Pontificia Universidad Católica de Chile
Facultad de Física
Instituto de Física

Impact of dose and grading uncertainties on Xerostomia prediction using Machine Learning classification

by

María Constanza Hormazábal González

Thesis submitted to the Faculty of Physics
of Pontificia Universidad Católica de Chile, as one of the requirements
to qualify for the academic Master's degree in Medical Physics.

Supervisor	:	Dr. Niklas Wahl (DKFZ, Germany)
Co-supervisor	:	Dr. Araceli Gago-Arias (IDIS, Spain)
		Dr. Paola Caprile (PUC, Chile)
Committee	:	Dr. Gustavo Düring (PUC, Chile)

October, 2020
Santiago, Chile

Contents

List of Figures

List of Tables

Abstract

Due to the sharp dose gradients present in IMRT treatments, small uncertainties can generate significant differences between the planned and delivered dose distribution for head-and-neck cancer patients. This thesis investigated the impact of dosimetric and toxicity grading uncertainties on the prediction of post-radiotherapy xerostomia development.

Dosimetric uncertainties were simulated by Gaussian shifts of the planned dose for a cohort of 77 patients. We analyzed demographic, radiomic and dosimetric features as predictor of xerostomia. Univariate and multivariate studies were carried out and compared with classic logistic regression prediction models. These models were based on the nominal features extracted from the planned dose and the mean and standard deviation of features extracted from the shifted dose. In the case of grading uncertainties, we introduced a sample weight in the predictive model.

The predictive power of the models was quantified in terms of AUC. In general, no change in the AUC values was observed when considering the mean of the features. Nevertheless, when considering the standard deviation, we obtained models with higher AUC, especially in the classical model analysis, corresponding to models based on mean doses. The model based on ipsilateral and contralateral mean doses improved its performance from an AUC of 0.47 (0.44-0.50) to 0.83 (0.80-0.85). In the univariate analysis we found that the dose gradient in the patient's right-left direction was highly correlated with xerostomia development. However, this predictor becomes a bad indicator of xerostomia when considering the dose uncertainties. In the multivariate study, developing machine learning models with good performance was possible, reaching AUC values close to 0.9.

The uncertainties in xerostomia grading showed to influence the probability space generated by the predictive model. However, these results were not significant, obtained the same AUC values as those calculated without the grading uncertainties.

Chapter 1

Introduction

Since the first clinical use of radiation shortly after the discovery of X-rays by Röntgen in 1895, the field of radiotherapy (RT) has developed considerably. Along with surgery and chemotherapy, RT is one of the three main treatments against cancer. It is estimated that approximately 50% of the population of patients with cancer receives at least one treatment of radiotherapy during their lifetime, which can be alone or in combination with other cancer treatments (Delaney et al. (2005)).

Cancer is the result of uncontrolled cell growth. Cancer cells can quickly invade surrounding tissues and spread to distant organs in the body. Therefore, RT aims to kill or slow the growth of these cells by using ionizing radiation directed at the affected tissue in order to damage their DNA (Bernier et al. (2004)).

Although the goal of radiotherapy is to attack the cancerous cells, inevitably, the radiation will also affect healthy or normal tissue because there are organs very close to the target and/or in the path of the radiation beam to the tumor. This is a particularly important issue when organs in the vicinity of the target, called organs at risk (OARs), decrease all or part of their functionality due to the amount of radiation received. When this happens, it is called a side effect of RT.

The great challenge in RT's field has been to deliver the correct dose of radiation to the tumor while avoiding the damage to healthy tissue near the cancer cells. With this purpose, RT treatment delivery modalities have been developed and improved over the years. In this regard, one of the most important advances has been the advent of Intensity Modulated Radiation Therapy (IMRT) as a technique to deliver the treatment. IMRT allows the radiation beam to adapt to the three-dimensional shape of the tumor by modulating its fluence, which results in the delivery of a high dose to the target while avoiding the critical structures in its close proximity. (Taylor (2004)).

According to Argirion et al. (2019), cancer of the head and neck (H&N) region is the sixth most common cancer in the world. In the treatment of this disease, one of the OARs to

be considered are the salivary glands, especially the parotids, which are the largest salivary glands and the main source of stimulated saliva (Humphrey and Williamson (2001), Dawes and Wood (1973)). One of the side effects of RT for H&N cancer is xerostomia, which is the sensation of dryness in the mouth resulting from a reduction in salivary flow (Dirix et al. (2007)). This symptom is produced because, despite the advances in RT delivery methods, salivary glands are often irradiated due to their proximity to the tumor, affecting the production of saliva or its constitution. The dryness of the mouth can severely impact a patient's quality of life, from causing difficulties in chewing, swallowing and speaking, to tooth decay, oral infections, and tooth loss (Teng et al. (2019)).

