir, Turkey*

ARTICLE INFO

Article history:
Received 25 February 2009
Received in revised form 15 December 2009
Accepted 10 January 2010

Keywords:
Data mining
Machine learning
UNOS
Thoracic Transplantation
Survival analysis
Prognostic index

ABSTRACT

Objective: The prediction of survival time after organ transplantations and prognosis analysis of different risk groups of transplant patients are not only clinically important but also technically challenging. The current studies, which are mostly linear modeling-based statistical analyses, have focused on small sets of disparate predictive factors where many potentially important variables are neglected in their analyses. Data mining methods, such as machine learning-based approaches, are capable of providing an effective way of overcoming these limitations by utilizing sufficiently large data sets with many predictive factors to identify not only linear associations but also highly complex, non-linear relationships. Therefore, this study is aimed at exploring risk groups of thoracic recipients through machine learning-based methods.

Methods and material: A large, feature-rich, nation-wide thoracic transplantation dataset (obtained from the United Network for Organ Sharing—UNOS) is used to develop predictive models for the survival time estimation. The predictive factors that are most relevant to the survival time identified via, (1) conducting sensitivity analysis on models developed by the machine learning methods, (2) extraction of variables from the published literature, and (3) eliciting variables from the medical experts and other domain specific knowledge bases. A unified set of predictors is then used to develop a Cox regression model and the related prognosis indices. A comparison of clustering algorithm-based and conventional risk grouping techniques is conducted based on the outcome of the Cox regression model in order to identify optimal number of risk groups of thoracic recipients. Finally, the Kaplan–Meier survival analysis is performed to validate the discrimination among the identified various risk groups.

Results: The machine learning models performed very effectively in predicting the survival time: the support vector machine model with a radial basis Kernel function produced the best fit with an R^2 value of 0.879, the artificial neural network (multilayer perceptron-MLP-model) came the second with an R^2 value of 0.847, and the M5 algorithm-based regression tree model came last with an R^2 value of 0.785. Following the proposed method, a consolidated set of predictive variables are determined and used to build the Cox survival model. Using the prognosis indices revealed by the Cox survival model along with a *k*-means clustering algorithm, an optimal number of “three” risk groups is identified. The significance of differences among these risk groups are also validated using the Kaplan–Meier survival analysis.

Conclusions: This study demonstrated that the integrated machine learning method to select the predictor variables is more effective in developing the Cox survival models than the traditional methods commonly found in the literature. The significant distinction among the risk groups of thoracic patients also validates the effectiveness of the methodology proposed herein. We anticipate that this study (and other AI based analytic studies like this one) will lead to more effective analyses of thoracic transplant procedures to better understand the prognosis of thoracic organ recipients. It would potentially lead to new medical and biological advances and more effective allocation policies in the field of organ transplantation.

1. Introduction

1.1. Motivation

Thoracic (heart and lung) transplantation has been accepted as a viable treatment for end-stage cardiac and pulmonary failure. The

[*] Corresponding author. Tel.: +1 918 594 8283; fax: +1 918 594 8283.
E-mail address: dursun.delen@okstate.edu (D. Delen).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.01.002

increased experience in cardiac and pulmonary transplantation, improvements in patient selection, organ preservation, and preoperative support have significantly reduced the early threats to patient survival [1]. Over the past decade, the thoracic transplant waiting time for a listed patient has markedly increased, but the number of transplants performed has declined. In addition, the research also found that there is a perceived inequity in access to organs. The organ allocation system needs to be improved since it may become a major factor negatively influencing the survivability of thoracic transplant [2].

The survivability prediction is becoming increasingly more important in medicine. When a resource is scarce, the need for accurate prediction becomes acute [3]. Especially *prediction of survival time* and *prognosis prediction of medical treatments* are clinically important and challenging problems [4]. Scarceness of organs necessitates the development of effective and efficient procedures to select the most optimal organ receiver since demand for organs of all patients might not be satisfied. To achieve this, one critical step is to reveal the knowledge underlying huge amount of data collected and stored from organ transplantation procedures performed in the past. The objectives are (1) to maximize the patients' survival time after the organ transplantation surgery, and (2) to optimize the prognosis for the organ recipients. These can be potentially achieved by discovering the knowledge that may be contained in large dataset consisting of more than hundreds of determinative variables regarding the donors, the potential recipients, and transplantation procedures. Therefore, in this study a data mining method is proposed to process large amount of transplantation data obtained from UNOS to identify the important factors as well as their relationships to the survival of the graft and the patient. Thereafter, a prognostic index [5,6] is developed to classify the patients into different risk groups for better understanding of the transplantation phenomenon. In short, this study will address the following questions: (1) what are the most important variables to be included in an effective prognostic index related to thoracic organ transplantations? (2) what are the most coherent risk groups that can be formed based on the prognostic index? Predicting the thoracic survivability and classifying the patients (potential thoracic organ receivers) into different classes of risks would help decision makers in determining patients' priority for transplantation source assignment.

1.2. Literature review

1.2.1. Related research in survival analysis for organ transplantation

In the recent past, a number of studies were conducted using data-driven analytics on various organ transplantation datasets. Closely related to the study reported herein, Hariharan et al. [7] focused on the analysis of improved graft survival rate using cyclosporine after renal transplantation in both short-term (less than 1 year) and long-term (more than 1 year). A regression analysis was used to predict the probability of the graft failure after kidney transplantation in both short-term and long-term period in the light of demographic characteristics, transplant-related variables, and post-transplantation variables. The study performed by Herrero et al. [8] included 116 patients who received a liver transplant between the years 1994 and 2000. Statistical tests are used to compare the demographic and characteristic variables, pretransplant, and intra-operative variables between the two groups, namely younger and older than 60. The results indicate that there is a clear trend showing that older patients have lower survival after liver transplantation. Hong et al. [9] presented a survival analysis of liver transplant patients in Canada by considering some factors such as age, blood type, donor type (cadaveric or alive), race, and gender of recipient and donors. However, having limited the variables with this scope, they also admitted that the clinical information lacks of many potential details.

Taking a data mining approach, Kusiak et al. [10] compared two rule-based data mining techniques, i.e. decision trees and rough sets, to predict survival time of kidney dialysis patients. This study achieved satisfactorily high prediction accuracy. The main limitation of the study was the utilization of a small dataset with only 188 patients in total and also many patient-related parameters were neglected in the problem formulation. Using more traditional methods, and specifically having focused on thoracic transplantation, Jenkins et al. [11] and Fernandez-Yanez et al. [12] had a rich pool of independent variables for survivability prediction. Their studies used popular statistical techniques such as Kaplan–Meier method of survival analysis with Mantel–Haenszel log-rank test. However, both of these techniques have been criticized with two major limitations: (1) linear relationships are assumed, which hence cannot capture the nonlinearity among the variables, and (2) the independent variables were selected solely based on the experiences and intuitions of the analysts who conducted these studies. Thus, many potentially significant variables might be left outside the scope of this study. Tjang et al. [13] added more explanatory variables to determine the survivability in heart transplantation, such as body mass index, waiting time on the list, and previous cardiac surgery, their study also ignored the nonlinear relationships among the pool of survivability-related variables. Similar limitations exist in some other studies focused directly or indirectly on thoracic transplantation [14–16].

