Predicting Xerostomia after IMRT treatments: a data mining approach

Abstract

Background and Purpose: Xerostomia is one of the most frequent long term side-effects experienced by head-and-neck cancer patients undergoing radiation therapy, reducing drastically the quality-of-life of patients. In the present study, a prediction model for xerostomia after radiotherapy is proposed.

Material and Methods: Model construction was based on a dataset of 138 patients with head-and-neck cancer treated at the Portuguese Institute of Oncology of Coimbra (IPOCFG) with Intensity Modulated Radiation Therapy, using different data mining predictors. The models considered dosimetric information and patient specific features known prior to treatment to estimate which patients will experience xerostomia (G0 vs G1/G2 according to RTOG/EORTC). The quality of the classifiers was assessed by applying cross-validation procedures and was validated by different datasets. ROC/AUC, precision and recall were the measures used to evaluate the models' performance.

Results: Age, gender, severity of xerostomia prior to radiation therapy and planned mean (physical) dose in both parotids revealed to be relevant predictors of xerostomia. The best model was the one based on random forests. The method produced an AUC equal to 0.73, a precision of 72% and a recall of 83% considering the threshold 0.5.

Conclusions: The ability to discriminate patients according to their features helps to achieve personalized radiation therapy treatments. Random forests revealed to be a good classification method for predicting the binary response "risk for xerostomia induced by radiation therapy at 12 months", showing a high discriminative ability.

Keywords: head-and-neck cancer, radiotherapy, IMRT, xerostomia, random forests, data mining.

1. Introduction

Radiation therapy is one of the main modalities used for cancer treatment, alone or in combination with surgery and chemotherapy. The objective of radiotherapy treatments is to be able to destroy all cancerous cells, but at the same time to spare all healthy cells and organs. Intensity Modulated Radiation Therapy (IMRT) is one of the forms of radiation therapy treatment. Its main feature is the ability to conform the delivered radiation to the volumes to treat. This is achieved through a discretization of the radiation beams into a set of individual beamlets that can have different intensities.

In radiation therapy, the medical doctor will prescribe the desired treatment for each patient. This prescription is composed of a set of radiation dose constraints that should be satisfied and that, in most cases, define maximum and minimum limits to the dose to be delivered to the patient. The dose delivered should be such that allows the best possible irradiation of the volumes to be treated but that also protects the organs at risk (namely, guaranteeing their proper functioning during and after the treatment). The treatment is delivered during a predefined number of treatment sessions. IMRT allows the achievement of a high degree of conformity between the volume to be treated and the prescribed dose distribution allowing high sparing of the surrounding healthy tissues when compared with conventional treatment techniques [1-5]. Despite improvements obtained with IMRT in head-and-neck cancer patients, sparing of the salivary glands is still challenging. The irradiation of these organs at risk can result in salivary dysfunction and consequently xerostomia, one of the most frequent long term side-effects. Xerostomia is characterized by the feeling of dry mouth due to the lack of saliva, reducing drastically the quality-of-life of patients due to difficulties in swallowing and in feeding [6-10]. There is thus a great interest in the development of accurate models capable of predicting whether a given patient will experience xerostomia after radiation therapy. If it was possible to predict the future occurrence of xerostomia, this could be taken into account during the treatment planning phase, trying to reach treatment plans that would spare as most as possible the salivary glands.

The first predictive models for radiation therapy outcomes, and in particular xerostomia on head-and-neck cancer patients, were mostly based on dosimetric information [11-14]. The work developed by El Naqa et al. [15] was the pioneering study incorporating not only dosimetric information but also other prognostic clinical factors in outcome prediction models for head-and-neck cancer patients treated with IMRT. After that, a proliferation of predictive models estimating xerostomia, mostly based on dosimetric data, was observed [11, 16-20]. Naqa et al. [20] and Blanco et al. [11] presented studies with a small number of patients. The authors measured the salivary flow at 6 and 12 months and predicted radiotherapy outcomes using support vector machines and multivariate logistic regression. According to these authors, the prediction of treatment response can be improved by discovering nonlinear interactions among model variables. They concluded that, with conventional fractionation, the incidence of xerostomia was significantly smaller when the mean dose of at least one parotid gland was below 25.8 Gy. However, even when patients were irradiated with doses lower than this threshold value, they experienced a delayed recovery of salivary function. Beetz et al. [17-19], using a larger dataset, concluded that the models developed in a population treated with a specific technique cannot be generalized and extrapolated to a population treated with another technique without external validation. Using logistic regression they concluded that the inclusion of predictive factors other than dose-volume histogram parameters can significantly improve model performance. Similar results were also accomplished by using logistic regression in our exploratory study [16]. The main drawback was the small number of patients and the great amount of predictive attributes. Therefore, there is an enormous need for the development of predictive models that include explanatory features beyond the dosimetry.

In the present work, several different data mining models were applied for prediction of radiation-induced complications in the salivary glands for head-and-neck cancer patients irradiated with IMRT. The models included dosimetric information but also patient specific features known prior to the treatment. The data mining approaches considered were classifiers that were trained to be able to predict the binary response “risk for xerostomia induced by radiation at 12 months”. The models were trained considering a database of head-and-neck cancer patients treated at the Portuguese Institute of Oncology of Coimbra. Testing and validation were made using different sets of patients.

