Supplement:

1. Predictive models

Support vector machine (SVM) with Radial basis function (RBF) kernel: The goal of SVM is to construct optimal separating boundaries between different classes of dataset by solving a constrained quadratic optimization problem [24]. Assume a training set is denoted by $\{x_i, y_i\}, i = 1, \cdots, N$, where x_i represents D-dimension input vector the attribute and $y_i \in$ Error! Bookmark not defined. is the label. The linear SVM seeks to find the discriminant with the maximum margin in feature space by mapping function $\Phi: \mathbb{R}^D \to \mathbb{R}^S$

$$f(x) =< w \cdot \Phi(x) > +b, \tag{1}$$

In the training step, the classifier is trained by solving the following problem [25]:

$$\min_{w,b,\xi} \frac{1}{2} < w \cdot w > +C \sum_{i=1}^{n} \xi_i, \tag{2}$$

$$\text{Subject to: } y_i(< w \cdot \Phi(x_i) > +b) + \xi_i - 1 \geq 0, \xi_i \geq 0. \tag{3}$$

where w represents the normal vector to the hyperlane, C represents a penalty parameter on training error, and ξ_i represents the non-negative slack variables. The parameter b is the bias term of the separating hyperplane. This problem is solved by introducing the Lagrange dual function:

$$max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \tag{4}$$

where $k(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >$ is considered as kernel function and α is the vector of dual variables corresponding to each separation constraint. After optimizing this function, we can obtain$w = \sum_{i=1}^{N} \alpha_i y_i \, \Phi(x_i)$. Assume x is a test sample in the testing step, the discriminant function is:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k < x \cdot x_i > +b. \tag{5}$$

Artificial neural network (ANN): ANN is inspired by biological neural networks, a group of interconnected artificial neurons closely interacting with one another [26]. A typical multilayer feedforward neural network structure is presented in Figure S1, which includes an input layer, hidden layers and an output layer. The training process is designed to adjust the connection weight. In each training process, the network calculates the response of the output neuron and if an error occurs after a particular input pattern, the weights and bias will be changed accordingly. The training will stop until no error occurs. Hence, a nonlinear mapping between input features and output classes can be constructed. In testing process, a probability for each label is obtained by the constructed nonlinear mapping and the corresponding label is predicted. In this work, the feed forward neural network is adopted, with a single hidden layer, and the number of nodes is set as 10 [27].

Logistic Regression (LR): LR provides a function to transform the parameters into a probability calculated in the following way [28]. Assume there is a binary outcome variable y and a vector $x = (x_1, \cdots x_I)$. The model relates y to x by assuming that [29]

$$Pr(y = 1|x) = \varphi(\beta_0 + \sum_{i=1}^{I} \beta_i x_i) \ , \tag{7}$$

where $\varphi(\mu) = \frac{1}{1+\exp(-\mu)}$ represents the logistic function. The regression coefficients β_i are calculated by using the maximum-likelihood estimation in the training process. In the testing process, the probability $Pr(y = 1|x)$ can be predicted by

$$Pr(y = 1|x) = \varphi(\hat{\beta}_0 + \sum_{i=1}^{I} \hat{\beta}_i x_i), \tag{8}$$

where $\hat{\beta}_i$ denotes the optimal coefficient. The outcome of the constructed model is represented as a binary variable, which corresponds to distant failure or not.

2. Feature selection methods

2.1 Clonal selection algorithm (CSA)

The feature selection process is critical to the performance of machine learning methods. In this work, clinical parameter selection is considered as a combinatorial optimization problem. Specifically, CSA [33] is used to select optimal clinical parameters for each predictive model. Compared to analytical optimization methods [34], one of the major theoretical advantages of CSA is the achievement of global combinatorial optimization. The details can be seen in [18].

In this work, the area under the ROC (receiver operating characteristic) curve (AUC) is used as the evaluation criterion for clinical parameter selection and model parameter training. AUC integrates the information of specificity and sensitivity, which provides more comprehensive information than the accuracy alone as the evaluation criterion [19] [35]. 5-cross-validation is utilized to validate the performance. The goal of CSA is to find optimal combination of selected clinical parameters and model parameters to maximize AUC. The detailed procedures to implement CSA are presented as follows:

Step 1: Initialization

In CSA, both clinical parameter selection and model parameter training are performed. Thus, hybrid initialization is needed for clinical parameters and model parameters. The initial solution set is generated randomly and it is called population. One particular solution from this population is called individual. For clinical parameters, it is encoded by a binary encoding method. Each locus in the coding represents one clinical parameter or one model parameter. The value of "1" in a particular locus indicates that the corresponding clinical parameter is selected, whilst "0" denotes that the corresponding clinical parameter is not selected. Model parameters are continuous and they are optimized directly. The coding length is the sum of the number of clinical parameters and model parameters.