The existing studies implicitly assume that the relationships among the predictive variables and output variable are linear and the predictor variables are independent of each other, which may not be valid in reality. Moreover, the abovementioned studies focus on small datasets with limited number of predictors for survivability of patients after transplantation. This limitation may cause incomprehensive modeling due to the insufficient information contents (i.e., omission of a number of potentially important predictor variables).

1.2.2. Related research in devising a prognostic index

Prognostic index (PI) provides compact prognosis information regarding a specific patient based on the results of a Cox proportional hazards model [5]. Cox proportional hazards model helps identify variables of prognostic importance and hence prognostic index can be used to define groups of individuals at different risk categories. Even though prognostic index is a convenient tool to measure how well the patients are doing after the transplantation, its use in the organ transplantation area has been limited mostly due to the lack of follow-up data. Some existing studies related to devising a PI in transplant area are summarized as follows.

In the study conducted by Christensen et al. [17], it is mentioned that primary biliary cirrhosis requires a liver transplantation operation at the end stage. Based on the prognosis analysis with as well as without transplantation, it is decided whether or not the transplantation is required, if so when. To achieve this goal, corresponding PIs and probabilities of surviving are computed for transplantation and non-transplantation cases. Yoo et al. [18] developed a similar index and revealed that socioeconomic status does not influence patient or graft survival that undergoes liver transplantation at the institute where they performed their study. Deng et al. [19] conducted a study with a national dataset in Germany, which discovers the effect of receiving a heart transplant for the patients in a waiting list. The results indicate that cardiac transplant is associated with survival benefit only for patients with a predicted high risk of dying on the waiting list. Ghobrial et al. [20] performed a study to determine prognostic factors for overall survival in 107 adult patients with post-transplantation lymphoproliferative disorders (PTLDs). It is validated that in discriminating the low and high scored patients the proposed prognostic scoring significantly performs better than the International Prognostic Index for the subset of the patients (56 out of 107) with lactate dehydrogenase.

The common limitation in all of these studies is similar to the limitations of the studies summarized in Section 1.2.1. Namely, they directly devise a prognostic index without determining if the variables used in prognostic index devising phase are necessary and sufficient. This motivates a machine learning-based initial step of variable selection procedure. Because, if the critical predictive factors are not captured effectively due to the intuition- and experience-based selection, the resulting prognostic indices developed based on the selected variables would be inaccurate and, in turn, related risk groups of patients would be deviated from the real classes. This may cause mistakes for decision maker in making organ transplantation policies.

2. Proposed method

Section 1.2 shows that the most of the existing studies for organ transplantation procedures utilize conventional statistical approaches such as Kaplan–Meier function and log-rank test along with expert-selected variables to predict the survivability. However, organ transplantation procedures consist of a large number of variables (several hundred) that may have nontrivial impact on modeling the prognosis of the grafts/patients. Using a somewhat comprehensive variable list may help discriminate patients from each other by placing them into proper risk groups. Unintentional omission of the important variables may lead to inaccurate classification of patient risk groups, which may, in turn, lead to suboptimal organ allocation policies and ineffective treatments.

This study is aimed at overcoming the abovementioned shortcomings by employing both machine learning techniques as well as statistical methods to identify the most critical factors affecting the survivability of thoracic transplant patients. To achieve this goal, this study proposes adopting a 5-step approach illustrated in Fig. 1. Step 1 involves data understanding and preparation, which is arguably the most time demanding step in the process. Step 2 employs various predictive modeling techniques such as support vector machines, artificial neural networks, and regression trees to develop survival time prediction models and to extract the most important variables by means of sensitivity analysis through the best performing model. Step 3 determines the consolidated candidate set of critical predictor variables. Step 4 develops a Cox regression model using the consolidated set of predictor variables and also devises a prognostic index. The last step, Step 5, classifies the patients into various risk categories by comparing and contrasting the clustering performance of algorithm-based and manually calculated groups. Then the resulting risk categories are validated by using the Kaplan–Meier survival curves. These steps will be further explained in details in Sections 2.1–2.5, respectively.

2.1. Step one: data source and data preparation

In this study, the data source that was used to validate the proposed method was thoracic organ transplant dataset provided by UNOS, which is a tax-exempt, medical, scientific, and educational organization that operates the national Organ Procurement and Transplantation Network under the contract to the Division of Organ Transplantation of the Department of Health and Human Services [21]. The data files were obtained from UNOS using a formal data requisition procedure (which includes submission of specific data needs, purpose of the study, and a data use agreement). These data files are named as UNOS Standard Transplant Analysis and Research (STAR) files for heart, lung, and simultaneous heart–lung transplants, namely thoracic transplants. Each transplant STAR file consists of information on all thoracic transplants that had been performed in the US and reported to UNOS since October 1, 1987. It includes both deceased- and living-donor transplants. None of the files include any specific patient or transplant hospital identifiers due to the privacy and security issues. However, there is a patient identification number, unique to each patient, which allows linking multiple files and tracking the patient. Considering these features, UNOS's data files are recognized as the most comprehensive source of information available in any single field of medicine and for organ transplantation in US [22].

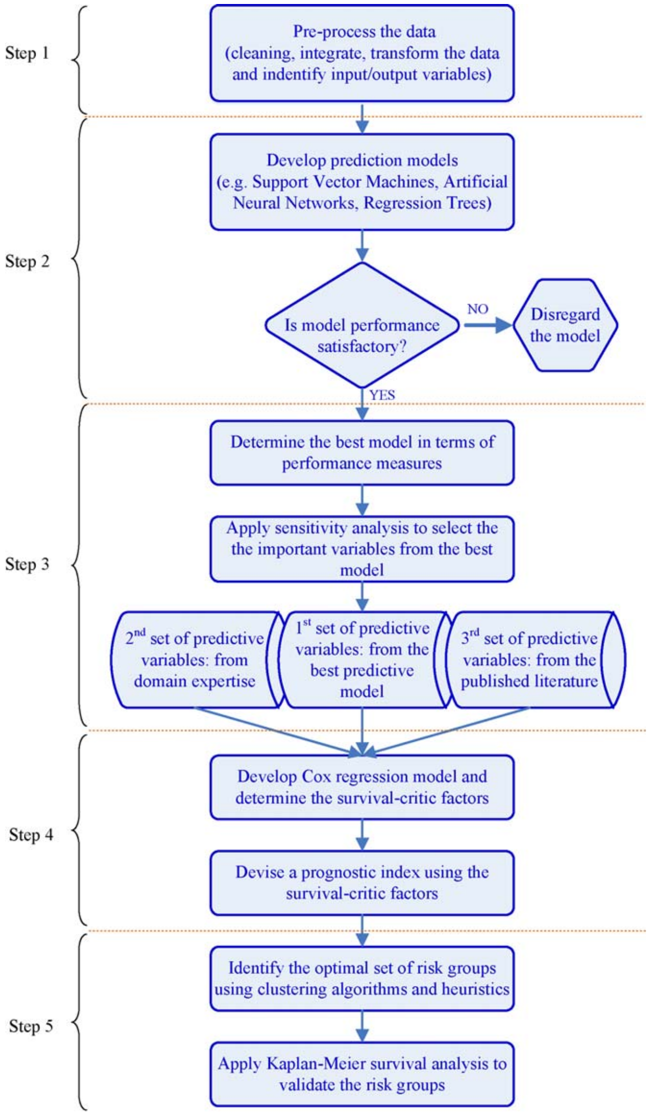
Step 1
Pre-process the data (cleaning, integrate, transform the data and indentify input/output variables)
Step 2
Develop prediction models (e.g. Support Vector Machines, Artificial Neural Networks, Regression Trees)
Is model performance satisfactory?
NO
Disregard the model
YES
Step 3
Determine the best model in terms of performance measures
Apply sensitivity analysis to select the the important variables from the best model
2nd set of predictive variables: from domain expertise
1st set of predictive variables: from the best predictive model
3rd set of predictive variables: from the published literature
Step 4
Develop Cox regression model and determine the survival-critic factors
Devise a prognostic index using the survival-critic factors
Step 5
Identify the optimal set of risk groups using clustering algorithms and heuristics
Apply Kaplan-Meier survival analysis to validate the risk groups

Fig. 1. A flowchart representation of the proposed method.