In the next section, a description of the materials and methods used is detailed. Section 3 presents the main results. Section 4 presents a critical discussion. Some conclusions and paths for future work are presented in section 5.

2. Material and Methods

2.1. Dataset

Population cohort included patients with head-and-neck tumors treated at IPOCFG with different forms of IMRT [21]. All clinical and treatment patient data was retrieved from the electronic health information system RESPONSE [22]. The system comprises a number of patient features and medical registrations: patient and tumor characteristics, treatment details and patient response to radiation therapy registered during the follow-up medical consultations.

This study considered the development of a model to predict xerostomia, based on the knowledge acquired from retrospective data. The model considered only predictive attributes known at the beginning of the treatment, capable of influencing the posterior development of xerostomia. The aptness for xerostomia will be related to the delivered treatment and not to the planned one. Therefore, patients that interrupted radiation therapy were not considered in this study, since the delivered doses may not correspond to those planned.

Classification of side-effects at IPOCFG is made using RTOG/EORTC guidelines. Thus, complications severity is ranked from 0 to 5, where 0 means no complication and 5 death from toxicity [23]. In this study, we were only interested in predicting patients' binary response, regardless of the complication degree. Thus, all severity degrees equal to 1 or 2 (where 2 was the maximum severity degree obtained at IPOCFG) were grouped. Only two severity classes were considered: "1" if the patient presented xerostomia; "0" otherwise. IPOCFG dataset had 138 patients: 52 belonging to class "0", i.e, being complication-free, and 86 belonging to class "1". The original dataset was thus unbalanced, since 62% of patients belonged to class "1" and 38% belonged to class "0". Patients demographic, clinical and treatment characteristics are listed in Table 1. In this table it is also possible to find information regarding patients that presented xerostomia at baseline, i.e. prior to the beginning of radiation therapy treatments. Xerostomia at baseline was an exploratory variable incorporated in the predictive models tested.

An overview of the corresponding calculated mean doses to organs at risk is depicted in Table 2.

2.2. Predictive Features

In order to develop a classification model that allow us to predict whether a given patient will experience xerostomia 12 months after the radiotherapy treatment, we should consider as valid inputs only those attributes that are known prior to the beginning of the radiation therapy or, at most, during the first weeks of treatment. Being able to predict xerostomia at early stages of the treatment will make it possible to adjust the treatment plan in order to minimize or avoid this complication.

There are a myriad of variable screening techniques that can be applied to help in the selection process of the most relevant features characterizing a specific dataset regarding the relation with a given outcome. A dataset of 138 patients can be considered a small dataset (see our previous work [24]). In such circumstances, the instability of the selection methods and the limited power to select relevant attributes can lead to a loss in predictive ability, because more information can be lost than gained. Therefore, alternative strategies should be taken into account for small datasets [24, 25]. The use of external information, such as clinical knowledge or information from other studies, is highly important in the variable selection and estimation processes [24-26]. This information can not only improve the predictive performance of the model, but also increase its clinical credibility. The medical team can feel more comfortable and rely more in the use of a model that includes the team's inputs rather than in a model constructed based on variable screening practices only [25, 27, 28]. In small datasets, the model should, as much as possible, be based on external knowledge that can be expected to describe the patterns in the dataset sufficiently well [25]. Thus, in this retrospective study, we incorporated clinical knowledge in the modeling process, by considering four attributes indicated by the medical team as probably being highly associated with xerostomia outcome:

(1) Patients data: age and gender;

(2) Severity of xerostomia prior to radiation therapy;

(3) Calculated mean (physical) dose on the contralateral and ipsilateral parotids.

These features were all known at the beginning of the treatment for every patient.

2.3. Predictive Models

The problem of predicting a response for a new patient based on a model derived from a dataset of previously treated patients can be seen as a machine learning problem, namely, a classification learning problem. In a classification problem, a training dataset consisting of *n* elements is available. Each element is characterized by a *p*-dimensional attribute vector *x*, belonging to a suitable space, and a class label (also known as response) $y \epsilon \{0,1,\dots\}$. The objective is to construct a decision or classification rule (also known as predictor, classifier or model) that will accurately predict the class labels of elements for which only the attribute vector is observed [29].

The output of a classifier will be a value belonging to [0,1] that corresponds to the probability of a patient belonging to a specific class. Classifying a patient will thus require this probability to be translated into a binary output: 1 or 0. This can be done by considering a threshold α: if the probability is greater than α, the assigned class should be "1", and "0" otherwise. Each α represents a decision boundary in the feature space. The most used threshold is the value 0.5.

Clinical knowledge was incorporated in the modeling process, by considering the four attributes indicated by the medical team.

We applied different data mining classifiers in order to classify new patients according to the probability of experience xerostomia 12 months after the beginning of IMRT treatments. The classifiers used were Random Forests, Stochastic Boosting, Support Vector Machines, Neural Networks, Model-based Clustering and Logistic Regression. In the following we will describe each methodology used.