IMRT is one of the main treatment techniques for patients with H&N cancer (Teng et al. (2019), Owosho et al. (2017)). The reproducibility and accuracy of patient setup is an important aspect to take care of for these type of treatment technique. Due to the steep dose gradients used in IMRT, little displacements in patient positioning may dramatically alter the dose distribution in the OARs, causing the development of post-radiotherapy diseases. It has been demonstrated that H&N treatments are affected by geometric uncertainties caused by setup error, posture changes, weight-loss, and tumor shrinkage (ODaniel et al. (2007), Heukelom et al. (2020)). These uncertainties lead to differences between the planned and delivered dose to the patient.

Over the years, different protocols and prediction models have been implemented to prevent and predict, respectively, the normal tissue toxicity following RT (Fiorino et al. (2009), Bentzen et al. (2010)). In the last years, the introduction of Artificial Intelligence (AI) in the medical field, most specifically the use of Machine Learning (ML) algorithms have been beneficial to predict clinical outcomes (Kang et al. (2015), Naqa et al. (2018)). This implementation has been possible because of the availability of imaging and dosimetric information in an RT treatment. Besides, the incorporation of patient characteristics (like age and sex) and the previous health history have improved the prediction of a clinical outcome making it more robust.

Patient data stored in clinics and hospitals make it possible to perform retrospective studies where the outcome is already determined. In the context of ML applications, this allows using supervised learning algorithms to predict clinical outcomes. In supervised ML, the class (or condition) of every patient is given to the algorithms to make them learn and find relationships between the given inputs and outputs (Bishop (2006)). Therefore, the correct or incorrect category of every patient, i.e., its class, could lead to a good or bad performance of the classification.

Different studies have been carried out to predict radiotherapy-induced xerostomia, using the mean dose to the parotid glands as a predictor of this condition. Some of them have shown that the mean dose is a good predictor of xerostomia (Beetz et al. (2012), Houweling et al. (2010), Nishimura et al. (2004)). Nevertheless, other studies

have demonstrated that the mean dose fails to predict this side effect, or there are other predictors with better results (Buettner et al. (2012), Lee et al. (2015)).

Gabryś et al. (2018) designed a xerostomia prediction model with features based on parotid shape (radiomics), dose shape (dosiomics), and demographic characteristics, using ML algorithms, for a cohort of 153 H&N cancer patients. They found that the incorporation of these features is beneficial for modeling this side effect. They used dosiomics features that are based on the planned dose, which, as mentioned before, is not equal to the dose received by the patient.

The aim of this thesis is to study the impact of uncertainties, related to the planned dose and to the toxicity grading on xerostomia prediction. The same cohort studied by Gabryś et al. (2018) is used in this work, and the prediction of xerostomia is performed applying ML classification.

To study the dose uncertainties, three different scenarios are implemented. The prediction of xerostomia is analyzed for univariate and multivariate models, which means considering one single feature and a group of features as predictors, respectively.

In this work, the xerostomia grade of each patient registered in each follow-up report is part of the data collected. Using this, the second part of this thesis focuses on the uncertainties in the toxicity grading system of this side effect. To carry out this study, the ML classification was performed under two conditions. First, using the average of grades registered as the final toxicity of xerostomia and second, considering the contribution of each of the grades registered in the follow-up reports to calculate the final xerostomia grade.

The next chapter presents the theoretical framework used in this thesis. It includes an explanation of the development of post-radiotherapy xerostomia and the toxicity grading of this side effect together with an introduction of machine learning classification. Characteristics of the patient cohort, endpoints, and a description of the features used as predictors of xerostomia, are given in Chapter 3. This chapter also describes the methodology applied to build the xerostomia prediction models.

Chapter 4 shows the results of this work, presenting the performance of the xerostomia prediction models considering uncertainties in the planned dose, followed by the performance of the models regarding the uncertainties in the xerostomia grading system. Chapter 5 analyzes and interprets the results of this thesis given in the previous chapter. Finally, the conclusions and an outlook on the impact of the dosimetric and grading uncertainties into xerostomia prediction are shown in Chapter 6.

Chapter 2

Theoretical framework

This work studies the impact of uncertainties in the prediction of post-radiotherapy xerostomia. This chapter begins with an introduction of radiotherapy, followed by an explanation of xerostomia and how it is related to head-and-neck radiotherapy treatment. To analize the uncertainty associated to the xerostomia grading, the toxicity grading system is also detailed. Finally, the normal tissue complication probability is presented, together with an introduction of machine learning classification models.