There are two datasets involved in our study, which are *regular dataset* and *follow-up dataset*. The regular dataset contains all information of donors and recipients before transplantation occurred, and the follow-up dataset provides all information of donors and recipients after the transplantation. The TRR_ID variable (transplant identifier) is the common variable between these two datasets and the one which is proposed by UNOS to merge and integrate these two datasets. Therefore, these two datasets were combined in a relational database environment using the link (a.k.a. primary key) of TRR_ID.

Overall, the complete dataset consists of 310,773 records and 565 variables. These variables include the socio-demographic and health-related factors with regard to both the donor and the recipients. There are also procedure-related factors among the dataset. To assign as an output (dependent variable), there are four possible variables which are called *pstatus*, *ptime*, *gstatus*, and *gtime*. These variables have the following meanings: whether or not the patient died after transplantation occurred (referring to *pstatus*, with dead = 1 and alive = 0). A very similar variable was *gstatus*, referring to whether or not graft has failed (1 denoting "failed" and 0 denoting "succeeded"). The variable *ptime* denoted patient follow-up time (in days) from transplant to death/last follow up time. Similarly, *gtime* is explained as graft lifespan from transplant to death/last follow up time. Since the goal of this study is to develop models to predict the survivability solely based on thoracic transplant, the dependent variable was assigned as *gtime*. This assignment was done to discriminate the patients who died solely due to the thoracic graft incompatibility from the ones who died from any other reasons. Therefore, the rest of the potential dependent variables (pstatus and ptime) were eliminated from the dataset. Besides, *gstatus* was kept inactive up to the stage where Cox regression model was implemented (Step 4 in Fig. 1).

Considering the *gtime* as the continuous dependent variable, the records for the patients whose *gtime* information were missing were removed from the dataset. The data set also includes some identification variables (e.g., Donor ID) which help track the recipient patient anonymously, track the thoracic transplant procedure, or link records from multiple data files to each other. Since these types of identification variables do not have any information content to enhance the prediction capability of the models, after linking and integrating the files they were also excluded from the analysis dataset. Moreover, the name of transplantation type was recorded in the dataset as a variable named *Dataset* which had one value (TH referring to thoracic) and the date of data processing is recorded as a variable named *Date of Run* which are useful for data integration purposes but has no information for contributing to the prediction of survivability and hence are also excluded from the analysis dataset. Similarly, other variables having only one possible value for all records in the dataset, which have no discriminating information, are also eliminated from the predictive modeling.

This dataset had excessive number of missing values which render most of the records and variables seemingly insignificant. However, in data mining studies one should be very reluctant to remove the candidate predictor variables while at the same time trying to avoid artificial data imputation procedures. There is an obvious trade-off here. As a rule of thumb, for column (variable) deletion, we were cautious to remove any variable from the analysis and assumed that if a variable has more than 95% missing values, only then it should be regarded as not having significant information content and hence should be deleted. Next step was to handle the missing values by following the general convention: for the categorical variables we filled the missing values with some heuristic values such as E (referring to empty) or NR (referring to not reported), and for the continuous variables we imputed the missing values with the average of the existing records. After adopting these data preparation strategies, the final dataset was reduced to 372 cleansed independent variables and one dependent variable (*gtime*) with the total record count of 106,398.

2.2. Step two: predictive modeling

Since the dependent variable herein was a continuous variable (graft survival time, which is the number of days from transplant to death or last follow-up), the problem refers to a *prediction* (or regression) problem (as opposed to a classification problem). Since the relationships between the dependent variable and the independent variables were not known in advance, this step was to develop various predictive models for graft survival time using all of the available independent variables. It is also required to check whether the models have passed the pre-specified threshold values of performance measures, specifically the R^2 and mean square error (MSE), to determine the *best* model that explains these unknown relationships between dependent and independent variables by ranking them according to these measures. The model which is deemed to be the most successful one would be kept for further modeling steps to determine the importance of the independent variables.

Support vector machines (SVMs) are supervised learning methods that generate input–output mapping functions from a set of training data. They belong to a family of generalized linear models which achieve a classification or regression decision based on the value of the linear combination of features. They are also said to belong to the kernel methods [23]. The mapping function in SVMs can be either a classification function (used to categorize the data) or a regression function (used to estimate the numerical value of the desired output, as is the case in this study). Nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data become more separable (i.e. linearly separable) compared to the original input space. Then, maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes, the better the generalization error of the prediction would be.

Artificial neural networks (ANNs) have been utilized to model complex relationships (such as nonlinear functions and multicollinearity) among the predictor variables and the dependent variable [24]. ANNs are highly sophisticated analytic techniques capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called "learning" from existing data [25]. ANNs have been one of the most popular artificial intelligence-based data modeling algorithms used in recent medical informatics studies due to their satisfactory predictive performance [26]. On the other hand, compared to other machine learning methods (such as ANNs), decision trees have the advantage of not being a black box model, namely having the capability to explain the inner structure of the model in the form of a graphically represented inverse tree or a collection of condition-action rules. This advantage has made them a viable and desirable alternative method in medical informatics [27]. If the dependent variable is continuous (as in the case in this study) the resulting decision tree is called a regression tree. Regression trees are known to be among the highly adaptable, relatively flexible, yet computationally intensive data mining techniques [28]. Popular regression tree algorithms are CART (or C&RT) [29], CHAID [30], and M5 [31] which can be used for both classification and regression type prediction problems. ID3 and its successors, C4.5 and C5 are also among the popular decision tree algorithms, but they can only work for classification type prediction problems.

2.2.1. Performance criteria

To compare the abovementioned prediction models, two performance criteria are considered: mean squared error (MSE) of the model on testing dataset and R^2 value between the actual observation for the target variable (Y_t) and the predicted value by the model (F_t). MSE which is given by Eq. (1) does not have a rule-of-thumb threshold cut-off value for acceptable models. It is a relative criterion to select the best model, namely the smaller the value the better the model has performed [32].

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2 \tag{1}$$

On the other hand, R^2 ($R^2{}_{F_t,Y_t}$ or shortly R^2) which is given by Eq. (2) can be considered as both an absolute measure and a relative measure to determine and rank the satisfactory models [33]. Unlike the MSE, the higher the R^2, the better the performance for the compared models.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y}_t)^2} \tag{2}$$

2.2.2. k-Fold cross-validation

In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, researchers tend to use *k*-fold cross-validation [34]. In *k*-fold cross-validation, also called rotation estimation, the complete dataset (*D*) is randomly split into *k* mutually exclusive subsets (the folds: D_1, D_2, . . ., D_k) of approximately equal size. The prediction model is trained and tested *k* times. Each time ($t \in \{1, 2, \ldots, k\}$), it is trained on all but one fold (D_t) and tested on the remaining single fold (D_t). The cross-validation estimate of the overall performance criteria is calculated as simply the average of the *k* individual performance measures as in Eq. (3),

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \tag{3}$$

where *CV* stands for cross-validation, *k* is the number of folds used, and *PM* is the performance measure for each fold [35].