Random Forests

A random forest consists in a collection of tree-structured classifiers [30], where leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The random forest classifier works as an ensemble of decision trees predictors, where each tree is constructed based on a random selection of observations of the working dataset (also called training set). The main essence of this procedure is to build multiple trees in randomly selected subspaces of the feature space, such that locally-optimal decisions are made at each node. The split in each node is made according to the best feature among all possible features on the subspace. The classification of a new observation corresponds to the class that is the mode of the classes outputted by individual trees.

Stochastic Boosting

The basic idea of stochastic boosting is to combine very simple classification rules to form an ensemble, with a significantly improved performance. All elements have an initial weight and, at each iteration, the weights are recalculated. The correct classified elements have their weights decreased and those incorrectly classified have the weight increased. The aim is to direct the classifier to these elements. The final classification is a weighted majority vote of all the trained weak learners where each weak learner has one vote. The most usual weak learners applied are classification trees and the most used algorithm is *AdaBoost* [31].

Support Vector Machines

Support vector machines efficiently perform a non-linear classification implicitly mapping the observations into a high-dimensional feature space using a set of mathematical functions known as kernels. The basic idea behind support vector machines is the construction of a hyperplane in a higher dimensional space defining a decision boundary to separate the set of elements having different class memberships. The algorithm selects prototypes from the training data lying on the board between two classes in order to derive the classification rule for new data [32]. Support vector machines implementations require the user to define some parameters, namely the kernel function and a cost parameter used to penalize the classifier for incorrect classifications of the training data. The error of misclassifications can be minimized by an adequate choice of the kernel function.

Neural Networks

A neural network is an interconnected group of nodes. This structure was inspired by the central nervous system and explored for addressing an array of problems [33]. Formally, a neural network is an information processing paradigm composed by a large number of highly interconnected processing elements (known as neurons), organized by layers and working in unison to solve specific problems. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer where the answer is finally yielded. Within each hidden layer neuron there is a sigmoidal activation function that polarizes network activity, as a function of a weighted sum of its inputs, and helps it to stabilize by modifying the weights of the connections according to the input patterns to decrease the differences between the neural network outputs and the true outputs of the training data. Neural network analysis often requires a large number of individual runs to obtain the best solution.

Model-based Clustering

Model-based clustering assumes that all elements of the original dataset are created by a mixture of components, each described by a density function and having an associated probability or “weight” in the mixture. The class of a new element will correspond to the group defined by the mixture component that most likely created it [34]. We can adopt any probability model for components, but typically it is assumed the Gaussian finite mixture model, where each component is modeled by a single Gaussian term with the same covariance structure among classes. This procedure is well-known as Eigenvalue Decomposition Discriminant Analysis (EDDA). Furthermore, the covariance matrix can assume several parameterizations, which leads to different models with different interpretations.

Logistic Regression

Logistic regression classifier (also known as logit model) measures the relationship between a dependent variable (also called response) and one or more independent variables, by using probability scores as the predicted values of the dependent variable. The probabilities are modeled as a function of the explanatory variables by using a logistic function [35].

R software was used, namely the packages “randomForest” [36], “ada” [37], “kernlab” [38], “nnet” [39] and “mclust” [40] and the “glm” function.

All the classifiers considered had parameters that should be fixed a priori. Several computational tests were done to try to find the best parameters. Random Forests were considered as having 500 trees, Stochastic Boosting model considered 100 trees, the Support Vector Machine model considered a radial basis function (with automatic sigma estimation) as the kernel function and a penalty of 1, the neural network predictor had a single inner layer with 5 nodes and the Model-based Clustering classifier was fitted with the Gaussian multivariate mixture model EII (spherical and equal volume).

2.4. Performance Measures

Bootstrapping is a technique used to iteratively improve the classifier performance where a sample with a size higher or equal to the original dataset is considered by random selection with replacement [41]. We used bootstrapping samples with 500 observations. To assess the suitability of the models, we used a cross-validation technique [42]. Cross-validation involves partitioning the available data sample into complementary subsets, performing the analysis on one subset (training set) and validating the analysis on the other subset (validation or testing set) [43].

After running all cross-validation iterations in the bootstrap sample, we created the Receiver Operating Characteristics Curve (ROC) and determined the Area Under the ROC Curve (AUC) in order to assess the performance of the classifiers and measure the discriminative ability of the models.

The results of correlation tests between attributes and classifications are not very useful, because strong correlations do not imply good predictors. A model is useful if it efficiently separates ‘responders’ from ‘non-responders’ and the metric that quantifies this ability is the AUC, which is the probability that the model will correctly rank sampled ‘responder’ and ‘non-responder’ pairs from the data set. We considered the intervals [0.5,0.7[, [0.7,0.9[and [0.9,1] to define the performance of the model as poor, moderate/good and excellent, according to [44]. Also, for a binary outcome, ROC and AUC are the most commonly recommended and used performance measures to judge the discriminative ability of a model between the observations with and without the characteristic [45-47]. The AUC can be interpreted as the probability that a patient with xerostomia 12 months after IMRT treatments is given a higher probability of the outcome by the model than a randomly chosen patient without the outcome.

3. Results

3.1. Predictive Models

The suitability of the predictive models was assessed by applying the leave-one-out cross-validation (LOOCV) procedure. This method uses all elements of the original dataset except one as training data, and the remaining single observation as validation data. Hence, all observations with exception of one are used to train the model, which is then used to predict the class for the remaining observation.

