


RESEARCH ARTICLE

Open Access



# Feature selection through validation and un-censoring of endovascular repair survival data for predicting the risk of re-intervention

Omneya Attallah<sup>1,2</sup>, Alan Karthikesalingam<sup>3</sup>, Peter J. E. Holt<sup>3</sup>, Matthew M. Thompson<sup>3</sup>, Rob Sayers<sup>4</sup>, Matthew J. Bown<sup>5</sup>, Eddie C. Choke<sup>6</sup> and Xianghong Ma<sup>1\*</sup> 

## Abstract

**Background:** Feature selection (FS) process is essential in the medical area as it reduces the effort and time needed for physicians to measure unnecessary features. Choosing useful variables is a difficult task with the presence of censoring which is the unique characteristic in survival analysis. Most survival FS methods depend on Cox's proportional hazard model; however, machine learning techniques (MLT) are preferred but not commonly used due to censoring. Techniques that have been proposed to adopt MLT to perform FS with survival data cannot be used with the high level of censoring. The researcher's previous publications proposed a technique to deal with the high level of censoring. It also used existing FS techniques to reduce dataset dimension. However, in this paper a new FS technique was proposed and combined with feature transformation and the proposed uncensoring approaches to select a reduced set of features and produce a stable predictive model.

**Methods:** In this paper, a FS technique based on artificial neural network (ANN) MLT is proposed to deal with highly censored Endovascular Aortic Repair (EVAR). Survival data EVAR datasets were collected during 2004 to 2010 from two vascular centers in order to produce a final stable model. They contain almost 91% of censored patients. The proposed approach used a wrapper FS method with ANN to select a reduced subset of features that predict the risk of EVAR re-intervention after 5 years to patients from two different centers located in the United Kingdom, to allow it to be potentially applied to cross-centers predictions. The proposed model is compared with the two popular FS techniques; Akaike and Bayesian information criteria (AIC, BIC) that are used with Cox's model.

**Results:** The final model outperforms other methods in distinguishing the high and low risk groups; as they both have concordance index and estimated AUC better than the Cox's model based on AIC, BIC, Lasso, and SCAD approaches. These models have  $p$ -values lower than 0.05, meaning that patients with different risk groups can be separated significantly and those who would need re-intervention can be correctly predicted.

**Conclusion:** The proposed approach will save time and effort made by physicians to collect unnecessary variables. The final reduced model was able to predict the long-term risk of aortic complications after EVAR. This predictive model can help clinicians decide patients' future observation plan.

**Keywords:** Survival analysis, Censoring, Feature selection, Model selection, Factor analysis, Cox's hazard proportional model, Endovascular aortic aneurysm repair

\* Correspondence: x.ma@aston.ac.uk

<sup>1</sup>School of Engineering and Applied Science, Aston University, B4 7ET, Birmingham, UK

Full list of author information is available at the end of the article



## Background

Endovascular aortic aneurysm repair (EVAR) is a surgical operation for patients suffering from aorta inflation known as aorta aneurysm. EVAR carries significantly lower preoperative risk than open repair surgery; therefore, it is preferred by patients and recommended by medical guidelines as the choice for treating abdominal aortic aneurysm AAA [1]. There is an obligatory need for life-long surveillance after this operation and it is considered to be expensive, varied, and poorly-calibrated [2]. However, the surveillance procedures are extensively various [3] and there is shortage of an indication to select the best timing or modality; Patients may be exposed to radiations and contrast nephropathy as a result of frequent surveillance. However, for some patients, complications required for treatment might be missed between surveillance [4–6]. For that reason, optimizing surveillance is very important [7, 8]. It is considered as an important issue in clinics and affects the long-standing cost-effectiveness of EVAR. By specifying which patients are more (high risk patients) or less likely (low risk patients) to require re-intervention within 5 years, a cost-effective and risk-stratified surveillance system could be achieved.

This study focuses on developing and validating a reduced predictive model for aortic complications after EVAR. In previous literature, models were usually built using only one dataset and validated with cross validation method [9–12]. However, cross-center testing is essential when the target is to validate the model for wider applications. Therefore in this work, two EVAR datasets were collected in two vascular centers in the United Kingdom which contain almost 91% of censored patients and which make the survival model construction and classification of difficult tasks. Censoring occurs when some patients could not be observed in the whole period of the survival study due to various reasons such as death, feeling better, or changing their residence location, leading to a type of missing data called censored data. The only information available for censored patients is the time until the last follow up or death (if it is not the event of interest). Therefore, the time until the event of interest is unknown [13]. The task here is to construct a predictive model using Center 1 EVAR data and testing it using other data collected from Center 2 with the presence of this high level of censoring. To deal with censoring, survival analysis techniques are adopted. It enables the use of the information available in the dataset even if it is censored. Actually, it does not omit censored patients [13]. Survival analysis techniques [14–16] were proposed with comparable datasets sizes to handle censoring.

Feature selection (FS) is very useful, especially in the medical area, as it reduces the time needed and the effort made by physicians to measure irrelevant and redundant features. It could avoid over-fitting that might occur

during the learning process of the predictive model. It may also lower the model's complexity and speed up the prediction process [17]. There are four primary approaches including filter, wrapper, embedded, and hybrid methods in FS [18]. These methods are used widely for standard data, though the task becomes more complex for survival data due the presence of censoring [17]. Therefore, the contribution of this paper is to select suitable feature selection methods for censored data, especially of a high level, which exist in the EVAR datasets. In FS using censored survival data, the most common approach used for modeling censored data is Cox's proportional hazard model. Cox's method uses the features information that might affect the hazard to build a predictive model [19]. Popular methods that perform FS using Cox's include; wrapper FS methods which wrapped FS around Cox's model and used several criteria such as Akaike or Bayesian information criteria, hazard ratio, and concordance index calculated from its prediction [20–23]. Some used Cox's model as a filter approach and perform a univariate analysis to calculate Cox's score metric for selection [24–28], while others used Wald test or likelihood test criteria instead [29] to quantify variables association with survival prediction. Several other metrics such as chi squared test [30], mutual information [31], and correlation information [32, 33] were also used to rank or filter variables. Other used penalized methods based on Cox's model such as; penalized L1 Least absolute shrinkage and selection operator (LASSO) [34, 35–38]. It was then extended to adaptive LASSO, weighted LASSO, and gradient LASSO [39, 40–42]. However, Lasso was used with another survival model called accelerated failure time model instead of Cox [43] for constructing a survival model. Other penalized methods which perform feature selection include; elastic-net [44, 45] and smoothly absolute clipped deviation (SCAD) [46] models.

Machine learning techniques [20] (MLT), such as Bayesian networks and Artificial Neural Networks (ANN), are usually favorable over the standard statistical models such as Cox due to their ability to identify complex relations between data that improve prediction [47]. However, they cannot be used directly on survival data due to the presence of censoring. Therefore, several scenarios have been proposed to handle censoring, but these methods were not applied to perform FS [48].

Some papers have discussed the use of MLT to perform FS in survival analysis data with the presence of censoring. Among them are the popular partial logistic artificial neural network (PLANN) and its extension PLANN with automatic relevance detection (PLANN-ARD). PLANN was used with backward feature elimination [49]. PLANN-ARD [11] method selects features based on their relevance to the model according to a Bayesian framework. These methods handle censoring by dividing observation time into  $n$

intervals and repeating patients to these intervals. Repeating leads to unbalanced and biased predictive models, especially with high censoring level [50]. It also increases the complexity and training time of these models and may increase the noise level existing in the datasets. Authors in [28] applied Cox's model to perform FS on censored data before entering a support vector machine (SVM) classifier. Others used wrapper FS methods [51, 52] with Bayes classifiers and K-nearest classifier. While in [30], a chi-square test was used to measure the degree of association between variables and observation time, then use ANN for prediction. Random survival forest classifier [53] is an extension to standard random forest classifier in which the output is survival time with a censor. It was used in [16, 54, 55] to perform FS. Gradient boosting with component-wise least squares was proposed by [56] based on cubic smoothing splines for L2 loss functions. It was then extended in [57] which illustrated that boosting technique works well in high-dimensional datasets for censored outcome data. Gradient boosting was used in [58] for survival analysis, however it differs from [57] in the choice of loss function, which in both cases is optimized via gradient boosting. The limitation of methods [11, 28, 30, 49–52] is the way they handled censoring which is carried out by deleting, ignoring, using only uncensored patients, or considering censored patients as event free. These methods of handling censoring are not appropriate to deal with a high level of censoring which exists in the suggested EVAR datasets. Moreover, the main drawback of tree based methods are instability, variable selection bias and over-fitting [59].

In the researcher's previously published paper [60], an uncensoring approach was proposed to deal with the high level of censoring in the EVAR datasets using MLT without performing FS. Moreover, in the researcher's other publication [61], the proposed uncensoring approach was combined with an existing ranking FS method to reduce the number of features in the datasets. Factor analysis is a feature transformation method used to transform data into a new domain so that most of the classification related information is compressed in a smaller number of features [62]. In this paper, it was used to group variables and to remove variables that are not related to any group, rather than to transform data into a new compressed domain used for classification. Moreover, a new FS technique was proposed to be combined with factor analysis and the researcher's uncensoring approach to select a reduced set of features and produce a stable predictive model. A stable model means that there may be a slight or no change in the prediction error when the data examples used for its training are replaced by other cases [63]. Prediction error is decomposed into two types; errors due to variance and or to bias. In order to produce a stable algorithm, a tradeoff between variance and bias must be made [64]. Therefore, the new proposed FS approach addresses the instability issue to

reduce bias and instability, which may be produced during FS and improves consistency of the feature selection using iterated nested cross validation method.

The proposed approach is divided into seven steps. The first step is feature reduction using factor analysis technique. The second is cross validation and permutation. The third is a stepwise feature selection which uses the  $p$ -value of the log rank test as a criterion. The fourth is the uncensoring step. The fifth is the iterated nested cross validation step followed by the sixth ANN model construction step. Steps two to six are repeated for each fold produced from the cross validation and permutation. They are repeated until a model with the minimum number of features is produced. The last step is the final model selection step done to choose among the different models constructed using each training fold generated in the cross validation step. The final model is the one which minimizes the  $p$ -value of the log rank test of the remaining censored test fold that was not used in training.