In this study, to estimate the performance of the prediction models a 10-fold cross-validation approach was used. Empirical studies showed that 10 seems to be an optimal number of folds (that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process) [34]. In 10-fold cross-validation the entire dataset is divided into 10 mutually exclusive subsets (or folds). Each fold is used once to test the performance of the prediction model that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates.

2.2.3. Sensitivity analysis

After selecting the best prediction model based on the performance criteria as explained in Section 2.2.1, it is required to determine the importance of the independent variables. In machine learning algorithms, sensitivity analysis is a method for extracting the cause and effect relationship between the inputs and outputs of a trained model [36]. In the process of performing sensitivity analysis, after the model is trained the learning is disabled so that the network weights are not affected. The fundamental idea is that the sensitivity analysis measures the predictor variables based on the change in modeling performance that occurs if a predictor variable is not included in the model. Hence, the measure of sensitivity of a specific predictor variable is the ratio of the error of the trained model without the predictor variable to the error of the model that includes this predictor variable [37]. The more sensitive the network is to a particular variable, the greater the performance decrease would be in the absence of that variable, and therefore the greater the ratio of importance. This method is followed in support vector machines and artificial neural networks to rank the variables in terms of their importance according to the sensitivity measure defined in Eq. (4) [38].

$$S_i = \frac{V}{C(F_t)} = \frac{V(E(F_t|X_i))}{V(F_t)} \tag{4}$$

where $V(F_t)$ is the unconditional output variance. In the numerator, the expectation operator *E* calls for an integral over X_{-i}; that is, over all input variables but X_i, then the variance operator V implies a further integral over X_i. Variable importance is then computed as the normalized sensitivity. Saltelli et al. [39] show that Eq. (4) is the proper measure of sensitivity to rank the predictors in order of importance for any combination of interaction and non-orthogonality among predictors. As for the decision trees, variable importance measures were used to judge the relative importance of each predictor variable. Variable importance ranking uses surrogate splitting to produce a scale which is a relative importance measure for each predictor variable included in the analysis. Further details on this procedure can be seen in Breiman et al. [29].

2.3. Step three: determining the candidate sets of predictor variables

Step 3 is to determine which predictor variables to be used in devising a prognostic index in Step 4. This step helps eliminate the insignificant variables and improves the accuracy of the model by optimizing the predictor variables list. The potential input variables to this step consist of three candidate sets of predictor variables. The first set is composed of variables selected by the predictive models. The predictive models explained in Section 2.2 rank the predictor variables based on their importance level in predicting the graft survival time. The predictive variables selected by the sensitivity analysis of the best-performing model (ranked in terms of R^2 and MSE) are chosen as the first set of predictive variables. The second set of predictive variables is obtained by considering the expert domain knowledge. This set includes variables which are logically related to heart and lung transplantation such as donor's history of cigarette usage. The third set of predictive variables is selected from the related literature. This set consists of the variables which have been commonly and repeatedly used in previous studies in the organ transplantation area. The second and third sets of predictive variables provide more comprehensive information for the next step, the Cox regression model, by including the variables that might have importance in the survival analysis but were determined to be insignificant by the predictive models in Step 2.

2.4. Step four: survival analysis and prognostic index devising

Step 4 takes all the three sets of predictive variables identified in Step 3, and then applies Cox regression to model the graft survivability and filter out the candidate predictive variables which do not have significant survival effect. Hence, in Step 4 the final critical predictive variables are determined by the Cox regression model. Cox regression model also enables devising a prognostic index to categorize the patients into various groups with different levels of risks.

Cox regression model is a semi-parametric model extensively used in survival analysis [6]. The survival time of each patient is assumed to follow the hazard function (h_i) given by Eq. (5) as follows:

$$h_i = h_0 \exp(x_i\beta) \tag{5}$$

where h_0 is the baseline hazard function and x_i is the vector of predictor variables for the *i*th patient. β is the vector of regression coefficients for the predictor variables and is assumed to be the same for all patients [40,41].

One important application of Cox regression model is to identify variables which may be of prognostic importance [5]. Once identified, knowledge from these variables will be combined and used to define a prognostic index, which in turn defines groups of organ recipients with different levels of risk. To use the prognostic index, key patient characteristics are recorded, from which a score is derived. This score gives an indication of whether a particular patient has high, intermediate, or low levels of prognosis for the disease [5,42]. Recalling Eq. (5), the prognostic index (PI) for each patient can be calculated by Eq. (6):

$$\mathrm{PI} = x_1\beta_1 + x_2\beta_2 + \ldots + x_n\beta_n \tag{6}$$

where x_1 to x_n are the patient's values for the variables in the Cox model, and β_1 to β_n are the corresponding regression coefficients determined by Cox regression model [42].

Note that PI in Eq. (6) represents the exponent portion in Eq. (5). Therefore, the smaller the PI, the smaller the hazard function value, and hence the smaller the risk associated with a particular recipient.

2.5. *Step five: determining risk groups of thoracic recipients*

An important question following Step 4 is "How many risk groups should the patients be classified into?" In Step 5, *k*-means clustering algorithm, two-step cluster analysis, and conventional heuristics-based approaches are used to answer to this question. As a statistical and/or pictorial verification mechanism for the number of groups determined by the best performing abovementioned clustering approaches, finally the Kaplan–Meier survival analysis [43] is adopted and corresponding survival curves are generated.

k-Means method is an extensively used, arguably the most popular clustering algorithm that searches for a nearly optimal partition with fixed number of clusters represented by the parameter *k* [44]. It proceeds by assigning *k* initial centroids to the multidimensional datasets. Each record in the dataset is allocated to the centroid which is nearest and hence forming a cluster. Each cluster centroid is then updated to be the center of its members, followed by a new assignment of records to the nearest centroids to re-construct the clusters. The algorithm converges when there is no further change in allocation of members to clusters or some predefined time-based stopping criteria is satisfied [45].

Another popular clustering algorithm is two-step cluster analysis (TSCA) [46,47]. It has two steps: (1) to *pre-cluster* the cases (or records) into many small sub-clusters, and (2) to *cluster* the sub-clusters resulting from pre-cluster step into the desired number of clusters. The *pre-cluster step* uses a sequential clustering approach. It scans the data records one by one and decides if the current record should be merged with the previously formed clusters or starts a new cluster based on the distance criterion. Then the *cluster step* takes sub-clusters resulting from the pre-cluster step as input, and groups them into the desired number of clusters. Since the number of sub-clusters is much less than the number of original records, the traditional clustering methods can be used effectively. This step uses the agglomerative hierarchical clustering method [46,47]. Although there are several other clustering algorithms (e.g. Kohonen networks) they do not allow the modeler to specify a desired number of clusters at the beginning of the clustering algorithm. *k*-Means and TSCA algorithms overcome this issue. The modeler can predefine a specific number of clusters to group the variables and compare them according to their clustering performances. Since this is the main focus of our study, we utilized *k*-means and TSCA algorithms for clustering the PIs and thus identfying the risk groups of thoracic patients.