The features and performance of the different predictive models were compared by generating the ROC curves for each classification model and computing the respective AUC. The best result was obtained when considering random forests, which results in an AUC of 73% (Figure 1A). The ROC curve generated by the stochastic boosting model resulted in an AUC of 65% (Figure 1B). For the support vector machine model, the AUC achieved was of 66% (Figure 1C). The neural network predictor resulted in an AUC of 61% (Figure 1D). Model-based Clustering classifier presented an AUC of 43% (Figure 1E). The last model tested was the logistic regression, which exhibited an AUC of 47% (Figure 1F).

For that prediction model that revealed a better performance (Figure 1A) complementary analyses were also carried out in order to confirm the robustness of the classifier. Such experiments are exposed in the following subsections.

3.2. Random Forest Predictive Model

3.2.1. n-factor Predictive Models

For the best classifier (random forest model), and in order to support clinical knowledge, all possible n-factor predictive models ($n\epsilon\{1,2,3,4\}$) were built and the results are shown in Table 3. The highest performance was accomplished by the model that considers the four attributes (Table 3). Considering $\alpha=0.5$ in the 4-factor random forest predictive model, in addition to reach an AUC of 73%, we also obtained an accuracy of 70%, a precision of 72% and a recall equal to 83%.

3.2.2. 6-fold Cross-Validation Analysis

The suitability and performance of the 4-factor random forest predictive model were also assessed by a 6-fold cross-validation procedure. This process consisted in the partitioning of the original dataset, into 6 complementary subsets, each with 23 patients. Each subset was used once as testing set, whereas the remaining patients were all used to train the model. This process was applied 6 times such that each complementary subset was used once as testing set. Similarly to the previous situation, the bootstrap samples had 500 instances. The process was repeated 100 times, obtaining an average AUC of 0.69 and standard deviation of 0.03.

3.2.3. Model Validation

In order to validate the predictive model, we randomly selected 24 elements from the dataset, obtaining two subsets from the original sample. The new sample composed by the remaining 114 observations was used to train the model, while the 24 separated elements were then used to validate the model. The validation process was repeated 100 times producing an average AUC of 0.69 with standard deviation of 0.03. Considering α=0.5, we obtained an average accuracy of 0.68 and standard deviation of 0.09 for the validation sample set. Figure 2 illustrates the histogram generated by the accuracy values obtained in 100 iterations considering α= 0.5. Each bar [*a*,*b*[of the histogram quantifies the number of iterations that resulted in an accuracy value between *a* (inclusive) and *b* (exclusive). It is worth noting that the worst accuracy value in the [0.6,0.7[interval was 0.62. Therefore, and as can also be seen in Figure 2, 80% of the total iterations (that is the sum of the frequencies of the last 6 bars of the histogram) resulted in an accuracy higher than 62% (proportion of patients in class “1”). This means that in 80 out of 100 run iterations the classifications obtained were better than classifying all patients as belonging to class “1” (class that comprises 62% from the total size of the dataset).

4. Discussion

We applied several data mining classifiers for predicting xerostomia induced by radiation treatments in head-and-neck cancer patients at 12 months. The basis of a predictive study is the set of explanatory features that characterize the original sample set. The majority of available radiobiological models only incorporate dose information in the analyses. In Naqa et al. [15], the authors determined that the analysis can be enriched and be more reliable by carefully select the most adequate explanatory features and increasing larger datasets. Naqa et al. performed analogous studies to investigate radiotherapy outcomes [15, 20]. They did not evaluate the performance of their methodologies by computing the AUC but correlated their approaches with other equivalent applications.

Beetz and his colleagues also constructed predictive models to study xerostomia and sticky saliva at 6 months of treatments [18, 48], focusing their attention into two different treatment techniques, precisely, three-dimensional conformal radiotherapy and IMRT. Concerning the first treatment, they obtained AUC values of 82% and 84%, respectively. When considering IMRT, the values were significantly lower, 66% and 63%, respectively. These values were also lower than those reached by our 4-factor random forest classifier (our forecasting model that resulted in a better performance), when considering the same treatment technique, which was equal to 73%. The studies performed by Beetz and his colleagues present two drawbacks: the use of small datasets and the elimination of patients with moderate to severe xerostomia or sticky saliva at baseline. This circumvented inherent complications in the construction of the classification model, but cannot reflect the real scenario of therapies and possibly not leading to reliable predictions.

The present work can also be compared with previous work of the same authors [16], where an AUC of 73% was also achieved, but considering a significant smaller dataset (only 49 patients) and a considerably higher number of attributes (15 predictive features), preventing further validation tests to be performed.

A compromise between the number of elements and the number of attributes used is of great importance in predictive analysis to guarantee a good performance of the model and to avoid overfitting. Commonly, a concession of 10 observations per predictive variable is assumed. Nevertheless, the 1:10 rule is somewhat arbitrary. Moreover, the total sample size is highly significant. According to Steyerberg et al. [25], for small datasets, the external clinical knowledge should be a priority. We used the valuable expertise of the medical team to select a small number of features that were considered as influencing most the future occurrence of xerostomia. The random forest predictive model, based on the set of four attributes suggested by clinicians, was the data mining approach that presented better results regarding the prediction of aptness for xerostomia induced by radiation at 12 months. The classifier produced results with high predictive capability, showing that clinical experience and knowledge was really an excellent support for predictive studies.