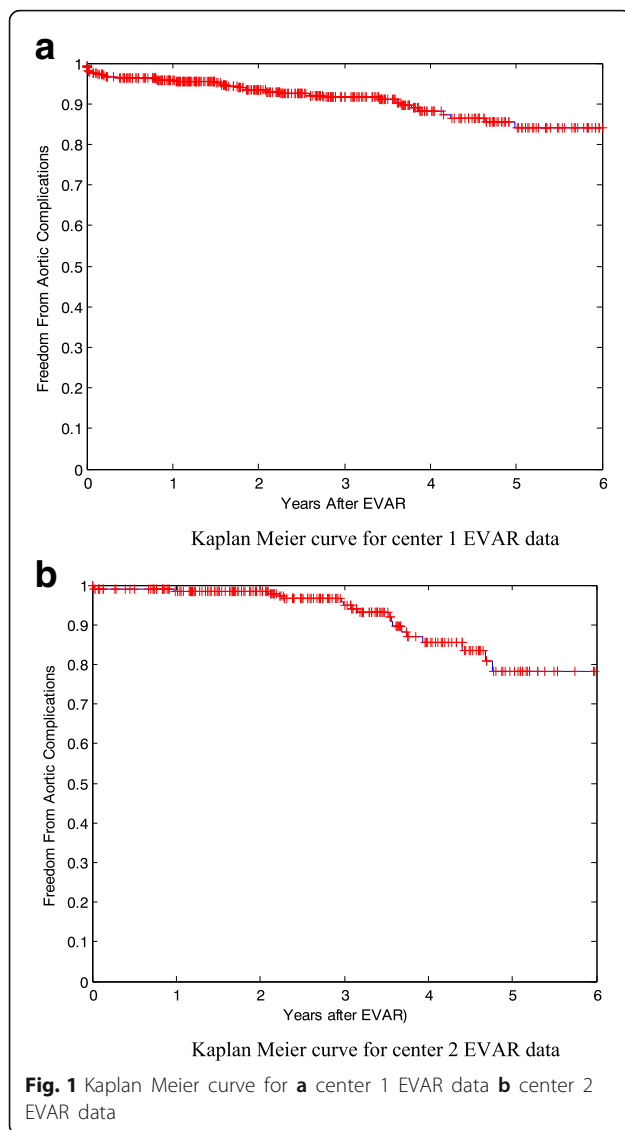
## Methods

### Data acquisition

From 2004 till 2010, follow up observations were collected to patients experiencing EVAR surgery from two different vascular centers located in the UK, at St George hospital (Center 1) and Leicester (Center 2). Both datasets consist of details of operative procedure and 47 patient morphological features. Three-dimensional computed tomography (CT) was employed to measure the pre-operational morphology features. The slice thickness of the CT images was 0.625 or 1.25 mm. They were obtained from the thoracic inlet to the level of the common femoral artery bifurcation. The total numbers of patients after removing missing values are 457 and 286 respectively, and the number of patients that experienced the EVAR re-intervention during the study period is 40 and 26 respectively for Centers 1 and 2. Details of these datasets can be found in a previous publication [65].

### Kaplan Meier curves

It is also known as product limit estimate of the survival function. It is one of the most well-known non-parametric survival analysis techniques. Non-parametric means that it does not take into account the information available by the predictive variables of a dataset when estimating survivability and does not assume survival distribution [66]. It calculates the probability of patient's survival at any time for the whole dataset even if it is censored. Clinical trials adopt survival analysis techniques for several purposes, such as predicting survivability after treatment of a disease, recurrent of cancer, or the risk of re-intervention after a surgical intervention [67]. Kaplan Meier curves of Centers 1 and 2 are plotted before using any feature selection method. They are shown in Fig. 1a and b.



## Feature transformation and reduction

### Factor analysis vs. principal component analysis

Principal Component Analysis (PCA) and Factor Analysis (FA) are the most well-known feature transformation and reduction techniques. PCA produces new features called components, which are the linear combinations of the original ones. These components are orthogonal to each other. They represent the variance in the data. Usually the first few components are the most important. However, FA seeks to determine the underlying structure among variables, which cannot be directly observed. It assumes that the attributes of data are produced by a linear combination of unobserved ones called latent factors. There are two types of factors, common and unique. The common factor is an unobserved variable that expresses the common variance between two or more observed variables (original measured

variables). The unique variance is an unobserved variable that explains the variance of only one observed variable. Features that are not correlated to any factor may be considered as unimportant and may be removed from the data. Therefore, FA is favorable than PCA as it aims to find the minimum number of factors that explain the underlying structure of the data [68]. Additionally, PCA does not take into account the effect of unique variance in determining principal components; it only regards the common variance. FA considers both the common and unique variances in calculating latent factors.

### Factor rotation

FA seeks to reduce the dimension of data in order to visualize and understand grouping and underlying structures between variables. Though, sometimes the primary latent factors are not able to clearly demonstrate this. Therefore, factor rotation may be needed to solve the problem. Rotation process is applied to the primary factors by rotating them into new axes called rotated factors [69]. These new axes produce large loading factors on one or two of the rotated factors, and small loading factors on others. The varimax and promax rotations are the most popular rotation techniques. A varimax rotation creates orthogonal uncorrelated rotated factors, while promax produces oblique ones [70]. In some cases, a promax rotation is favorable over varimax as it is a faster and simpler method [71]. However, in this paper varimax rotation was used to generate uncorrelated rotated latent factor in order that the variable related to different latent factors will be uncorrelated.

### Cox proportional hazard model

Cox-proportional hazard or Cox regression is the most popular semi-parametric method used in clinical trials. It builds a hazard model using covariates values by assuming that their effect on survival function is constant over time [72]. The relationship between patients' survival and covariates are investigated using Cox's model. In other words, it is used to estimate the risk of event given the prognostic variables at a time  $t$ . The model's output will be a hazard as a function of time and specific covariates.

### Bayesian networks

Bayesian network follows a particular structure of probabilistic graphical models known as directed acyclic graph (DAG). A DAG consists of a number of nodes (vertices)  $V = \{V_1, V_2, V_3, \dots, V_n\}$  representing variables and arcs  $A \subseteq V \times V$  connecting them and representing conditional and unconditional dependencies between these variables. Each node  $V$  represents a specific random variable and is drawn as a circle with its name on it. Arcs connecting the nodes are drawn as arrows and must be directed in only one direction which means that when an arc leaves a



node, it does not return to it again [73]. Each node  $V_i$  can be considered as a parent node when an arrow comes out of it. It can also be a child node when an arrow points towards it. BN illustrates relations and joint probabilities among variables of a dataset. This interpretability and its ability to provide reasoning with uncertainty make it appropriate to be used in the proposed approach.

### Artificial neural network

Artificial neural networks (ANNs) is one of the most popular and widely used ML techniques, especially in the medical area, as it has high ability to get good prediction even with noisy data [74] like the endovascular aortic repair (EVAR) datasets used in this paper. The artificial neural networks (ANN) are models stimulated by the biological nervous system of animals and specifically the neural networks of the brain [75]. Multilayer perception MLP-ANN is the commonly used ANN in the medical area. It consists of a number of neurons referred as nodes connected together by weights. These nodes are gathered together to form N layers consisting of one input, one output, and one or more hidden layers. The learning process in ANN is done by changing the weights of the connected neurons with the help of the training data and a learning algorithm.

### The proposed algorithm

The proposed technique used Center 1 (457 patients) for feature selection and Center 2 (286 patients) for its validation and assessment.

The proposed algorithm is divided into seven main steps. The flowchart in Fig. 2 describes the subsequent procedures. It also illustrates the main three areas of contributions of the proposed approach which are feature selection, uncensoring, and classification with their interactions which are highlighted in blue. Step 1 is reducing features using factor analysis. It applies factor analysis technique in order to lower the dimension of the data as an initial step. This was done after using Kaiser-Meyer-Olkin and Bartlett's tests to examine whether FA is needed for the data [76, 77]. Step 2 is the cross validation which splits Center 1 data into five outer training and testing folds and this is called the outer loop. It is used in the feature selection process for choosing the final reduced model, which produces the smallest  $p$ -value of the log rank test. Step 3 is the feature selection process which uses the  $p$ -value of the log rank test to select the variables used to build the ANN model from the uncensored training folds of Center 1 data. Step 4 is the uncensoring of the EVAR data. Step 5 is the iterated nested cross validation; which re-splits each outer training fold into five inner (nested) training and testing partitions. This is called the inner loop and it was used to overcome overoptimistic prediction that

may occur during feature selection, uncensoring and ANN construction steps. Moreover, this process produces a stable algorithm that tunes the tradeoff between variance and bias [78, 79]. Step 6 is the ANN model construction used for predicting the risk of re-intervention. Step 7 is to select the final model with the optimal number of features; the chosen model is the one which minimizes the  $p$ -value of the log rank test on the remaining testing fold which has not been used in the FS and ANN construction process. This  $p$ -value is used to determine whether the ANN model built with the selected features was capable of differentiating between the two risk groups of the censored Center 2 EVAR data. Usually, when it is lower than a significance level of 0.05, the two risk groups are considerably different, discriminative and separable. In the following subsection, each step will be discussed in detail.

### Feature reduction using factor analysis step

FA was employed in this work as a first step to reduce the dimension of the datasets by eliminating the variables that are not related to any latent factor. This will reduce the computational cost required later in the proposed feature selection algorithm. The number of latent factors has to be determined first in order to start FA. In order to determine the number of latent factors used in factor analysis, a scree plot is produced which shows the Eigen values accompanied with principle components or latent factor listed in descending order versus the number of components or factors as shown in Fig. 3, only 6 components showed almost 80% of the variance in the data. Therefore, the number 6 is adopted to do FA on Center 1 data.

After applying FA on the data, insignificant variables should be removed from the dataset. The most common criterion used to drop these variables is communality which is the part of the variance generated from the common factors. It is calculated by summing the squared factor loadings of all common factors for a given variable [80]. Factor loadings are the values or weights (called loadings) multiplied by each factor, then added together to form a linear combination of factors that produce observed variables. As a rule of thumb, when a variable has large loadings to some factors, this means that this variable can be represented by particular factors. Therefore, it should be retained as it is considered an influential observed variable [81]. For this reason, communality of high values is better. If the communality of a variable is high, this means that the factors describe a large percentage of the variable's variance. Hence, this indicates that this specific variable is well explained by the common factors, and therefore that the factor analysis is reliable. Low communality of a variable indicates that this variable does not load to any factor (have low loading factors), can be removed, and is considered a non-influential variable [81, 82]. Communality equals to (1- uniqueness).