The Kaplan–Meier analysis is a non-parametric technique used to test the statistical significance of differences between the survival curves associated with two different circumstances [43]. The analysis expresses the distribution of patient survival times in terms of the proportion of patients still alive up to a given time. On the other hand, the Kaplan–Meier survival curves plot the proportion of patients surviving against time which has a characteristic decline. In biostatistics, a typical application of Kaplan–Meier survival curves involves grouping patients into risk groups such as low, medium, and high risks.

3. The case study and discussion

In order to demonstrate and validate the proposed methodology in Section 2, two most popular data mining toolkit are used, namely SPSS PASW Modeler® [48] and SAS 9.1.3® [49] statistical software package. Using the UNOS data set, Sections 3.1–3.5 discuss the results obtained by following the above mentioned modeling procedures presented in Section 2. The prediction performance results reported herein are all based on the test (or holdout) dataset.

3.1. *Predictive model results*

To reveal the initially unknown relationship between the thoracic input/independent variables and the continuous output/dependent variable (*gtime*), due to the high computational time required for 10-fold cross-validation of each model we only used two most popular models from each family of machine learning techniques. Radial basis function (RBF) and polynomial functions as Kernel methods in support vector machine were deployed. We used multilayer perceptron (MLP) and RBF type of network structures for ANNs. The most recent algorithms C&RT and M5 were utilized for prediction with the decision trees. The 10-fold averaged prediction results in terms of MSE and R^2 for each model are tabulated in Table 1. The acceptance of predictive models is first evaluated based on their coefficient of determination (R^2) values. It is widely accepted that if R^2 is higher than 0.6, the predictive model has performed fairly well [50,51]. Therefore, we set this as a threshold value for the model sufficiency. Since all the models have passed this threshold, we kept the one with the highest R^2 and the smallest MSE for further analyses, which came out to be the support vector machine model with radial basis Kernel function in this case study.

3.2. *Determination of the candidate covariates for Cox regression model*

Step 3 in the proposed method provides three different sets of candidate covariates to be used in the Cox model. Since the best performing model to explain the relationships of independent and

Table 1
Comparison of machine learning prediction model results.

	Performance measures	
Prediction models	MSE	R^2
Support vector machine		
RBF	0.023	0.879
Polynomial	0.793	0.643
Artificial neural network		
MLP	0.031	0.847
RBF	0.146	0.835
Decision trees		
M5	0.324	0.785
C&RT	0.578	0.766

dependent variables was found to be RBF-SVM, the sensitivity analysis as explained in Section 2.2.3 by Eq. (4) was conducted on the predictor variables to rank them in terms of their importance in predicting the *gtime* output variable. This first set consists of the predictor variables which are presented in Table 2.

The second set of predictor variables were selected by the authors through brainstorming sessions with medical professionals. The second set of candidate covariates are tabulated in Table 3.

Table 2
The 1st set of candidate covariates generated from RBF-SVM.

Variables	Explanation
Citizenship	Recipient citizenship @ registration
Contin_alcohol_old_don	Deceased donor-history of alcohol dependency + recent 6 months use
Contin_iv_drug_old_don	Deceased donor-history of iv drug use + recent 6 months use
Creat2_old	Most recent creatinine >2.0 mg/dl y/n
Da2	Donor a2 antigen
Dantiarr_old	Deceased donor given antiarrythmics 24 h prior to cross-clamp
Dayswait_chron	Active days on waiting list
Dobut_don_old	Deceased donor-dobutamine w/in 24 h pre-cross-clamp
Education	Recipient highest educational level @ registration
Ethcat_don	Donor ethnicity category
Fluvaccine	Anti-viral treatment—fluvaccine
Func_stat_tcr	Recipient functional status @ registration
Func_stat_trr	Recipient functional status @transplant
Gender	Recipient gender
Hbsab_don	Deceased donor hbsab test result
Hemo_pa_dia_tcr	Most recent hemodynamics pa (dia) mm/hg @ registration
Hemo_pa_mn_tcr	Most recent hemodynamics pa (mean) mm/hg @ registration
Heparin_don	Deceased donor management—heparin
Hgt_cm_tcr	Recipient height @ registration
Hist_alcohol_old_don	Deceased donor-history of alcohol dependency
Htlv2_old_don	Deceased donor-antibody to htlv ii result
Impl_defibril_after_list	Implantable defibrillator inserted between listing and transplant
Inotrop_agents	Deceased donor—three or more inotropic agents at time of incision
Inotrop_support_don	Deceased donor inotropic medication at procurement (y/n)
Ischtime	Ischemic time in hours
Med_cond_tcr	Recipient medical condition @ registration
Med_cond_trr	Recipient medical condition pretransplant @ transplant
Physical_capacity_tcr	Physical capacity at listing
Pretreat_med_don_old	Deceased donor medication(s) from brain death to 24 h prior to procurement
Prior_lung_surg_tcr	Recipient prior lung surgery (non-transplant) at listing
Pst_airway	Events prior to discharge: airway dehiscence
Pst_cardiac	Events prior to discharge: cardiac re-operation
Pst_dial	Events prior to discharge: dialysis
Pst_drug_trt_infect	Events prior to discharge: any drug treated infection
Pst_surgical	Events prior to discharge: other surgical procedures
Pt_t4_don	Deceased donor-thyroxine-t4 b/n brain death w/in 24 h of procurement
Sternotomy_tcr	Events occurring prior to listing: sternotomy
Sternotomy_trr	Events occurring between listing and transplant: sternotomy
Steroid	Chronic steroid use y/n/u @ transplant
Trtrej1y	Treated for rejection within 1 year
Trt_pulm_sepsis	IV treated pulmonary sepsis y/n/u @ registration
Vad_tah_tcr	Recipient on life support—ventilator @ registration (1 = yes, 0 = no)
Vad_tah_trr	Recipient on life support—ventilator @ transplant

The third set of candidate covariates was determined through the recent literature [52]. This set includes the variables commonly used in the previously published studies related to organ transplantation. The third set of candidate covariates are shown in Table 4.

The second and third set of candidate covariates can be perceived as the expert component of the method. If the predictive models in Step 3 do not reveal some very critical predictor variables (such as the age of the recipient in our case study), the method proposes to *force* the Cox model once more to review the significance of this kind of predictor variables.

3.3. Deployment of Cox regression model and devising the prognostic indices

All the candidate covariates as determined in Section 3.2 were assigned to Cox regression model at this step. The stepwise variable selection procedure was applied with 0.05 for entry and

Table 3
The 2nd set of candidate covariates.

Variables	Explanation
Antiarry	Heart medical factors: antiarrythmics at registration
Contin_Alcohol_Old_Don	Deceased donor-history of alcohol dependency + recent 6 months use
Contin_Cig_Don	Deceased donor-history of cigarettes in past and >20 pack years + recent 6 months use
Contin_IV_Drug_Old_Don	Deceased donor-history of iv drug use + recent 6 months use
Contin_Oth_Drug_Don	Deceased donor-history of other drugs in past + recent 6 months use
Eint	Ethnicity interaction between recipient and donor (in the same ethnic group, y/n)
Gint	Gender interaction between recipient and donor (having the same sex, y/n)
Hist_Alcohol_Old_Don	Deceased donor-history of alcohol dependency
Hist_Cancer_Don	Deceased donor-history of cancer (y/n)
Hist_Cig_Don	Deceased donor-history of cigarettes in past and >20pack yrs
Hist_Cocaine_Don	Deceased donor-history of cocaine use in past
Hist_Diabetes_Don	Deceased donor-history of diabetes, incl. Duration of disease
Hist_Hypertens_Don	Deceased donor-history of hypertension
LOS	Recipient length of stay post-transplant
Oth_Tobacco	Other tobacco use
Pack_Yrs	If history of cigarette use, number of pack years

Table 4
The 3rd set of candidate covariates.