Figure S2.A shows an example of the encoding method. Assume that there are 6 clinical parameters and 2 model parameters. The upper level represents the locus and the lower level represents the corresponding values. In this example, clinical parameter 2, 3 and 5 are selected. Two model parameters are set as "-5" and "3", respectively.

Step 2: Clonal operator

Based on the evaluation criterion, AUC is calculated for each individual (i.e., solution) in the population (P). Individuals in a population are then ranked in decreasing order. The top 30% of individuals are cloned and those ranked as highest are cloned more times. The number of clones for each individual is calculated as:

$$N_i = round(\frac{\beta \cdot P}{i}), \quad (9)$$

where N_i represents the clonal number for ith cloned individual and β is the scale factor. When this part is complete, the new cloned population is generated for next step.

Figure S2.B shows an example of clonal selection operation. Assume there are 8 individuals in a population. The objective function for each individual was calculated and ranked in decreasing order. Assume that β is set as 1, it is evident that 4 individuals can generate 8, 4, 3, 2 new clonal individuals, respectively. The clonal population consists of the original population and newly generated individuals.

Step 3: Mutation operator

In this step, the mutation operator is executed in the clonal population. For each locus, a random mutation probability $RP_i, i = 1, \cdots, CL(CL\ is\ the\ coding\ length)$ is initially generated. Assume that the general mutation probability for each locus is $GMP_i, i = 1, \cdots, CL$. If $GMP_i > RP_i$, the corresponding locus will mutate. Figure S2.C illustrates an example of the mutation process. For the example we still assume that there are 6 clinical parameters and 2 model parameters in an individual. Assume that $GMP_i = 0.5, i = 1, \cdots, 8$, and the randomly generated probabilities are [0.81, 0.75, 0.44, 0.52, 0.31, 0.28, 0.64, 0.37]. The upper layer in Fig. 2.C represents the individual before mutation, while the lower represents the individual after mutation. The mutation will occur only if the RP_i is larger than 0.5. After this step, the mutation population will be generated.

Step 4: Evaluation and selection.

In this step, after the AUC of the individuals in the mutation population were calculated, they were ranked in decreasing order. To maintain the population size, top P individuals are selected and a new population is generated.

Step 5: Termination test.

If the algorithm reaches the maximal generation G, it will stop and the individual with the maximal AUC in population is considered as the output. Otherwise, step 2-4 are performed.

2.2 Sequential forward feature selection (SFS)

SFS method is a classic feature selection method [31]. Assuming there are N features in original feature space, SFS begins at an initially empty feature set. In each forward step, a feature from N is selected if the prediction accuracy achieved by the constructed predictive model is improved. The algorithm will stop until no further feature is found. It is important to note that SFS can only select features, while CSA can select the optimal feature set and optimize model parameter simultaneously.

2.3 Statistical analysis based method (SA)

In this work, statistical analysis based feature selection was performed using logistic regression. In the logistic regression model, backward elimination procedure was used with the entering p-value criterion from the univariate analysis set at 0.20, allowing any potentially significant features to be included in the initial model. All features remained in the final model have a p-value less than 0.05. For certain continuous features that we also considered dichotomization, additional analyses were performed for each data type. Based on logistic analysis, four features were selected including ethnicity, gender, stage, dose per fraction.

3. Model parameters

The clinical and model parameters were selected by CSA in three predictive models. In the training step, both the population (P) and the maximal generation (G) in CSA were set as 100. For SVM, Radial Basis Function (RBF) was adopted as the kernel function [16]. In the SVM training, the coding length was 20, consisting of 18 clinical parameters and 2 model parameters. As ANN and LR do not require the optimization of model parameters, the coding length was 18. Clonal probability was set as 0.3 and β was set as 1 in the clonal operator, while $GMP_i = 0.5, i = 1, \cdots, CL$ in mutation operator. When comparing three clinical parameter selection methods, SVM is taken as the predictive model and the model parameters are set to equal values.

In addition to AUC, the performance of different predictive models and feature selection strategies was also evaluated by sensitivity and specificity.

As ANN and LR don't have model parameters to train, we only reported model parameters for SVM. For RBF in SVM, two parameters were trained: 1) the kernel parameter that defines the influence of a single training sample; and 2) the cost efficient that defines the tradeoff of misclassification between training samples and the simplicity of the decision surface. In this study, the optimized kernel parameter and cost efficient were -4 and -2 respectively.