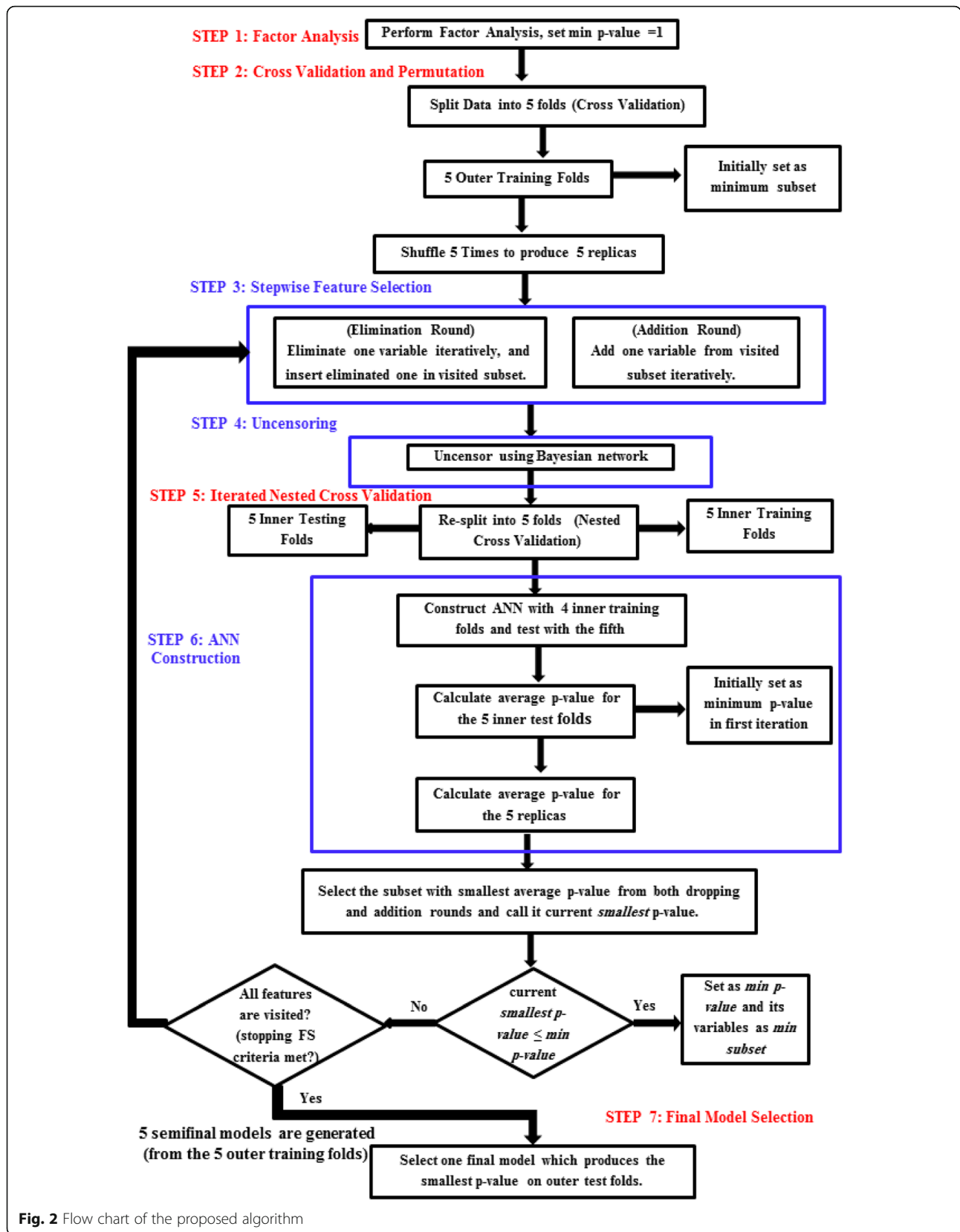
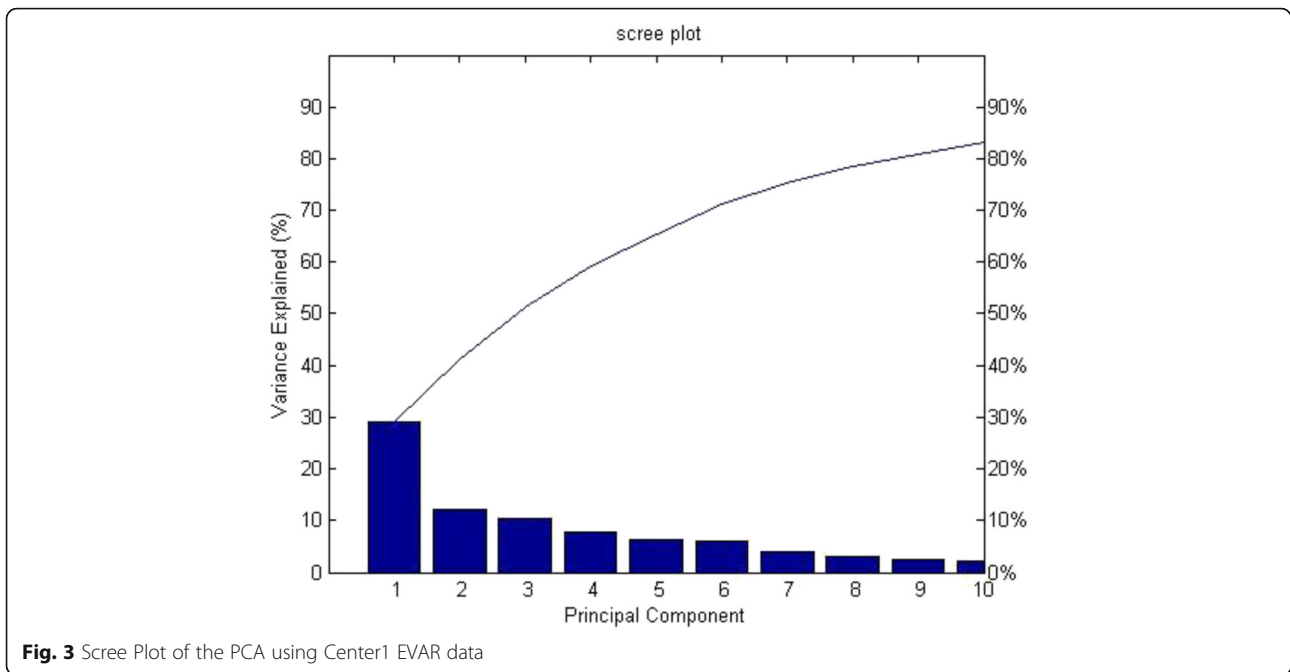


Fig. 2 Flow chart of the proposed algorithm



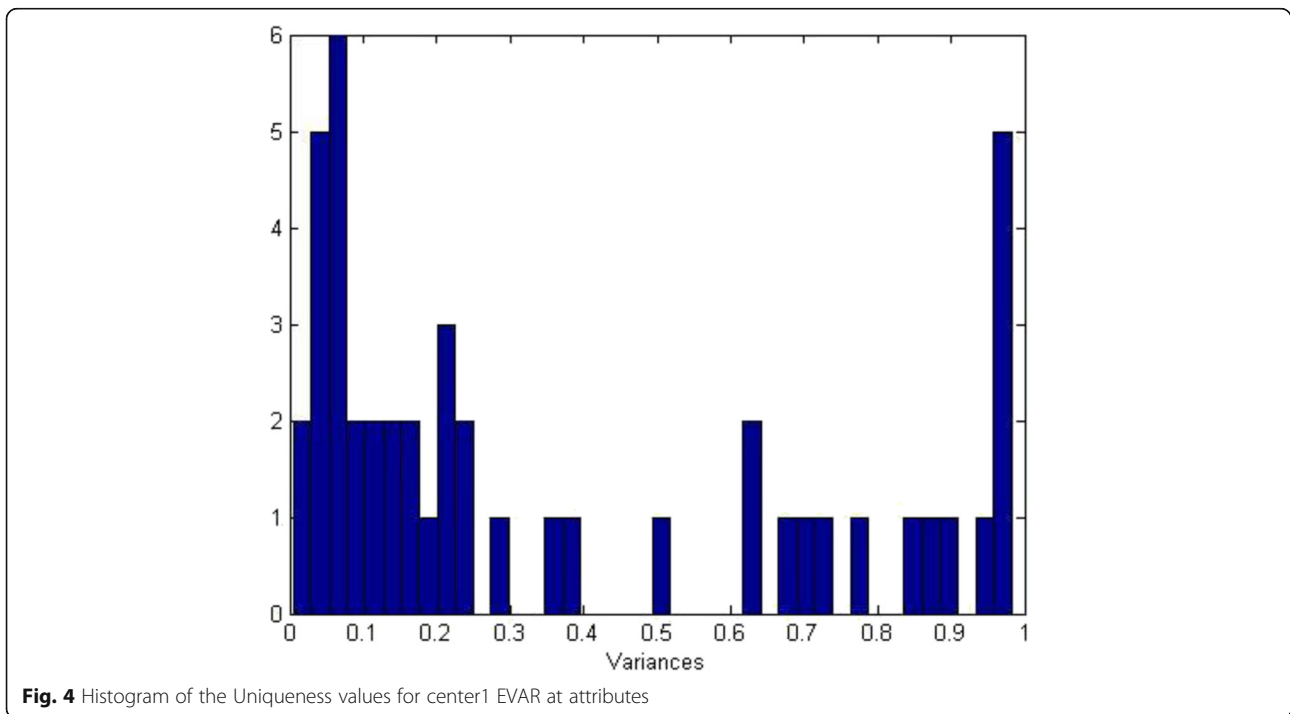
**Fig. 3** Scree Plot of the PCA using Center1 EVAR data

The uniqueness is a metric which measures the portion of the variance produced due to unique factors. In this work, the threshold of communality according to which features are dropped is determined from the histogram of unique variances (uniqueness), produced by unique factors. Variables with uniqueness values greater than this threshold will be removed, as they correspond to low communalities. As shown in Fig. 4, the first sudden drop in the uniqueness

occurs at 0.25. The researcher has chosen this threshold as other greater values correspond to high uniquenesses and low communalities.

**Cross validation and permutation step**

Center 1 EVAR data is divided into five folds using five folds cross validation method. Each separate four folds are called the outer training fold and the fifth is



**Fig. 4** Histogram of the Uniqueness values for center1 EVAR at attributes

called the outer testing fold. Each outer training fold is used in nested cross validation, stepwise feature selection, uncensoring, and ANN construction steps later. During the stepwise feature selection step, it is permuted five times to produce its replicas after being uncensored (in the uncensoring step).

**Stepwise feature selection step**

In the stepwise feature selection (SWFS) step, a canonical greedy stepwise search was used to select features using the  $p$ -value of the log-rank test as a performance metric. It is divided into two stages; feature elimination and addition. The main advantage of this strategy is that the eliminated features are given another chance to re-enter the feature selection process. It starts with the elimination stage in which all the features of each outer training fold is eliminated iteratively and the rest are used in the uncensoring and ANN construction steps after being permuted and re-split into inner nested folds to calculate the average of the  $p$ -value of the log-rank test. This prevents over-fitting and overoptimistic  $p$ -value predictions and enables the production of a stable model. The subset with the smallest averaged  $p$ -value is the one chosen and called “minimum subset” and its average  $p$ -value is called the “minimum averaged  $p$ -value”. Features eliminated will be inserted in a subset called the “visited subset” and will be given another chance to enter the FS process again in the addition stage. This is done by adding features of the visited subset iteratively, and then repeating the uncensoring and ANN construction steps. If the subset with the smallest average  $p$ -value has a  $p$ -value lower than the “minimum averaged  $p$ -value”, then this subset is set as the “minimum subset” and its  $p$ -value as the new “minimum averaged  $p$ -value”. This procedure is repeated until all features are visited.

**Uncensoring step**

The censoring time variable for each outer training loop was used to divide patients into three groups. The first one belongs to patients that experience the re-intervention at a time shorter than or equal to 5 years (re-intervened or high risk patients). The second group refers to patients that did not need the re-intervention for longer than 5 years (low chance of re-intervention or low risk patients). Finally, the rest of the patients are considered as the censored group which is the third group (those who died or left the follow up observations before 5 years).