As far as the authors know, this was the first time that random forests encompassing only pre-treatment predictors were applied with the aim of determining a potential adverse-effect of xerostomia in radiation treatments. The majority of available approaches was only based on dosimetry data [11, 15, 20]. Consequently, these predictive models might not be as accurate for individualized clinical decision-support system for routine care [49]. Our 4-factor random forest approach, being based on a small set of four pre-treatment known attributes revealed a good performance. Thus, it might have a great clinical utility being capable of correctly estimate the class for new patients.

Our best predictive method was also evaluated using resampling techniques and validated on independent datasets. All possible combinations of the four chosen attributes were tested. As can be seen in Table 3, using AUC as the measure of the discriminative ability of the models, the best model was the one considering all four attributes simultaneously. Although the dosimetric feature revealed the most significant univariate prognostic ability, these results highlight the importance of considering other attributes in addition to dosimetric information.

The 4-factor classifiers yielded a probability consisting in a numerical value that represents the degree to which a patient is a member of class “1”. The obtained results by LOOCV for the random forest approach displayed a high AUC value of 0.73, which revealed a high performance of the predictive model, highly consistent with the true classifications. All similar systems and classification approaches investigated exhibit poor values of AUC measure. The high accuracy value (70%) as well as the high precision (72%) and recall (83%) values, obtained for the most commonly used threshold value of 0.5, reinforced the great performance and consistence of the 4-factor random forest model. These values mean that our 4-factor random forest model was able to correctly predicting the xerostomia class for 70% of the patients. Additionally, from all patients estimated by the 4-factor random forest model as having aptness for xerostomia 12 months after IMRT, 72% really developed this complication. For the 86 patients that really presented xerostomia after 12 months of radiation therapy, our predictive 4-factor random forest model correctly predicted 83%.

When applying a 6-fold cross-validation procedure to the best reached forecasting model, the average AUC presented also a good value, 0.69. Moreover, the very small average standard deviation, resulting in an AUC of 0.69±0.03, revealed once again that the model was able of correctly estimating the aptness for xerostomia complication after 12 months of the beginning of radiation therapy for new head-and-neck cancer patients. Validation analysis produced similar AUC results and thus equally high performance and discriminative ability of the 4-factor random forest model to make the predictions for new patients. In the scope of validation analyses, we also explored the accuracy results obtained when considering the most commonly used threshold, 0.5, which resulted in an accuracy of 0.68±0.09. Indeed, in 80% of the studied cases the obtained accuracy produced a value higher than 0.62, which means that in 80% of the analyzed cases, the predictions were better than classifying all patients as belonging to class “1” (which comprises 62% from the total 138 head-and-neck cancer patients).

5. Conclusion

Random forests proved to be good classifiers for predicting the binary response “risk for xerostomia at 12 months induced by radiation therapy treatments”, showing a high discriminative ability. The role of the four attributes: age, gender, severity of xerostomia prior to radiation therapy and planned mean physical dose in the contralateral and ipsilateral parotids appeared to be of main importance to the development of the radiation therapy side-effect xerostomia. The corresponding 4-factor random forest model can make highly reliable predictions of xerostomia complication. The importance of detecting prior to treatment, the radiation-induced complications has, as major advantage, the possibility of optimizing treatment plans trying to avoid this complication or at least minimize such side-effect.

Future work will examine other aspects of nonlinear modeling outcomes, such as applying data mining algorithms to address not only the short term and long term estimates of radiation treatment-induced complications but also the tumor response prediction problem. The developed models, as well as the obtained results can, in the future, be integrated in the optimization processes of radiation treatment planning.

Acknowledgements

This work was supported by FCT, Fundação para a Ciência e a Tecnologia, under project UID/MULTI/00308/2013, INESC Coimbra.

Conflict of interests

All authors declare that they have no conflicts of interest.