Variables	Explanation
ABO	Recipient blood group at registration
ABO_Don	Donor blood type
ABO_Mat	Donor-recipient ABO match level
Age	Recipient age (years)
Age_Don	Donor age (years)
Dayswait_Chron	Active days on waiting list
Don_TY	Donor type—deceased/living
Ethcat	Recipient ethnicity category
Ethcat_Don	Donor ethnicity category
Gender	Recipient gender
Gender_Don	Donor gender
Hbsab_Don	Deceased donor hbsab test result
Ischtime	Ischemic time in hours
Med_Cond_Tcr	Recipient medical condition at registration
Med_Cond_Trr	Recipient medical condition pretransplant at transplant
Wgt_kg_Don	Donor weight (kg)
Wgt_kg_Tcr	Recipient weight (kg) at registration

Table 5
The variables kept in the Cox regression model.

Variable	SE	Chi_square test	DF	Significance	exp(β)	95% CI for exp(β)	
						Lower	Upper
LOS	0.0002	385.8701	1	<.0001	1.004	1.004	1.005
Eint	0.0178	56.9447	1	<.0001	0.844	0.844	0.905
Gint	0.0183	11.8644	1	0.0006	0.906	0.906	0.973
Age_Don	0.0006	247.3162	1	<.0001	1.009	1.009	1.011
Wgt_kg_Tcr	0.0004	5.5091	1	0.0189	0.998	0.998	1.000
Wgt_kg_Don	0.0005	21.3483	1	<.0001	0.997	0.997	0.999
Acyclovir	0.0300	14.6651	1	0.0001	0.840	0.840	0.945
Citizenship	0.0554	5.5538	1	0.0184	0.787	0.787	0.978
Dayswait_Chron	0.0002	7.5318	1	0.0061	1.000	1.000	1.000
Fluvaccine	0.0189	15.9915	1	<.0001	1.039	1.039	1.119
Ischtime	0.0058	239.5080	1	<.0001	1.081	1.081	1.105
Med_Cond_Tcr	0.0109	75.6231	1	<.0001	1.076	1.076	1.123
Vad_Tah_Trr	0.0002	5.7861	1	0.0162	1.000	1.000	1.001
Vad_Tah_Tcr	0.0077	48.9955	1	<.0001	1.040	1.040	1.072

0.1 for removal as significance threshold criteria. The predictor variables determined to be significant by Cox regression model are listed along with their corresponding statistics in Table 5. The rest of the variables (which were in Tables 2, 3, or 4 but not in Table 5) were eliminated since they were found to be insignificant by Cox regression model.

As listed in Table 5, 14 of the variables had prognostic value which are determined by the Cox model as significant and kept in the Cox equation. Therefore, they were used to calculate the PIs by means of Eq. (6). The PI values received here were ranging between 0 and 3.

3.4. Clustering the prognostic indices

Once the prognostic indices (PIs) for each recipient calculated, the next step was to cluster the recipients through these PIs. However, the problem of defining these clusters and deciding which value to cut off and categorize the recipients should be solved first. Two commonly used clustering algorithms as described in Section 2.5, namely *k*-means and TSCA were used to determine these clusters. We also compared these algorithm-based clusters to conventional PI devising methods in medicine. Two potential ways to do the clustering are constructing *equal-width* PIs and *equal-percentile* PIs in this research domain. In the former one, the PIs are separated in groups so that the increments of PI in each group are equal whereas the latter method focuses on allocating the patients equally to each group. The algorithms *k*-means and TSCA were run by changing the value for *k* (number of clusters to be formed). The value of *k* with 2, 3, 4, and 5 were tried because it was considered that having clusters more than 5 would not provide logical risk groups to categorize and would probably not be easy to name and interpret medically afterwards. The results for each run are represented in Table 6.

The performances of these entire four approaches with different number of clusters (*k* = 2–5) were compared using *intraclass inertia* as the performance measure to decide which one to adopt. It is a measure which shows how compact each cluster is. Intraclass inertia is the average of the distances between the means and the observations in each cluster. Eq. (7) indicates this value for given *k* number of clusters [53].

$$F(k) = \frac{1}{n}\sum_{k}\sum_{i\in C_k}\sum_{P=1}^{m}(X_{iP}-\mu_{kP})^2 \tag{7}$$

where n is the number of total observations, C_K is the set of kth cluster, X_{iP} is the value of the attribute P for observation i and μ_{kP} is the mean of the attribute P in the kth cluster. Note that in our case there is only one attribute which is PI, and hence $m = 1$.

The intraclass inertia values for each possible cluster are also summarized in Table 6. Prognostic indices were clustered best with *k = 3* with *k*-means clustering algorithm in our case as seen in Table 6 considering its low intraclass inertia value. As seen in Table 6, this classification not only gives the lowest intraclass inertia value but also provides an even distribution of the thoracic patients for our nation-wide dataset (38%, 16%, and 46% for low, medium, and high risk groups of patients, respectively). Although 5 clusters with *k*-means algorithm and 3 clusters in two-step cluster analysis perform very close to *k*-means algorithm with 3 clusters, neither of them provides such an even distribution of patients. Note that in addition to considerably higher inertia scores, heuristic calculation with equal-width PIs distribute the nation-wide patients highly skewed to lower tails of risk groups for all five potential cluster formations. Therefore, we conclude that the *k*-means algorithm based clustering performs better than the other potential groupings in terms of both objective and subjective aspects.

3.5. Validation of risk groups by Kaplan–Meier survival analysis

To validate the established prognostic indices with 3 clusters and hence the various risk groups in Section 3.4, Kaplan–Meier survival analysis [43] was conducted. The corresponding PI clusters were matched with the patients and their predictor variables from Table 5. In Kaplan–Meier survival analysis the predictor variables were used as explanatory variables and the PI-based clusters were used as the strata variable to label the patients with different risks. The main objective here was to compare survivor functions for different risk groups of thoracic recipients. If the survivor function for one risk group is always higher than the survivor function for another risk group, than the first group clearly lives longer than the second one. The less the survivor functions cross, the better the discrimination of the patients would be. Fig. 2 shows this clear distinction for *k*-means algorithm-based PIs.

In order to show that there is a statistically significant difference among these three risk groups, the test of equality over strata was also conducted. Test of equality over strata contains rank and likelihood-based statistics for testing homogeneity of survivor functions across strata. The rank tests with the log-rank test and Wilcoxon test indicate a significant difference between the risk groups. These results are also supported by likelihood-based statistics. These statistical test results are summarized in Table 7.

Table 6
The comparative results for clustering and heuristic-based algorithms.