Afterwards, each outer training fold was discretized using equal width unsupervised discretization technique. Then, each risk group was employed to build two Bayesian networks called low and high risk networks  $B^{low}$  and

$B^{high}$  respectively. Each Bayesian network is a directed acyclic graph (DAG) given a symbol  $\xi$ . They are constructed using the input variables  $V$ . Each variable  $V_i$  represents a node in this network and its parent is defined by  $\pi$ . The Bayesian networks were learned with Hill climbing structure learning algorithm [83]. The scoring function used for choosing the structure of the network has a minimum description [84]. Parameter learning was done using a maximum likelihood procedure to determine relation between nodes of a network [85].

The output class of each risk group is then removed as it is already known in the network. Next, each censored patient from the censored group is compared with the intrinsic distribution  $p$  of each network  $p^{high}$  and  $p^{low}$  correspondingly. Likelihood  $\ell(x_c/p)$  of the censored patients belongs to which network is used to uncensor the patient and determine to which group he/she belongs. It is calculated using Eqs. (1) and (2)

$$\hat{\ell}(x_c/p^{high}) = \ell(x_c/B^{high}) = p(x_c/\xi^{high}, p^{high}) = \prod_{i=1}^n p^{high}(V_i/\pi(V_i)). \tag{1}$$

$$\hat{\ell}(x_c/p^{low}) = \ell(x_c/B^{low}) = p(x_c/\xi^{low}, p^{low}) = \prod_{i=1}^n p^{low}(V_i/\pi(V_i)). \tag{2}$$

Afterwards, the posterior probability that outcome predictions that patients belong to each network given that they are censored ( $x_c$ )  $P(O/x_c)$  in Eq. (5) is calculated using Eqs. (3) and (4).

$$P(O^{high}/x_c) = \hat{P}(O^{high}) * \frac{\hat{\ell}(x_c/p^{high})}{P(x_c)}. \tag{3}$$

$$P(O^{low}/x_c) = \hat{P}(O^{low}) * \frac{\hat{\ell}(x_c/p^{low})}{P(x_c)}. \tag{4}$$

$$P(O/x_c) = \frac{P(O^{high}/x_c) + P(O^{low}/x_c)}{P(x_c)} = \frac{\hat{P}(O^{high}) * \hat{\ell}(x_c/p^{high}) + \hat{P}(O^{low}) * \hat{\ell}(x_c/p^{low})}{P(x_c)}, \tag{5}$$

Equation (5) is then normalized to ignore the effect of probability of a censored instance  $P(x_c)$  by dividing Eq. (5) by  $P(O/x_c)*P(x_c)$  to get Eq. (6).

$$P(O^{high}/x_c) + P(O^{low}/x_c) = 1. \tag{6}$$

Finally, a threshold named censoring correction  $P_{Th}$  is selected to determine whether each censored patient belongs to the high or low risk groups. If  $P(O^{high}/x_c)$  is bigger than  $P_{Th}$ , the censored patient will be considered as high risk and vice versa.



### Iterated nested cross validation step

After uncensoring all the replicas of each outer training partition, each replica is divided into five nested folds. Every different four folds are called inner training folds and are used to build an ANN (in the ANN construction step), while the fifth one is called inner test fold and is used for measuring the performance of the ANN using the  $p$ -value of the log-rank test of this inner test set. This is done for the different inner training and testing folds. Afterwards, the average of the  $p$ -value is then calculated. This process prevents overoptimistic results. It is worthy of note that each outer training fold was used to construct the ANN with different number of neurons. The process is repeated for every replica of the outer training fold, and then the average of all of them is calculated to produce the  $p$ -value of the nested cross validation which is used in the stepwise feature selection step as a criterion to select the attributes. It produces a stable model that tunes the tradeoff between variance and bias. Nested cross validation is also used as a method of avoiding over-fitting used due to the small sample size available in the data.

### ANN construction step

A three layer MLP-ANN trained with gradient descent with momentum back propagation algorithm was used to build a model for every outer training fold with different number of neurons. Sigmoid was used as an activation function. Other parameters are kept with their default values. Each model was used in the feature selection process to determine which one has the optimal features that minimize the  $p$ -value. The model has the ability to distinguish between the two risks groups of Center 1 data used in the ANN construction and can be validated with Center 2 data. The  $k$ -fold cross validation procedure was used to select the number of hidden neurons that reduces chances of over-fitting. In addition, it was employed and combined with feature selection and iterated nested cross validation to reduce chances of over-fitting.

### Final model selection step

Finally, five models will be produced from the five outer training folds that were used to build an ANN with different numbers of neurons. Each one is tested with the outer test fold which is not used in the ANN construction and the feature selection steps. The model that produces the minimum  $p$ -value in the outer test fold is the one chosen as a final model. The number of features and neurons produced in this model is used to train an ANN using Center 1 data. Center 2 data is used to validate the selected model.

## Results

### Results of the proposed feature selection method

One common approach for feature selection is to utilize the whole data set for the selection procedure. Resampling techniques such as cross validation and bootstrapping can split data into separate parts. Parts will be used for feature selection and evaluation. The other part that is not used in model construction will be used for validation and assessing the performance of the final feature selected model. In the proposed algorithm, Center 1 data were used for the feature and model selection (number of hidden neurons of ANN), while Center 2 data were employed for the validation and assessment. Cross and nested cross validation re-sampling techniques were used to split Center 1 data. The inner nested loop was used for producing an unbiased, stable algorithm and overcoming over-fitting and the overoptimistic predictions that might be produced. The outer loop was used to choose the model that produces the smallest  $p$ -value of the log-rank test on the outer test set. It is worth mentioning that the ANN of final chosen model has the number of neurons equal to seven.

Table 1 shows the results of using all the features available in Center 1 data and the results after FA reduction and the stepwise FS steps. It is clear from Table 1 that the number of features has been reduced from 45 to 27 after the FA step. The CI for Center 2 prediction has increased from 0.6 to 0.61. The  $p$ -value for Center 2 has improved from 0.036 to 0.034. It is also obvious from the table that the number of features selected in the final model after the stepwise selection step is lowered to 7. The  $p$ -value and CI for Center 2 predictions have improved to 0.022 and 0.63, respectively.

### Results of the final model of the proposed method compared to Cox's models based on AIC, BIC, lasso, and SCAD

In this section, the results of the final model produced by the proposed algorithm are compared with the most well-known, semi-parametric method used in clinical trials, Cox's Hazard proportional model. AIC, BIC, Lasso, and SCAD are popular feature and model selection techniques that are used with Cox's model to produce a stable algorithm with reduced number of features. It is well known that the Cox's output is continuous. In order to convert this output to binary outcome indicating the risk group,

**Table 1** Results of the proposed algorithm with the real dataset

The proposed algorithm	Number of features	$p$ -value (Logrank)	CI(standard error SE)	Sensitivity
All Features	45	0.036	0.6 (0.0677)	0.46
FA Features	27	0.034	0.61(0.0715)	0.57
Stepwise selection Features	7	0.022	0.63 (0.0739)	0.73

the parameter estimates from the final models were multiplied by each variable to generate a risk score and a threshold value is used to separate risk groups. A value above the threshold indicates high risk (class value of 1) and vice versa. The threshold that separated the two risk groups for Center 1 was 3.1 using AIC which is equivalent to the mean value of the risk score. The thresholds for BIC, Lasso, and SCAD are 2.4, 6.3, and 1.8, respectively. The same threshold is applied to Center 2 data. Tables 2 and 3 show the results of the proposed algorithm compared with the results of AIC, BIC, Lasso, and SCAD for Centers 1 and 2, respectively. The hyper parameter of Lasso and SCAD methods was optimized using ten-fold, cross validation.

It is obvious from Table 2 that the proposed algorithm, AIC, BIC, Lasso, and SCAD Cox's algorithm have all successfully distinguished the two risk groups of Center 1 as they all have  $p$ -value lower than 0.00001 except for SCAD which has a  $p$ -value of 0.0078. These  $p$ -values are beyond the significant level 0.05. The CI of AIC, BIC, Lasso, SCAD Cox's models are 0.79, 0.76, 0.7382, and 0.6271, which are greater than the proposed algorithm 0.74 except for the SCAD (0.6271). However, the proposed algorithm outperforms the AIC and Cox algorithm in the number of features selected in the final model which is 7, while in the other two methods it is 14. The proposed algorithm has the same number of selected features as Lasso [7], but greater than SCAD [3]. The advantage of using the proposed algorithm over the other methods in predicting the risk of re-intervention appears clearly in the predictions of Center 2 as shown in Table 3. First, it has the same CI 0.63 as the BIC model and outperforms the AIC, Lasso, and SCAD with CI equals to 0.61, 0.615, and 0.592. Moreover, it has an AUC (0.612) which is better than that of AIC, BIC, and SCAD (0.6, 0.605, and 0.58, respectively). However, the AUC of the proposed method is lower than that of Lasso 0.627. Finally, the  $p$ -values of the proposed method, AIC, BIC, Lasso, and SCAD algorithms are 0.022, 0.034, 0.029, 0.0068, and 0.2759. These  $p$ -values indicate that the results of all models can be separated significantly to distinguish between the high and low risk groups

**Table 2** Results of center1 data using the proposed algorithm compared with Cox's model using AIC, BIC, Lasso, and SCAD

Algorithm used	Number of features	$p$ -value (Logrank)	CI(SE)	Sensitivity
Proposed Algorithm	7	<0.00001	0.74 (0.0439)	0.76
AIC Algorithm	14	<0.00001	0.79 (0.0408)	0.69
BIC Algorithm	14	<0.00001	0.76 (0.0465)	0.38
Lasso Algorithm	7	<0.00001	0.7382 (0.0426)	0.714
SCAD Algorithm	3	0.0078	0.6271 (0.0518)	0.643

**Table 3** Results of center2 data using the proposed algorithm compared with Cox's model using AIC, BIC, Lasso, and SCAD