References

1. Lee NY, Terezakis SA. Intensity-modulated radiation therapy. Journal of Surgical Oncology 2008;97:691-6.
2. Yovino S, Poppe M, Jabbour S et al. Intensity-Modulated Radiation Therapy Significantly Improves Acute Gastrointestinal Toxicity in Pancreatic and Ampullary Cancers. Int J Radiat Oncol Biol Phys 2011;79:158-62.
3. Wang TJC, Riaz N, Cheng SK, Lu JJ, Lee NY. Intensity-modulated radiation therapy for nasopharyngeal carcinoma: a review. J Radiat Oncol 2012;1:129-46.
4. Kollmeier MA, Zelefsky MJ. Intensity-Modulated Radiation Therapy for Clinically Localized Prostate Cancer. Radiotherapy in Prostate Cancer (Eds.) Medical Radiology, Springer Berlin Heidelberg 2014 95-102.
5. Poitevin-Chacón MA, González GR, Zermeño AA et al. Implementation of intensity modulated radiotherapy for prostate cancer in a private radiotherapy service in Mexico. Reports of Practical Oncology & Radiotherapy 2015;20:66-71.
6. Jellema AP, Slotman BJ, Doornaert P, Leemans CR, Langendijk JA. Impact of radiation-induced xerostomia on quality of life after primary radiotherapy among patients with head and neck cancer. Int J Radiat Oncol Biol Phys 2007;69:751-60.
7. Messmer MB, Thomsen A, Kirste S, Becker G, Momm F. Xerostomia after radiotherapy in the head & neck area: long-term observations. Radiother Oncol 2011;98:48-50.
8. Vissink A, Jansma J, Spijkervet FK, Burlage FR, Coppes RP. Oral sequelae of head and neck radiotherapy. Crit Rev Oral Biol Med 2003;14:199-212.
9. Wijers OB, Levendag PC, Braaksma MM, Boonzaaijer M, Visch LL, Schmitz PI. Patients with head and neck cancer cured by radiation therapy: a survey of the dry mouth syndrome in long-term survivors. Head Neck 2002;24:737-47.
10. Dreizen S, Daly TE, Drane JB, Brown LR. Oral complications of cancer radiotherapy. Postgrad Med 1977;61:85-92.
11. Blanco AI, Chao KSC, El Naqa I et al. Dose–volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy. Int J Radiat Oncol Biol Phys 2005;62:1055-69.
12. Roesink JM, Moerland MA, Battermann JJ, Hordijk GJ, Terhaard CHJ. Qantitative dose-volume response analysis of changes in parotid gland function after radiotheraphy in the head-and-neck region. Int J Radiat Oncol Biol Phys 2001;51:938-46.

13. Lyman JT. Complication probability as assessed from dose-volume histograms. Radiat Res Suppl 1985;8:S13-9.
14. Kutcher GJ, Burman C. Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method. Int J Radiat Oncol Biol Phys 1989;16:1623-30.
15. El Naqa I, Bradley J, Blanco AI et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. Int J Radiat Oncol Biol Phys 2006;64:1275-86.
16. Soares I, Dias J, Rocha H, Lopes MC, Ferreira B. Predicting Xerostomia induced by IMRT treatments: a logistic regression approach. Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference. IEEE. 2014;72-77.
17. Beetz I, Schilstra C, van der Schaaf A et al. NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: the role of dosimetric and clinical factors. Radiother Oncol 2012;105:101-6.
18. Beetz I, Schilstra C, van Luijk P et al. External validation of three dimensional conformal radiotherapy based NTCP models for patient-rated xerostomia and sticky saliva among patients treated with intensity modulated radiotherapy. Radiother Oncol 2012;105:94-100.
19. Beetz I, Schilstra C, van Luijk P et al. Role of minor salivary glands in developing patient-rated xerostomia and sticky saliva during day and night. Radiother Oncol 2013;109:311-6.
20. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol 2009;54:S9-30.
21. Ferreira BC, Marques RV, Khouri L, Santos T, Sá-Couto P, Lopes MC. Assessment and topographic characterization of locoregional recurrences in head and neck tumours. Radiother Oncol 2015;10:41.
22. Ferreira BC, Khouri L, Lopes MC, Ferreira H. RESPONSE, an Electronic Health Patient Information Software for Radiation Therapy. Proceedings of the 6th Europ Conf of the Int Fed for Med and Biol Engin 2015;45:691-4.
23. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys 1995;31:1341-6.
24. Soares I, Dias J, Rocha H, do Carmo Lopes M, Ferreira B. Feature Selection in Small Databases: A Medical-Case Study. XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016, IFMBE Proceedings 2016;57:808-813.
25. Steyerberg EW, Eijkemans MJC, Jr FEH, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19:1059-79.
26. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421-33.
27. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373-9.
28. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488-94.
29. Culp M, Johnson K, Michailidis G. ada: An R Package for Stochastic Boosting. J Stat Soft 2006;17:
30. Breiman LEO. Random Forests. Mach Learn 2001;45:5-32.
31. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 1997;55:119-139.
32. Belousov AI, Verzakov SA, von Frese J. A flexible classification approach with optimal generalisation performance: support vector machines. Chemometrics and Intelligent Laboratory Systems 2002;64:15-25.
33. Zhang GP. Neural networks for classification: a survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 2000;30:451-462.
34. Fraley C. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal 1998;41:578-588.
35. Bishop C. Pattern Recognition and Machine Learning. (Eds.) Springer 2006
36. Liaw A, Wiener M. Package 'randomForest'. 2015;
37. Culp M, Johnson K, Michailidis G. Package "ada". 2016;
38. Karatzoglou A, Smola A, Hornik K. Package "kernlab". 2016;
39. Ripley B. Package 'nnet'. 2016;
40. Scrucca L. Package 'mclust'. 2017;
41. Efron B, Tibshirani R. An introduction to the bootstrap. 1st CRCPress reprint (Eds.) Boca Raton: Chapman & Hall/CRC 1998 237-81.
42. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21:3301-7.
43. Yang Y. Consistency of cross validation for comparing regression procedures. The Annals of Statistics 2007;35:2450-2473.

44. Rutkowska E. Parameter Estimation. Interdisciplinary ESTRO pre-meeting course: "Statistics for radiotherapy data" 19th April 2013;Geneva, Switzerland:

45. Haley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiol 1982;143:29-36.

46. Fawcett T. ROC graphs: notes and practical considerations for data mining researchers. Technical report hpl-2003-4 (Eds.) HP Laboratories 2003 Palo Alto, CA, USA.

47. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pat Recog 1997;30:1145-59.

48. Beetz I, Schilstra C, Burlage FR et al. Development of NTCP models for head and neck cancer patients treated with three-dimensional conformal radiotherapy for xerostomia and sticky saliva: the role of dosimetric and clinical factors. Radiother Oncol 2012;105:86-93.

49. Lambin P, van Stiphout RG, Starmans MH et al. Predicting outcomes in radiation oncology - multifactorial decision support systems. Nat Rev Clin Oncol 2013;10:27-40.

Table 1. Demographic, clinical and treatment information for the population cohort used in this study. The average patients age is 56.8, with a standard deviation of 11.9 (the ages are comprised between 17.8 and 84.8).

Characteristics	N (%)
Gender	
Female	20 (14)
Male	118 (86)
Stage AJCC[1]	
I	3 (2)
II	10 (7)
III	31 (23)
IV	94 (68)
Surgery	
Yes	60 (43)
No	78 (57)
Type of radiotherapy	
Non-Concomitant	64 (46)
Concomitant	74 (54)
Overall treatment time (days)	
≤44	85 (62)
>44	53 (38)
Xerostomia at baseline	
Yes	9 (7)
No	129 (93)
Total	138 (100)

[1] American Joint Committee on Cancer

Table 2. Dosimetry for the parotid glands.

Parotid Glands	Dmean±SD (Gy) (min–max)	
	Patients with xerostomia	Patients without xerostomia
Ipsilateral	39.0±7.4 (24.8 – 61.6)	37.1±8.1 (17.6 – 56.4)
Contralateral	35.2±7.4 (12.5 – 55.1)	33.1±8.3 (4.9 – 47.1)

Table 3. AUC obtained for all possible n-factor random forest predictive models (n∈{1,2,3,4}). The mean dose in the parotid glands is the average between both planned mean (physical) doses in the contralateral and ipsilateral parotids.

Features	AUC
Gender	0.50
Xerostomia at baseline	0.52
Age	0.63
Mean dose in parotid glands	**0.68**
Xerostomia at baseline and Gender	0.50
Age and Gender	0.50
Xerostomia at baseline and Age	0.59
Mean dose in parotid glands and Age	0.65
Mean dose in parotid glands and Xerostomia at baseline	0.66
Mean dose in parotid glands and Gender	**0.68**
Mean dose in parotid glands, Xerostomia at baseline and Gender	0.67
Mean dose in parotid glands, Age and Gender	**0.69**
Mean dose in parotid glands, Xerostomia at baseline and Age	**0.70**
Mean dose in parotid glands, Xerostomia at baseline, Age and Gender	**0.73**

Figure 1. ROC curve generated by the following predictive models: A. Random Forest, B: Stochastic Boosting, C. Support Vector Machine, D. Neural Network, E. Model-based Clustering, F. Logistic Regression. All models were applied to the dataset by a LOOCV technique. The diagonal line produces an AUC of 0.5. TPR and FPR are, respectively, the True and False Positive Rates.