		By clustering algorithms					
		k-Means algorithm			Two-step cluster analysis		
Number of clusters	Risk group	Prognostic index	Number of patients	Intraclass inertia	Prognostic index	Number of patients	Intraclass inertia
Cluster 1	Low	0–0.69	21163 (58%)	12.4×10^{-8}	0–1.09	34199 (94%)	866.30×10^{-8}
Cluster 2	High	0.70–3	15262 (41%)		1.1–3	2226 (6%)	
Cluster 1	Low	0–0.56	13766 (38%)	1.68×10^{-8}	0–1.04	33529 (92%)	2.20×10^{-8}
Cluster 2	Medium	0.57–0.91	5834 (16%)		1.05–1.83	2807 (7.7%)	
Cluster 3	High	0.92–3	16825 (46%)		1.84–3	89(0.3%)	
Cluster 1	Low	0–0.49	15227 (42%)	11.2×10^{-8}	0–0.41	6410 (17%)	445.39×10^{-8}
Cluster 2	Low–medium	0.50–0.77	1764 (5%)		0.42–0.70	15163 (42%)	
Cluster 3	Medium–high	0.78–1.12	9542 (26%)		0.71–1.04	11892(33%)	
Cluster 4	High	1.13–3	9892 (27%)		1.05–3	2960 (8%)	
Cluster 1	Very low	0–0.44	13266 (36%)	3.02×10^{-8}	0–0.36	2960 (8%)	720.71×10^{-8}
Cluster 2	Low	0.45–0.69	451(1%)		0.37–0.53	10475 (29%)	
Cluster 3	Medium	0.70–0.95	4449 (12%)		0.54–0.73	10815 (29%)	
Cluster 4	High	0.96–1.39	7814 (22%)		0.74–1.04	4674(13%)	
Cluster 5	Very high	1.40–3	10445 (29%)		1.05–3	7501 (21%)	
		By heuristics-based calculation					
		With equal PI widths			With equal percentiles		
Number of clusters	Risk group	Prognostic index	Number of patients	Intraclass inertia	Prognostic index	Number of patients	Intraclas s inertia
Cluster 1	Low	0–1.5	36154(99%)	713.68×10^{-8}	0–0.64	18212(50%)	7.02×10^{-6}
Cluster 2	High	1.6–3	271 (1%)		0.65–3	18213 (50%)	
Cluster 1	Low	0–0.9	32571 (89%)	1678.65×10^{-8}	0–0.53	12142 (33.5%;	2.01×10^{-6}
Cluster 2	Medium	1–1.9	3794 (10%)		0.54–0.76	12141 (33%)	
Cluster 3	High	2–3.0	60 (1%)		0.77–3	12142 (33.5%;	
Cluster 1	Low	0–0.7	26087 (72%)	12961.43×10^{-8}	0–0.47	9106 (25%)	2755.48×10^{-6}
Cluster 2	Low–medium	0.8–1.5	10153 (28%)		0.48–0.64	9106(25%)	
Cluster 3	Medium–high	1.6–2.3	162(0.4%)		0.65–0.82	9106 (25%)	
Cluster 4	High	2.4–3	23 (0.06%)		0.83–3	9107 (25%)	
Cluster 1	Very low	0–0.5	15605 (43%)	457.67×10^{-8}	0–0.43	7285 (20%)	3.16×10^{-6}
Cluster 2	Low	0.6–1.1	19608 (54%)		0.44–0.58	7285 (20%)	
Cluster 3	Medium	1.2–1.7	1109(3%)		0.59–0.71	7285 (20%)	
Cluster 4	High	1.8–2.3	80 (0.2%)		0.72–0.87	7285 (20%)	
Cluster 5	Very high	2.4–3	23 (0.06%)		0.88–3	7285 (20%)	

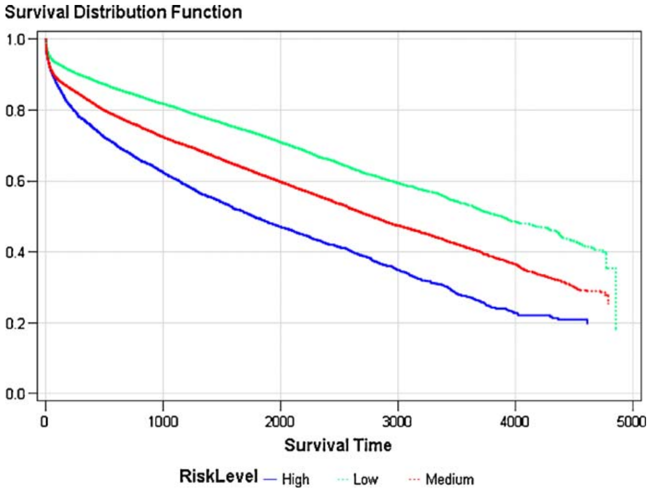

Fig. 2. Kaplan–Meier survival curves for three PIs.

Table 7
Tests of equality over risk groups for *k*-means based three PI cluster.

Test	Chi-square	DF	Pr > Chi-square
Log-rank	1002.6135	2	<.0001
Wilcoxon	939.7492	2	<.0001
−2 log(LR)	1013.3153	2	<.0001

4. Conclusions and future research directions

This study demonstrates that machine learning-based methodology for selecting predictor variables in survivability and prognostic modeling of thoracic organ transplantation is superior to the approaches adopting only expert-selected variables. The study showed that of the comprehensive list of predictors, some have been included in the previous studies (such as gender and age of the recipient, his/her medical condition at registration) while some others (which are found to be critical) have been absent from the related literature. These variables (e.g. such as recipient length of stay post-transplant and the interaction of gender and ethnicity between the recipient and the donor) should be combined with the factors identified in previous studies to better understand and improve the organ transplantation process.

The study revealed that based on *k*-means clustering algorithm the thoracic organ recipients should be allocated into an optimal number of "three" risk groups, namely low, medium, and high. This finding confirms the conventional medical discrimination commonly used in this field of study. However, it also proves that this grouping should be better performed through a data mining perspective rather than a heuristics-based approach because the latter one gives more skewed distribution of patients for our US nation-wide dataset. This is the point where the medical professionals should be advised to handle the problem in the future.

Some of the research extensions to the study reported in this article includes analysis of other organ types as well as the analysis of multiorgan scenarios where the correlations among the organs coming from the same donor are also included in the formulation of the problem. Another potential further research direction of this study is to validate the patterns obtained from the data mining models with a comprehensive simulation model of the organ transplantation process. Using actual cases, a comprehensive discrete-event simulation model can be developed and used as a test-bed where the potential benefits and limitations of these novel patterns are tested and validated for a sufficiently long period of time in the computer simulation environment.

References

[1] Trigt PV, Davis D, Shaeffer GS, Gaynor JW, Landolfo KP, Higginbotham MB, et al. Survival benefits of heart and lung transplantation. Annals of Surgery 1996;223:576–84.

[2] Pierson RN, Barr ML, McCullough KP, Egan T, Garrity E, Jessup M, et al. Thoracic organ transplantation. American Journal of Transplantation 2004;4:93–105.

[3] Sheppard D, McPhee D, Darke C, Shretha B, Moore R, Jurewitz A, et al. Predicting cytomegalovirus disease after renal transplantation: an artificial neural network approach. International Journal of Medical Informatics 1999;54: 55–76.

[4] Lin RS, Horn SD, Hurdle JF, Goldfarb-Rumyantzev S. Single and multiple time-point prediction models in kidney transplant outcomes. Journal of Biomedical Informatics 2008;41:944–52.

[5] Parmar MKB, Machin D. Survival analysis: a practical approach. Cambridge, UK: John Wiley & Sons; 1996.