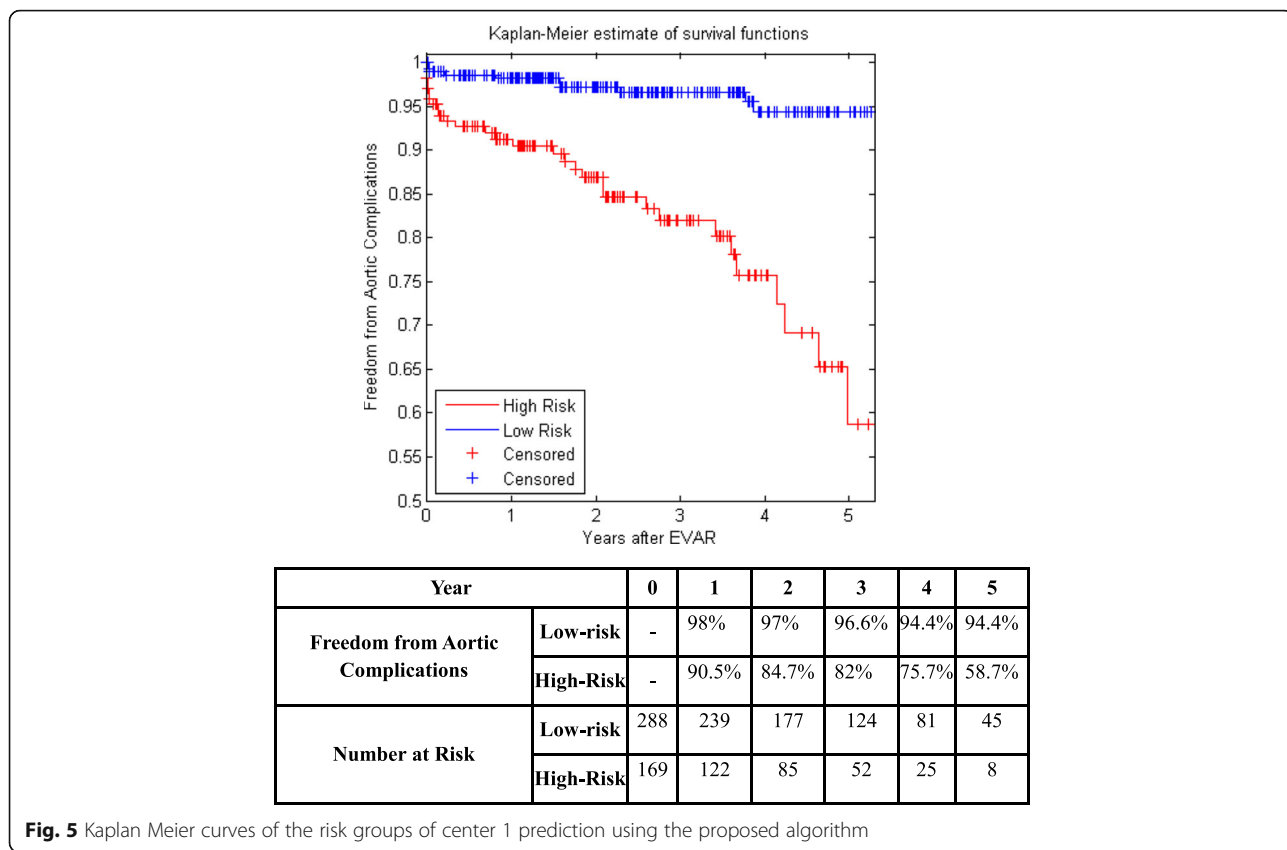
Algorithm used	Number of features	$p$ -value (Logrank)	CI(SE)	Uno's AUC	Sensitivity
Proposed Algorithm	7	0.022	0.63 (0.0739)	0.612	0.73
AIC Algorithm	14	0.034	0.61 (0.0725)	0.60	0.35
BIC Algorithm	14	0.029	0.63 (0.0685)	0.605	0.23
Lasso Algorithm	7	0.0068	0.615 (0.0864)	0.627	0.5
SCAD Algorithm	3	0.2759	0.592 (0.0658)	0.58	0.6923

except for SCAD which has  $p$ -value of 0.2759 which is greater than 0.05.

Figure 5 shows the KM curves of the two risk groups using the final selected model of the proposed algorithm for Center 1 compared with AIC (Fig. 6), BIC (Fig. 7), Lasso (Fig. 8), and SCAD (Fig. 9), respectively. Figure 10 shows the KM curves of the two risk groups using the final selected model of the proposed algorithm for Center 2 compared with AIC (Fig. 11), BIC (Fig. 12), Lasso (Fig. 13), and SCAD (Fig. 14), correspondingly. Moreover, these figures include the probability of freedom from aortic complications within the 5 years after EVAR along with the number of patients at risk of each group (low and high risks).

Figure 5 also shows that the proposed algorithm classified 169 of the Center 1 patients as high risk, which is equivalent to 37% of the patients. Freedom from aortic complications of the high risk patients reached 58.7% vs. 94.4% in low risk patients in the fifth year as shown in Fig. 5 ( $p < 0.00001$  log-rank test). However, the AIC, BIC, Lasso, and SCAD Cox's model predicted 104, 58, 196, and 208 patients as high risk as shown in Figs. 6, 7, 8 and 9. These are equivalent to 23, 13, 42, and 45.5% of the patients. Freedom from aortic complications of the high risk patients reached using AIC, BIC, and Lasso are 52% vs. 93%, 57.6% vs. 87.4%, and 91.7% vs. 73% in low risk patients in the fifth year ( $p < 0.00001$  log-rank test). For, SCAD, the high risk is 79.2% vs. 97.5% of the low risk in the fifth year ( $p = 0.0078$  log-rank test).

Figure 10 shows that the proposed algorithm classified 136 of the Center 2 patients as high risk, which is equivalent to 47.5% of the patients. Freedom from aortic complications of the high risk patients reached 69% vs. 92.6% in low risk patients in the fifth year ( $p < 0.022$  log-rank test). However, the AIC, and BIC Cox's model predicted 25 and 41 patients as high risk as shown in Figs. 11 and 12, which is equivalent to 8 and 14.3% of the patients. Freedom from aortic complications of the high risk patients reached 39.5% vs. 83% and 53.7% vs. 84.6% in low risk patients in the fifth year ( $p = 0.034$



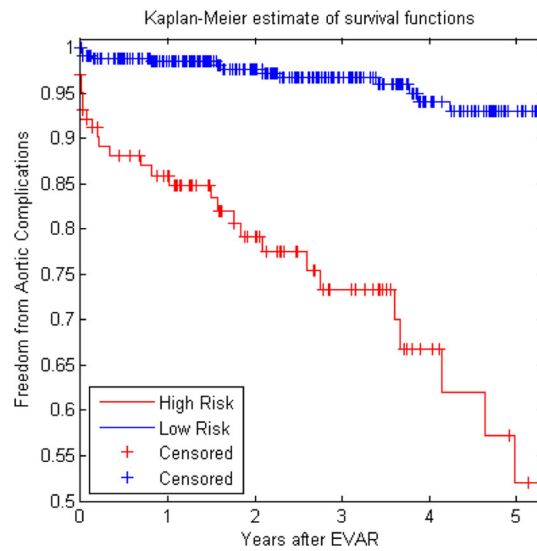
and 0.029 log-rank test). Lasso and SCAD algorithms predicted 76 and 121 patients as high risk as shown in Figs. 13 and 14, which is equivalent to 26.5 and 42.3% of the patients. Freedom from aortic complications of the high risk patients reached 60.7% vs. 87.5% for low risk in Lasso and 72.5% vs. 89.2% in low risk patients in SCAD in the fifth year ( $p = 0.034$  and 0.029 log-rank test).

**Discussion**

The influence of aortic morphology on long term prediction of EVAR is complicated and suitable to be analyzed with ANN, with significant possible interface between aortic volume, shape, diameter, and angulation. Current approaches have shown that aneurysm diameter predicts re-intervention after EVAR [86, 87], however evidence also recommended that other features of aneurysm morphology influence long-term clinical success [88–90]. Further complex concerns such as endograft configuration and deployment, or intermediate markers of patients’ cardiovascular risk phenotype, could possibly be used to train ANN in prospective studies, which increases the clinical significance of prediction and makes it more reliable. Also, adding more operative factors such as graft size, and endoleak at completion or post-operative variables, such as endoleak at early surveillance scans, could significantly

enhance the discriminatory power of ANNs [61]. As an evidence, the proposed FS approach using ANN has selected maximum aneurysm neck diameter, diameter of the left common iliac artery 1 and 5 mm below internal iliac ostium, maximum common iliac artery diameter 5 mm proximal to internal iliac origin, maximum iliac tortuosity index, maximum common iliac thrombus volume, and right common iliac artery non luminal volume. These features were examined by clinicians who approved that they have outstanding validation terms for the prediction morphology for current endografts available. Results were compared with current clinical method such as; SGVI (St George Vascular Institute), which showed that the proposed method has superior performance. The seven features ANN of the proposed method have greater predictive ability in classifying low and high risk patients than the SGVI score [91]. Concordance index, estimated AUC, and  $p$ -value of the predictive model using the selected features show good clinical sense. In addition, they potentially indicate an increase in the event detection (EVAR re-intervention in this study) and risk group separation without affecting the cost of collecting unnecessary additional variables and surveillance cost.

Details about the uncensoring algorithm, could be found in [60]. Both Bayesian and neural network were



Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.5%	97.7%	96.7%	94%	93%
	High-Risk	-	86%	79%	73.4%	66.73%	52%
Number at Risk	Low-risk	353	284	211	144	90	42
	High-Risk	104	77	51	32	16	9