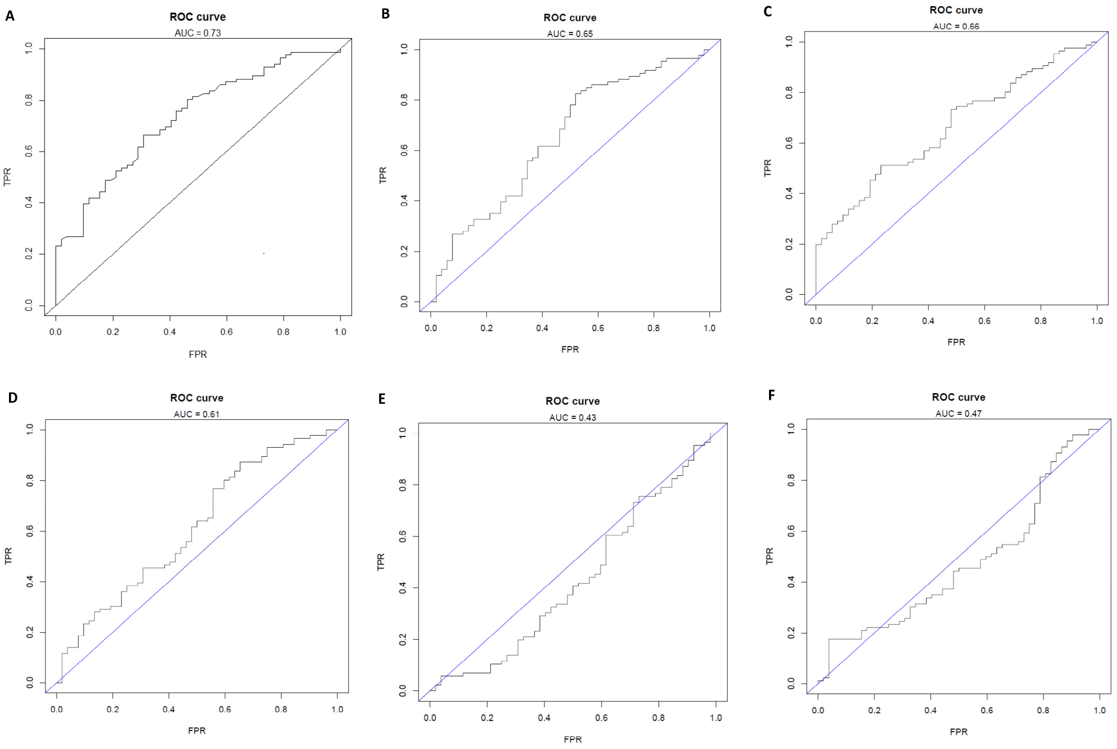

Figure 2. Histogram generated by the accuracy values, obtained for the validation dataset with 24 random samples, in the 100 iterations for the threshold value of 0.5, when considering the 4-factor random forest model. The histogram cells are left-opened and right-closed intervals.

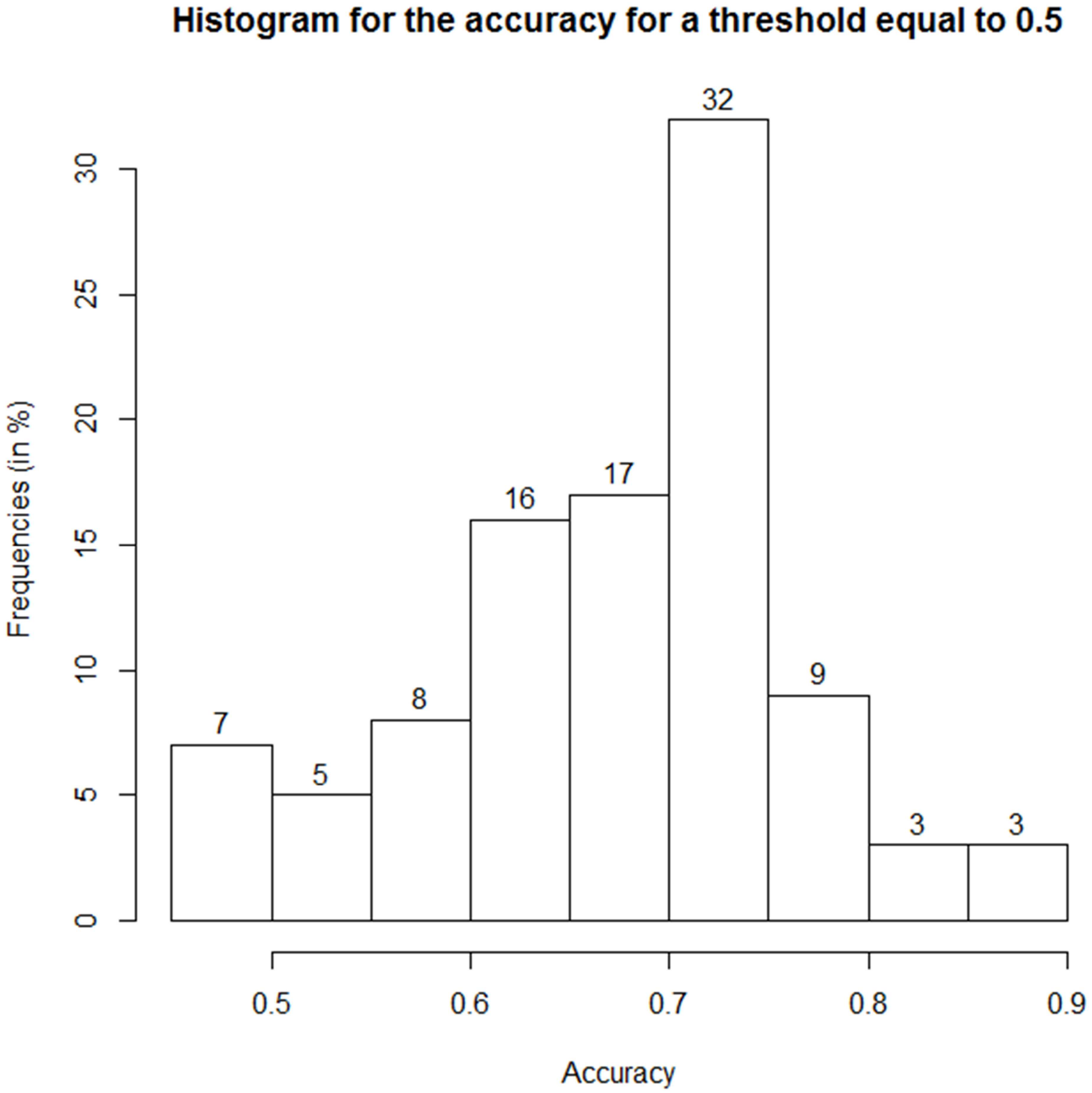

Table Legends

Table 1. Demographic, clinical and treatment information for the population cohort used in this study. The average patients age is 56.8, with a standard deviation of 11.9 (the ages are comprised between 17.8 and 84.8).

Table 2. Dosimetry for the parotid glands.

Table 3. AUC obtained for all possible n-factor random forest predictive models ($n \epsilon \{1,2,3,4\}$). The mean dose in the parotid glands is the average between both planned mean (physical) doses in the contralateral and ipsilateral parotids.