[6] Cox DR. Analysis of survival data. London: Chapman&Hall; 1984.

[7] Hariharan S, Johnson CP, Bresnahan BA, Taranto SE, McIntosh MJ, Stablein D. Improved graft survival after renal transplantation in the United States, 1988 to 1996. The New England Journal of Medicine 2000;342:605–12.

[8] Herrero JI, Lucena JF, Quiroga J, Sangro B, Pardo F, Rotellar F, et al. Liver transplant recipients older than 60 years have lower survival and higher incidence of malignancy. American Journal of Transplantation 2003;3:1407–12.

[9] Hong Z, Wu J, Smart G, Kaita K, Wen SW, Paton S, et al. Survival analysis of liver transplant patients in Canada. Transplantation Proceedings 2006;38:2951–6.

[10] Kusiak A, Dixon B, Shah S. Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine 2005;35:311–27.

[11] Jenkins PC, Flanagan MF, Jenkins KJ, Sargent JD, Canter CE, Chinnock RE, et al. Survival analysis and risk factors for mortality in transplantation and staged surgery for hypoplastic left heart syndrome. Journal of the American College of Cardiology 2000;36:1178–85.

[12] Fernandez-Yanez J, Palomo J, Torrecilla EG, Pascual D, Garrido G, de Diego JJG, et al. Prognosis of heart transplant candidates stabilized on medical therapy. Revista Espanola de Cardiologia 2005;58:1162–70.

[13] Tjang YS, Heijdan GJMG, Tenderich G, Grobbee D, Korfer R. Survival analysis in heart transplantation: results from an analysis of 1290 Cases in a single center. European Journal of Cardio-Thoracic Surgery 2008;33:856–61.

[14] Lin HM, Kaufmann HM, McBride MA, Davies DB, Rosendale JD, Smith CM, et al. Center-specific graf and patient survival rates: 1997 UNOS report. JAMA 1998;280:1153–60.

[15] Cope JT, Kaza AK, Reade CC, Shockey KS, Kern JA, Tribble CG, et al. A cost comparison of heart transplantation versus alternative operations for cardiomyopathy. Annual thoracic Surgery 2001;72:1298–305.

[16] Aguero J, Almenar L, Martinez-Dolz L, Moro J, Izquierdo MT, Cano O, et al. Differences in clinical profile and survival after heart transplantation according to prior heart disease. Transplantation Proceedings 2007;39:2350–2.

[17] Christensen E, Gunson B, Neuberger J. Optimal timing of liver transplantation for patients with primary biliary cirrhosis: use of prognostic modeling. Journal of Hepatology 1999;30:285–92.

[18] Yoo HY, Galabova V, Edwin D, Thuluvath PJ. Socioeconomic status does not affect the outcome of liver transplantation. Liver Transplantation 2002;8:1133–7.

[19] Deng MC, DeMeester MJ, Smiths JMA, Heinecke J, Scheld HH. Effect of receiving a heart transplant: analysis of a national cohort entered on to waiting list, stratified by heart failure severity. British Medical Journal 2000;321:540–5.

[20] Ghobrial IM, Habermann TM, Maurer MJ, Geyer SM, Ristow KM, Larson TS, et al. Prognostic analysis for survival in adult solid organ transplant recipients with posy-transplantation lymphoproliferative disorders. Journal of Clinical Oncology 2005;23:7574–82.

[21] Harper AM, Taranto SE, Edwards EB, Daily OP. An update on a successful simulation project: the UNOS liver allocation model. In: Joines JA, Barton RR, Kang K, Fishwick PA, editors. Proceedings of the winter simulation conference. New York, NY: ACM Publications; 2000. p. 1955–62.

[22] Cupples SA, Ohler L. Transplantation nursing secrets. St. Louis, MO: Hanley & Belfus Publication; 2002.

[23] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other Kernel-based learning methods. London: Cambridge University Press; 2000.

[24] Mitchell T. Machine learning. New York, NY: McGraw-Hill; 1997.

[25] Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River, NJ: Prentice Hall; 1998.

[26] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics 2008;77:81–97.

[27] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 2002;35:352–9.

[28] Efron B, Tibshirani R. Statistical data analysis in the computer age. Science 1991;253:390–5.

[29] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.

[30] Kass GV. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 1980;29:119–27.

[31] Quinlan JR. Learning with continuous classes. In: Adams, Sterling, editors. Proceedings of 5th Australian joint conference on artificial intelligence. Singapore: World Scientific; 1992. p. 343–8.

[32] Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting: methods and applications. New York, NY: John Wiley and Sons; 1998.

[33] Everitt BS. Cambridge dictionary of statistics. Cambridge, UK: Cambridge University Press; 2002.

[34] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Boutilier C, editor. Proceedings of the 14th international conference on AI (IJCAI). San Mateo, CA: Morgan Kaufmann; 1995. p. 1137–45.

[35] Olson DL, Delen D. Advanced data mining techniques. New York, NY: Springer; 2008.

[36] Davis G. Sensitivity analysis in neural net solutions. IEEE Transactions on Systems Man and Cybernetics 1989;19:1078–82.

[37] Principe JC, Euliano NR, Lefebvre WC. Neural and adaptive systems. New York, NY: John Wiley and Sons; 2001.

[38] Saltelli A. Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications 2002;145:280–97.

[39] Saltelli A, Tarantola S, Campolongo F, Ratto M. Sensitivity analysis in practice—a guide to assessing scientific models. New York, NY: John Wiley and Sons; 2004.

[40] Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. Journal of Biomedical Informatics 2001;34:428–39.

[41] Grambsch P, Therneau T. Proportional hazards rates and diagnostics based on weighted residuals. Biometrika 1994;81:515–26.

[42] Christensen E. Multivariate survival analysis using Cox's regression model. Hepatology 1987;7:1346–58.

[43] Kaplan E, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:187–220.

[44] MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Proceedings of the fifth symposium on math, statistics, and probability. Berkeley, CA, USA: University of California Press; 1967. p. 281–97.

[45] Krishna K, Murty MN. Genetic *k*-means algorithm. IEEE Transactions on Systems Man and Cybernetics-Part B Cybernetics 1999;29:433–9.

[46] Chiu T, Fang D, Chen J, Wang Y, Jeris C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Lee D, editor. Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY: ACM Publications; 2001 p. 263.

[47] Li ZH, Luo P. Statistical analysis lectures of SPSS for windows. Beijing, China: Beijing Publishing House of Electronics Industry; 2004.

[48] SPSS Inc. PASW Modeler Data Mining Toolkit, Version 13.0, http://www.spss.com/software/modeling/modeler/ 2009 (accessed: June 5, 2009).

[49] SAS Institute Inc. Statistical Analysis Systems, Version 9.1.3, http://www.sas.com/technologies/analytics/statistics/stat/ 2008 (accessed: May 11, 2009).

[50] Hair JF, Anderson RE, Tatham RL, Black W. Multivariate data analysis. Upper Saddle River, NJ: Prentice Hall; 1998.

[51] Johnson DE. Applied multivariate methods for data analysts. Pacific Grove, CA: Duxbury Press; 1998.

[52] Oztekin A, Delen D, Kong ZJ. Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. International Journal of Medical Informatics 2009;78:84–96.

[53] Michaud P. Clustering techniques. Future Generation Computer Systems 1997;13:135–47.