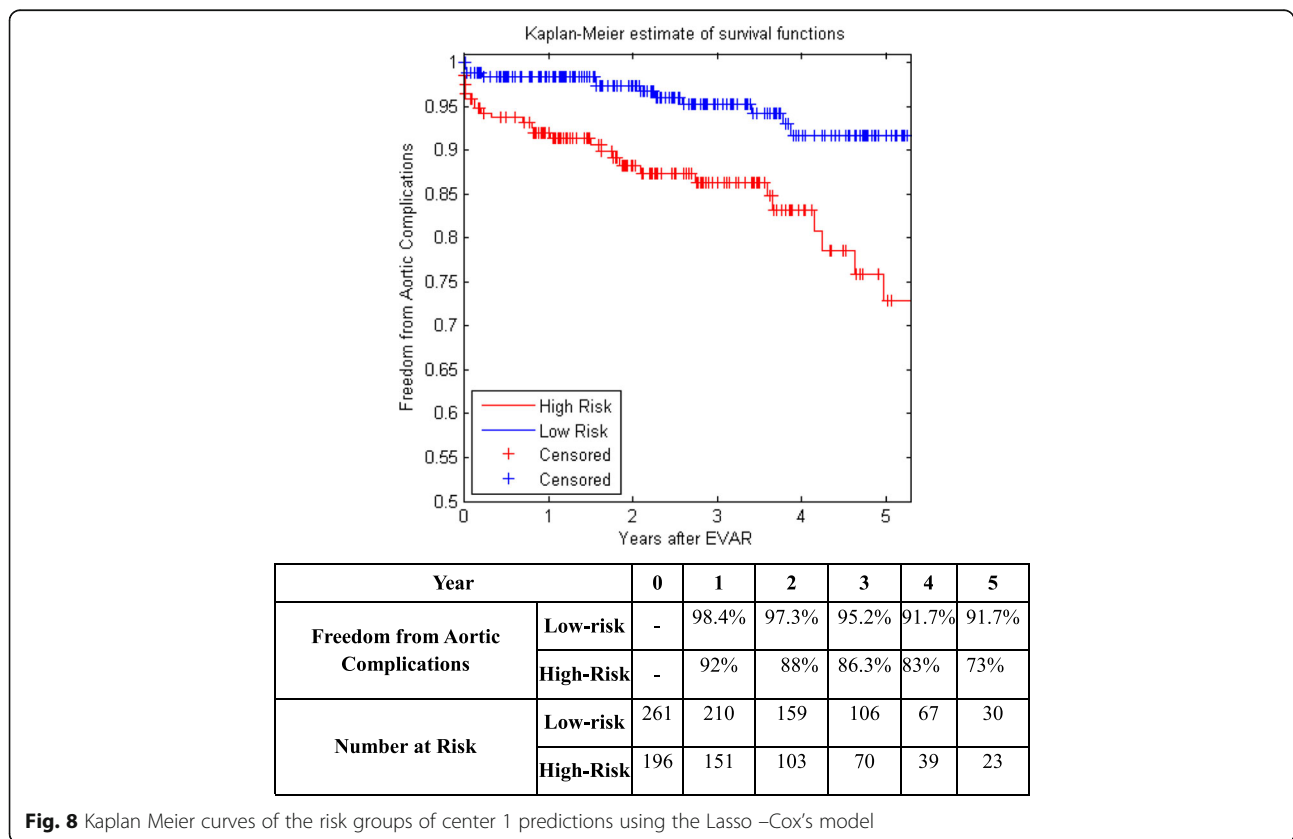
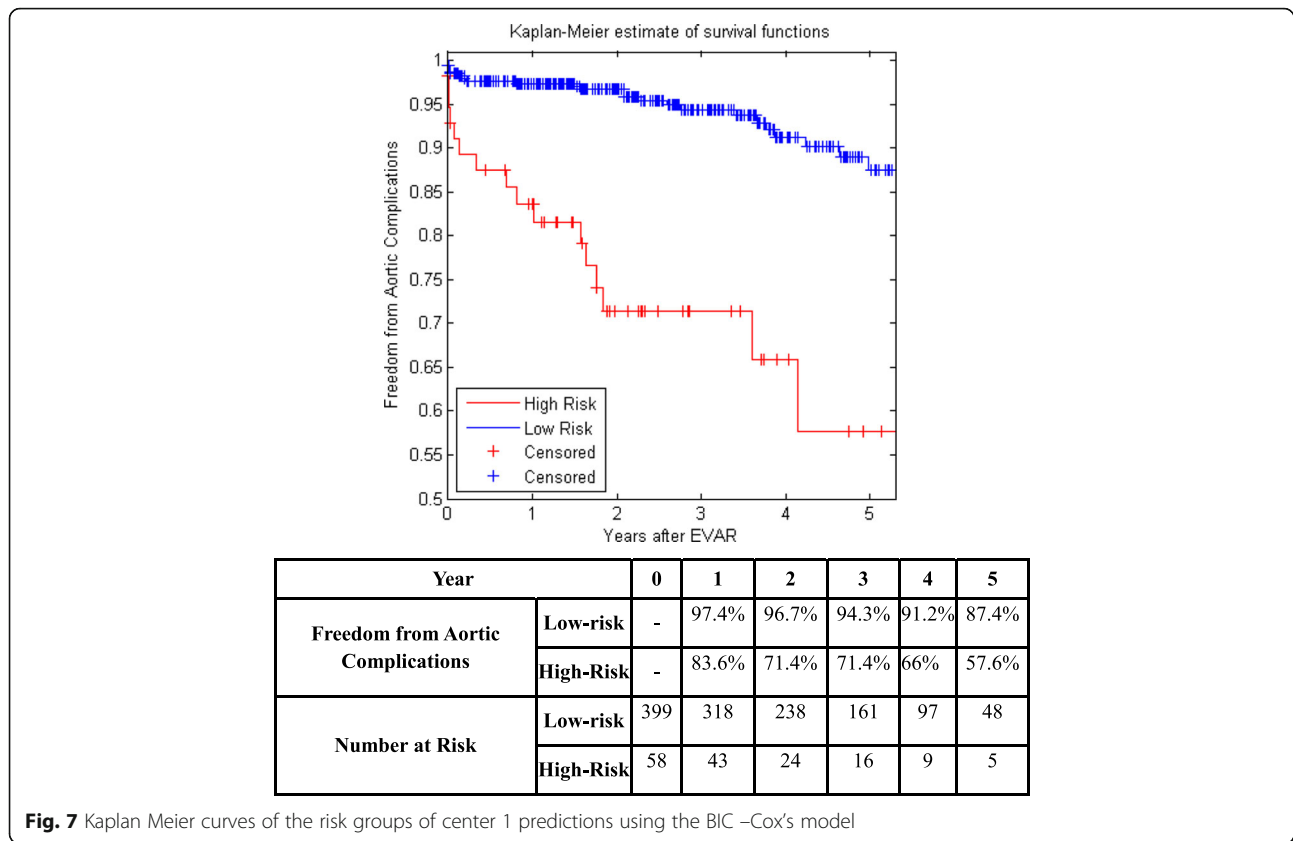
**Fig. 6** Kaplan Meier curves of the risk groups of center 1 predictions using the AIC –Cox's model

constructed using Weka software [92]. Bayesian network is used in this paper as it helps physicians and clinicians to understand the relations between variables. It is a graphical probabilistic network and it determines the joint probabilities between the variables of a dataset. Each variable is given as a node in the network. The final graph makes it easier for doctors to figure out the relations between the variables and how predictions are performed, as they are less aware of data mining and machine learning techniques [31]. The structures of the Bayesian networks generated after feature selection can be found in Additional file 1.

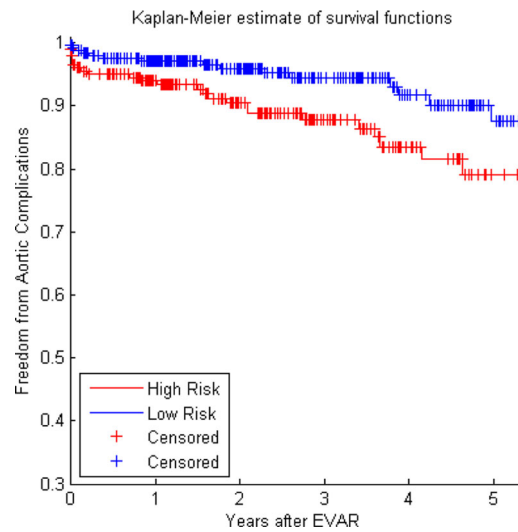
This work has several contributions; the first one is to use machine learning techniques such as Bayesian networks and Artificial Neural networks to solve censoring issue and build a predictive model. As mentioned before, previous techniques that used machine learning techniques dealt with low to medium level censoring. Moreover, the way they handled censoring may lead to a biased predictive model, especially with high censoring. The second one is to deal with the high censoring issue which is a major difficulty in datasets that prevent building a predictive model capable of predicting the risk of EVAR re-intervention. Since most of the previous studies used only one dataset split using k-

fold cross validation or bootstrapping to validate the results. The third contribution is to justify the model in future cross-center applications by constructing a predictive model using Center 1 EVAR data and testing them using other data collected from Center 2. Earlier work was mainly concerned with dealing with the censoring issue without performing FS. Existing FS methods dealing with censored data have several drawbacks and cannot be used with the highly censored EVAR datasets as mentioned before in the introduction. The fourth one is to apply the appropriate feature selection techniques to the highly censored endovascular repair survival data.

The proposed algorithm used Bayesian network to uncensor the highly censored EVAR datasets and ANN for prediction of risks. It employed FA reduction and stepwise FS to reduce the number of features used for building the predictive model. It used cross validation for model selection (choosing the number of neurons of ANN) and iterated nested cross validation to generate a stable feature selection algorithm and produce the final reduced model. A simulation study was performed and included in the Additional file 2 to show the effectiveness of the proposed algorithm compared with other variable selection methods.

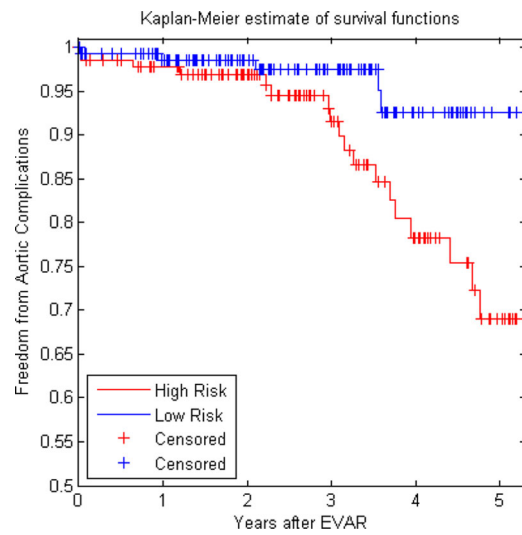






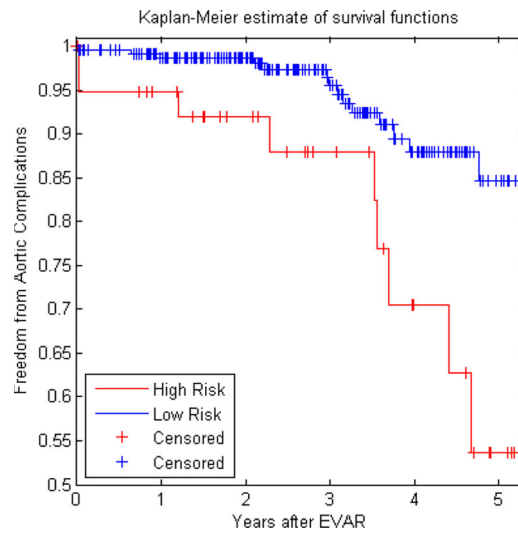
Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	97%	96%	94.4%	91.6%	97.5%
	High-Risk	-	93%	90.5%	87.7%	83.5%	79.2%
Number at Risk	Low-risk	249	205	150	100	60	27
	High-Risk	208	156	112	74	46	25

Fig. 9 Kaplan Meier curves of the risk groups of center 1 predictions using the SCAD –Cox’s model



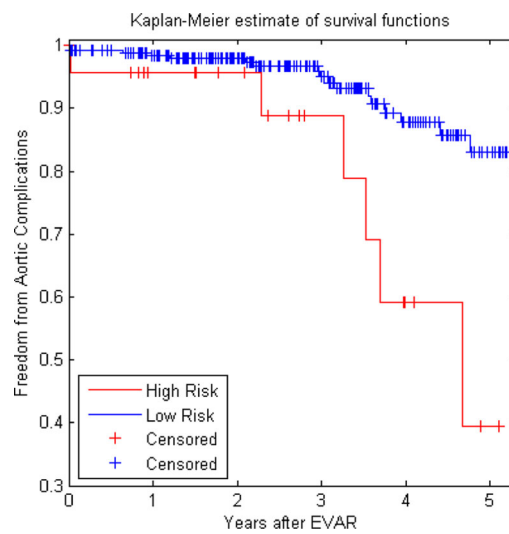
Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.5%	98.5%	97.5%	92.6%	92.6%
	High-Risk	-	97.7%	97%	91.5%	78.3%	69%
Number at Risk	Low-risk	150	128	93	60	27	5
	High-Risk	136	120	91	60	35	10

Fig. 10 Kaplan Meier curves of the risk groups of center 2 predictions using the proposed algorithm



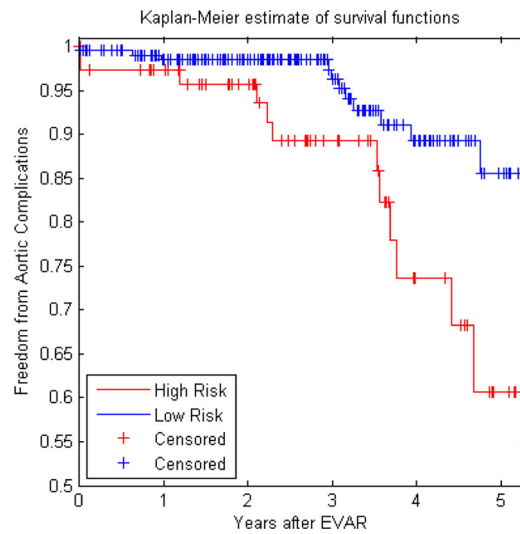
Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.7%	98.7%	95.5%	88%	84.6%
	High-Risk	-	94.8%	92%	88%	70.5%	53.7%
Number at Risk	Low-risk	245	214	159	102	53	15
	High-Risk	41	34	25	18	10	1

**Fig. 11** Kaplan Meier curves of the risk groups of center 2 predictions using the AIC –Cox’s model



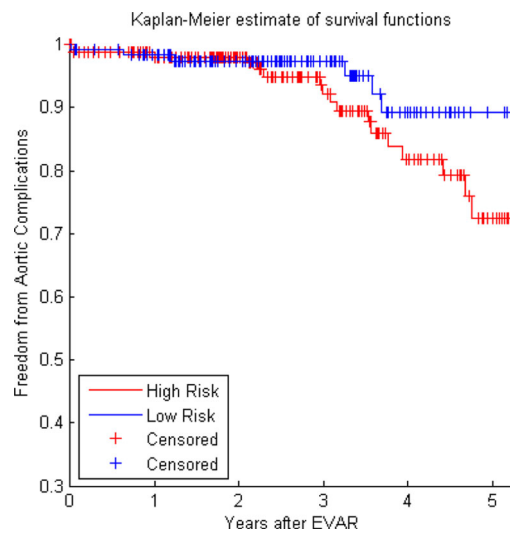
Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.4%	98%	95%	87.8%	83%
	High-Risk	-	95.6%	95.6%	88.5%	89.2%	39.5%
Number at Risk	Low-risk	261	230	169	111	58	15
	High-Risk	25	19	15	9	4	1

**Fig. 12** Kaplan Meier curves of the risk groups of center 2 predictions using the BIC –Cox’s model



Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.5%	98.5%	96.3%	89.2%	85.5%
	High-Risk	-	97.3%	95.7%	89.2%	73.5%	60.7%
Number at Risk	Low-risk	210	182	134	89	47	15
	High-Risk	76	66	50	31	15	3

**Fig. 13** Kaplan Meier curves of the risk groups of center 2 predictions using the Lasso –Cox’s model



Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.3%	97.4%	97.4%	89.2%	89.2%
	High-Risk	-	98%	98%	92.2%	81.8%	72.5%
Number at Risk	Low-risk	165	139	103	68	38	9
	High-Risk	121	109	81	52	24	6

**Fig. 14** Kaplan Meier curves of the risk groups of center 2 predictions using the SCAD –Cox’s model

## Conclusion

A new feature selection technique was proposed to build and validate a predictive reduced model using the two EVAR datasets (743 patients) collected during 2004 to 2010 from two vascular centers. The final reduced model was able to predict the long-term risk of aortic complications after EVAR. Only morphology features were used for constructing the model as they have greater effect on aortic complications than physiology ones [6, 93, 94]. The proposed feature selection technique has successfully reduced the final model to 7 features only instead of the full model of 45 attributes. The final reduced model was validated using Center 2 data and the results showed that it was capable of predicting the risk of EVAR re-intervention. It was also able to successfully distinguish between the two risk groups of each center as the  $p$ -value of the log rank test was lower than 0.00001 for Center 1 and 0.022 for Center 2. This proves that the model can be used in cross-center predictions. This will help clinicians to put a future different follow up surveillance plan for different risk groups of EVAR patients.

Four other popular feature selection techniques AIC, BIC, Lasso, and SCAD were compared to the proposed algorithm. The reduced predictive model constructed using the proposed approach has higher ability in discriminating and distinguishing between risk groups of patients than other variable selection methods based on Cox's model, since it has better concordance index, and estimated AUC. In addition, the number of patients of Center 1 that were classified as high risk using the proposed method, AIC, BIC, Lasso, and SCAD Cox's models are 169, 104, 58, 196, and 208 patients versus 136, 25, 41, 76, and 121 for Center 2. This means that the proposed algorithm better identifies the risk of EVAR re-intervention. Therefore, it may be preferred by doctors to decide which surveillance plan each patient should undertake. Clinicians will put a more regular monitoring schedule in the future follow up and surveillance plan for those who have high risk of needing re-intervention, and the lower risk patients can be monitored less regularly. This would help in balancing and developing a cost-effectiveness surveillance system.

## Additional files

**Additional file 1:** The structures of the high and low bayesian networks after feature selection. (PDF 208 kb)

**Additional file 2:** Comparative variable selection survival models and the generated Simulation Study. (PDF 180 kb)

## Abbreviations

AAA: Abdominal aortic aneurysm; AIC: Akaike information criteria; ANN: Artificial neural network; BIC: Bayesian information criteria; BN: Bayesian network; BP: Back propagation; CI: Concordance index; CT: Computed

tomography; DAG: Direct acyclic graph; EVAR: Endovascular aortic repair; FA: Factor analysis; FR: Feature ranking; FS: Feature selection; KM: Kaplan Meier; Lasso: Least absolute shrinkage and selection operator; MLP: Multilayer perceptions; MLT: Machine learning techniques; MPL: Maximum partial likelihood; PCA: Principal component analysis; PH: Proportional hazard; PLANN: Partial logistic regression artificial neural network; REINT: Risk of endovascular aortic repair intervention; SCAD: Smoothly clipped absolute deviation; SWFMS: Stepwise feature model selection

## Acknowledgments

We would like to acknowledge Prof. David Lowe at Aston University for his guidance and useful discussions on this research.

## Funding

The authors have no support or funding to declare.

## Availability of data and materials

The data cannot be made available outside of the English National Health Service due to existing data agreements covering patients treated in the authors' healthcare system and local/national data governance policies. Data are available upon request from the author Alan Karthikesalingam.

## Authors' contributions

All authors made substantial contributions to the algorithm design and methods. OA implemented the methods, analyzed and interpreted the data, prepared the draft of the paper. XM revised the manuscript critically for important intellectual content and approved the final version to be submitted. AK, PH, and MMT collected Center 1 dataset. RS, MJB, EC collected Center 2 dataset. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

The study (including anonymised imaging data from Leicester and St George's hospital patients) was classified as service evaluation by communication with Wandsworth Research Ethics Committee REC in 2012, when the data analysis plan was conceived. Wandsworth REC waived the need for approval as this was a retrospective observational study of anonymised data.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>School of Engineering and Applied Science, Aston University, B4 7ET, Birmingham, UK. <sup>2</sup>Department of Electronics and Communications, College of Engineering and Technology, Arab Academy for Science and Technology, Alexandria, Egypt. <sup>3</sup>St George's Vascular Institute, London SW17 0QT, UK. <sup>4</sup>St George's Vascular Institute, St George's University Hospitals NHS Foundation Trust, Blackshaw Road, London SW17 0QT, UK. <sup>5</sup>Vascular Surgery Group, University of Leicester, Leicester, UK. <sup>6</sup>Vascular Surgery Group, Robert Kilpatrick Clinical Sciences Building, Leicester Royal Infirmary, University of Leicester, Leicester LE2 7LX, UK.

Received: 5 July 2016 Accepted: 24 July 2017

Published online: 03 August 2017

## References

- Moll FL, et al. Management of abdominal aortic aneurysms clinical practice guidelines of the European society for vascular surgery. *Eur J Vasc Endovasc Surg.* 2011;41:S1–S58.
- Hay N, McCracken F, Richardson J, George E, Barnett D. Endovascular stent-grafts for the treatment of abdominal aortic aneurysms: NICE technology appraisal guidance. *Heart.* 2009;95:1798–800. doi:10.1136/hrt.2009.176362.

3. Karthikesalingam A, et al. Heterogeneity in surveillance after endovascular aneurysm repair in the UK. *Eur J Vasc Endovasc Surg.* 2011;42:585–90. doi:10.1016/j.ejvs.2011.06.053.
4. Weerakkody RA, et al. Radiation exposure during endovascular aneurysm repair. *Br J Surg.* 2008;95:699–702. doi:10.1002/bjs.6229.
5. Walsh SR, Tang TY, Boyle JR. Renal consequences of endovascular abdominal aortic aneurysm repair. *J Endovasc Ther.* 2008;15:73–82. doi:10.1583/07-2299.1.
6. Karthikesalingam A, et al. Risk of reintervention after endovascular aortic aneurysm repair. *Br J Surg.* 2010;97:657–63. doi:10.1002/bjs.6991.
7. Brown LC, et al. The UK Endovascular aneurysm repair (EVAR) trials: design, methodology and progress. *Eur J Vasc Endovasc Surg.* 2004;27:372–81. doi:10.3310/hta16090.
8. Chambers D, Epstein D, Walker S, Fayter D, Paton F, Wright K, Michaels J, Thomas S, Sculpher M, Woolcott N. Endovascular stents for abdominal aortic aneurysms: a systematic review and economic model. *Health Technol Assess.* 2009;13:215–318. doi:10.3310/hta13480.
9. Brown SF, Branford AJ, Moran W. On the use of artificial neural networks for the analysis of survival data. *IEEE Trans Neural Netw.* 1997;8:1071–7. doi:10.1109/72.623209.
10. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med.* 2005;34:113–27.
11. Lisboa PJ, Wong H, Harris P, Swindell R. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artif Intell Med.* 2003;28:1–25.
12. Kalderstam J, et al. Training artificial neural networks directly on the concordance index for censored data using genetic algorithms. *Artif Intell Med.* 2013;58:125–32.
13. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: basics and must know areas. *Perspect Clin Res.* 2011;2:145–8. doi:10.4103/2229-3485.86872.
14. Neuberger J. Predicting the prognosis of primary biliary cirrhosis. *Gut.* 1989;30:1519–22.
15. Zhu M, Fan G. Variable selection by ensembles for the Cox model. *J Stat Comput Simul.* 2011;81:1983–92.
16. Pang H, George SL, Hui K, Tong T. Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Trans Comput Biol Bioinform.* 2012;9:1422–31.
17. Liu H, Motoda H. Feature selection for knowledge discovery and data mining, vol. 454: Springer Science & Business Media; 1998. <http://www.springer.com/gp/book/9780792381983>.
18. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell.* 1997;19:153–8.
19. Prinja S, Gupta N, Verma R. Censoring in clinical trials: review of survival analysis techniques. *Indian J Community Med.* 2010;35(2):217–21. doi:10.4103/0970-0218.66859.
20. Yu S-L, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, Cheng CL, Yu CJ, Lee YC, Chen HS. MicroRNA signature predicts survival and relapse in lung cancer. *Cancer Cell.* 2008;13:48–57.
21. Choi I, Wells BJ, Yu C, Kattan MW. An empirical approach to model selection through validation for censored survival data. *J Biomed Inform.* 2011;44: 595–606.
22. Heng DY, Xie W, Regan MM, Warren MA, Golshayan AR, Sahi C, Eigl BJ, Ruether JD, Cheng T, North S. Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor–targeted agents: results from a large, multicenter study. *J Clin Oncol.* 2009;27:5794–9.
23. Adabag S, Rector TS, Anand IS, JJ MM, Zile M, Komajda M, RS MK, Massie B, Carson PE. A prediction model for sudden cardiac death in patients with heart failure and preserved ejection fraction. *Eur J Heart Fail.* 2014;16:1175–82.
24. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med.* 2002;8(8):816–24.
25. Ueda T, Volinia S, Okumura H, Shimizu M, Taccioli C, Rossi S, Alder H, Liu CG, Oue N, Yasui W. Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol.* 2010;11:136–46.
26. Hu X, Schwarz JK, Lewis JS, Huettner PC, Rader JS, Deasy JO, Grigsby PW, Wang X. A microRNA expression signature for cervical cancer prognosis. *Cancer Res.* 2010;70:1441–8.
27. Roessler S, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 2010;70:10202–12.
28. Tan Q, Thomassen M, Jochumsen K, Mogensen O, Christensen K, Kruse T. Gene selection for predicting survival outcomes of cancer patients in microarray studies. In: Sobh T, editor. *Advances in Computer and Information Sciences and Engineering.* Dordrecht: Springer; 2008. [https://link.springer.com/chapter/10.1007/978-1-4020-8741-7\\_73](https://link.springer.com/chapter/10.1007/978-1-4020-8741-7_73).
29. Yau C, et al. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Res.* 2010;12:R85.
30. Chen Y-C, Ke W-C, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med.* 2014;48:1–7.
31. Blanco R, Inza I, Merino M, Quiroga J, Larrañaga P. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *J Biomed Inform.* 2005;38:376–88.
32. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One.* 2011;6:e24709.
33. Wu T, Sun W, Yuan S, Chen C-H, Li K-C. A method for analyzing censored survival phenotype with gene expression data. *BMC Bioinform.* 2008;9:417.
34. Lee S, Rahnenführer J, Lang M, De Preter K, Mestdagh P, Koster J, Versteeg R, Stallings RL, Varesio L, Asgharzadeh S. Robust selection of cancer survival signatures from high-throughput genomic data using two-fold subsampling. *PLoS One.* 2014;9:e108818.
35. Ambler G, Seaman S, Omar R. An evaluation of penalised survival methods for developing prognostic models with rare events. *Stat Med.* 2012;31:1150–61.
36. Wu TT, Gong H, Clarke EM. A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival. *J Bioinforma Comput Biol.* 2011;9:63–73.
37. Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. *Ann Stat.* 2002;30(1):74–99.
38. Johnson BA. On lasso for censored data. *Electronic J Stat.* 2009;3:485–506.
39. Garcia RI, Ibrahim JG, Zhu H. Variable selection in the cox regression model with covariates missing at random. *Biometrics.* 2010;66:97–104.
40. He Z, Tu W, Wang S, Fu H, Yu Z. Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics.* 2014;71:178–87.
41. Sohn I, Kim J, Jung S-H, Park C. Gradient lasso for Cox proportional hazards model. *Bioinformatics.* 2009;25:1775–81.
42. Goeman JJ. L1 penalized estimation in the cox proportional hazards model. *Biom J.* 2010;52:70–84.
43. Cai T, Huang J, Tian L. Regularized estimation for the accelerated failure time model. *Biometrics.* 2009;65:394–404.
44. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Ser B.* 2005;67:301–20.
45. Khan MHR, Shaw JEH. Variable selection for survival data with a class of adaptive elastic net techniques. *Stat Comput.* 2016;26:725–41.
46. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc.* 2001;96:1348–60.
47. Leung KM, Elashoff RM, Afifi AA. Censoring issues in survival analysis. *Annu Rev Public Health.* 1997;18:83–104. doi:10.1146/annurev.publhealth.18.1.83.
48. Bandyopadhyay S, et al. Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Min Knowl Disc.* 2014;29(4):1–37.
49. Spelt L, Nilsson J, Andersson R, Andersson B. Artificial neural networks—a method for prediction of survival following liver resection for colorectal cancer metastases. *Eur J Surg Oncol.* 2013;39:648–54.
50. Dezfouli HN, Bakar MRA. 2012 International Conference on Statistics in Science, Business, and Engineering (ICSSBE). Langkawi: IEEE; 2012. p. 1–5.
51. Liu Y, Aickelin U, Feyereis J, Durrant LG. Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data. *Knowl-Based Syst.* 2013;37:502–14.
52. Neuvirth H, et al. Toward personalized care management of patients at risk: the diabetes case study 17th ACM SIGKDD. San Diego: ACM; 2011. p. 395–403.
53. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2:841–60.
54. Hsieh E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ Cardiovasc Qual Outcomes.* 2011;4:39–45.
55. Chen X, Ishwaran H. Pathway hunting by random survival forests. *Bioinformatics.* 2013;29:99–105.



56. Bühlmann P, Yu B. Boosting with the L2 loss. *J Am Stat Assoc.* 2003;98:324–39. doi:10.1198/016214503000125.
57. Buehlmann P. Boosting for high-dimensional linear models. *Ann Stat.* 2006;34:559–83.
58. Chen Y, Jia Z, Mercola D, Xie X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput Math Methods Med.* 2013;34:873595.
59. Koziol JA, Feng AC, Jia Z, Wang Y, Goodison S, Mc Clelland M, Mercola D. The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis. *Bioinformatics.* 2009;25:54–60.
60. Attallah O, Ma X. Bayesian neural network approach for determining the risk of re-intervention after endovascular aortic aneurysm repair. *Proc Inst Mech Eng H.* 2014;228:857–66. doi:10.1177/0954411914549980.
61. Karthikesalingam A, Attallah O, Ma X, Bahia SS, Thompson L, Vidal-Diez A, Choke EC, Bown MJ, Sayers RD, Thompson MM, Holt PJ. An artificial neural network stratifies the risks of Reintervention and mortality after endovascular aneurysm repair; a retrospective observational study. *PLoS One.* 2015;10:e0129024.
62. Janeczek A, Gansterer WN, Demel M, Ecker G. FSDM 90–105. Antwerp: On the Relationship Between Feature Selection and Classification Accuracy; 2008.
63. Ting KM, Quek RJY. Third IEEE international conference on data mining, ICDM 2003: IEEE. p. 653–6.
64. Han L. I. e. K. Model selection and model averaging for neural networks PhD thesis, Carnegie Mellon University Pittsburgh, (1999).
65. Ghatwary T, Karthikesalingam A, Patterson B, Hinchliffe R, Morgan R, Loftus I, Salem A, Thompson MM, Holt PJ. St George's vascular institute protocol: an accurate and reproducible methodology to enable comprehensive characterization of infrarenal abdominal aortic aneurysm morphology in clinical and research applications. *J Endovasc Ther.* 2012;19:400–14. doi:10.1583/11-3731MR.1.
66. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53:457–81.
67. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res.* 2010;1(4):274.
68. Kim H-J. Common factor analysis versus principal component analysis: choice for symptom cluster research. *Asian Nurs Res.* 2008;2:17–24.
69. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat.* 2010;2:433–59.
70. Jolliffe IT. Principal component analysis. 2nd ed: Springer; 2002. <http://www.springer.com/gp/book/9780387954424>.
71. Suhr DD. SUGI 30 203–230. Philadelphia: Principal component analysis vs. exploratory factor analysis; 2005.
72. Cox DR. Regression models and life-tables. *J Royal Stat Soc B.* 1972;34:187–220.
73. Cheng J, Greiner R. In advances in artificial Intelligence, 14th biennial conference of the Canadian Society for Computational Studies of intelligence, AI: Springer. p. 141–51.
74. Karegowda AG, Manjunath A, Jayaram M. Comparative study if attribute selection using gain ratio and correlation based feature selection. *Int J Inf Technol.* 2010;2:271–7.
75. Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian J Comput Sci Eng.* 2011;2:188–95.
76. Dziuban CD, Shirkey EC. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychol Bull.* 1974;81:358.
77. Tobias S, Carlson JE. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivar Behav Res.* 1969;4:375–7.
78. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 2006;7:1.
79. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics.* 2005;21:3301–7.
80. Matsunaga M. How to factor-analyze your data right: do's, don'ts, and how-to's. *Int J Psychol Res.* 2015;3:97–110.
81. Fabrigar LR, Wegener DT. Exploratory factor analysis. Oxford: Oxford University Press; 2011.
82. Costello A, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Pract Assess Res Eval.* 2005; 10. [pareonline.net/getvn.asp?v=10:t=2011](http://pareonline.net/getvn.asp?v=10:t=2011).
83. Perry BB. A genetic algorithm for learning Bayesian network adjacency matrices from data, PhD thesis. Manhattan: Kansas State University; 2003.
84. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn.* 1997;29:131–63.
85. Hassan AM. A probabilistic relaxation framework for learning Bayesian network structures from data master of science thesis. Giza: Cairo University; 2007.
86. Brown L, Greenhalgh R, Powell J, Thompson S. Use of baseline factors to predict complications and reinterventions after endovascular repair of abdominal aortic aneurysm. *Br J Surg.* 2010;97:1207–17.
87. Peppelenbosch N, et al. Diameter of abdominal aortic aneurysm and outcome of endovascular aneurysm repair: does size matter? A report from EUROSTAR. *J Vasc Surg.* 2004;39:288–97.
88. Lo BW, Macdonald RL, Baker A, Levine MA. Clinical outcome prediction in aneurysmal subarachnoid hemorrhage using bayesian neural networks with fuzzy logic inferences. *Comput Math Methods Med.* 2013;2013:1–11.
89. Wyss TR, Dick F, Brown LC, Greenhalgh RM. The influence of thrombus, calcification, angulation, and tortuosity of attachment sites on the time to the first graft-related complication after endovascular aneurysm repair. *J Vasc Surg.* 2011;54:965–71.
90. Stather PW, Wild JB, Sayers RD, Bown MJ, Choke E. Endovascular aortic aneurysm repair in patients with hostile neck anatomy. *J Endovasc Ther.* 2013;20:623–37.
91. Karthikesalingam A, Holt PJ, Vidal-Diez A, Choke EC, Patterson BO, Thompson LJ, Ghatwary T, Bown MJ, Sayers RD, Thompson MM. Predicting aortic complications after endovascular aneurysm repair. *Br J Surg.* 2013;100:1302–11. doi:10.1002/bjs.9177.
92. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed: Morgan Kaufman; 2005. <http://www.sciencedirect.com/science/book/9780123748560>.
93. Patterson BO, Hinchliffe RJ, Holt PJ, Loftus IM, Thompson MM. Importance of aortic morphology in planning aortic interventions. *J Endovasc Ther.* 2010;17:73–7. doi:10.1583/09-2967C.1.
94. Patterson BO, et al. Existing risk prediction methods for elective abdominal aortic aneurysm repair do not predict short-term outcome following endovascular repair. *J Vasc Surg.* 2010;52:25–30. doi:10.1016/j.jvs.2010.01.084.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