2.1 Radiotherapy

Radiation therapy or radiotherapy is one of the main treatments against cancer. The goal of RT is to shrink the tumor and kill cancer cells by using ionizing radiation, which works by damaging their DNA. This damage can be lethal, leading to cell death of the tumor tissue (Joiner and van der Kogel (2009)).

In this medical treatment, typically, a beam of high-energy photons in the order of mega-electronvolts (MeV) is directed from outside the patient's body into the tumor. These treatments are usually performed by a clinical linear accelerator (LINAC), which accelerates electrons to kinetic energies in the range of mega-electronvolt that impact a target and emit a fraction of its kinetic energy in the form of x-ray photons. These photons form the radiation beam that is pointed at the tumor inside the patient. When this radiation beam interacts with the patient's body, some of its energy is absorbed by the medium, in this case, the tissue of the patient. This amount can be quantified by the absorbed dose, which is defined as the energy absorbed (ΔE_{ab}) per unit mass of the medium (Δm) (Podgorsak (2010)), and is given by:

$$D = \frac{\Delta E_{ab}}{\Delta m} \tag{2.1}$$

The international system unit for absorbed dose is the Gray (Gy), defined as 1 J of absorbed energy in 1 kilogram of the medium.

Even though the goal of RT is to eliminate tumor cells, healthy tissue adjacent to the tumor is also affected because the radiation beam deposits its energy as it passes through the patient's body. Consequently, a good RT treatment plan delivers a sufficiently high dose to the tumor to kill the cancerous cells, while minimizing the dose to surrounding healthy tissue.

2.1.1 IMRT for head and neck cancer

In the early days of RT, treatments were delivered using opposing radiation fields, a modality known as Conventional Radiotherapy. This 2D RT technique used rectangular treatment fields yielded box-shaped dose distributions. However, tumors are rarely shaped this way. This limitation was most critical for tumors with concave shapes. Advances in this field, in particular the development of computer tomography (CT), overcame this limitation with the introduction of the 3D Conformal Radiotherapy technique (3DCRT), in which the shape of the radiation beams of uniform fluence is adapted to the shape of the tumor. A significant improvement from 3DCRT was achieved by the advent of Intensity Modulated Radiation Therapy (IMRT). IMRT uses an optimized modulation of the beams fluences to generate highly conformal dose distributions, which allow delivering a high dose to the macroscopic tumor or gross target volume (GTV) and better sparing of the normal tissue (Taylor (2004)) (Figure 2.1). Like Conformal RT, IMRT beams conforms the planned dose to the three-dimensional shape of the tumor, becoming very helpful in treating complex targets, either due to the shape or its location within the patient anatomy.

The irradiation of these structures can lead to severe complications for the patient (Chou et al. (2005)). The implementation of IMRT is suitable for the treatment of H&N cancer patients (Lee et al. (2007), Budgell (2002)). This technique's ability to precisely deliver the intended dose to the complex shape of the tumor while sparing most of the surrounding structures to the target allows to improve the outcome and lead to lower toxicities.

Different studies with head and neck cancer patients have shown that IMRT achieves a reduction in side effects and a better quality of life for the patients, as compared to conventional radiotherapy. In the case of xerostomia (presented in the next section), it has been shown to be prevalent in 75-80% of patients with conventional radiotherapy, while it

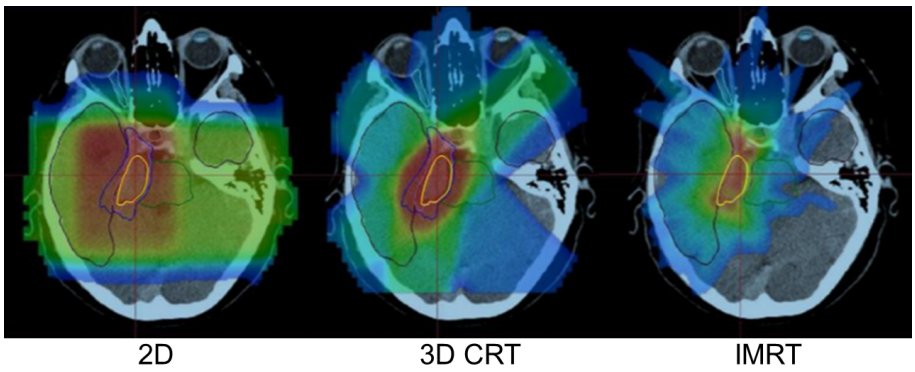


Figure 2.1: Three RT treatment delivery techniques. The first image on the left corresponds to the 2D-conventional radiation technique, the middle image to 3DCRT, and the image on the right corresponds to an IMRT plan. High-dose and low-dose regions are indicated in red and blue, respectively. It can be seen that the development of more advanced delivery techniques (from left to right) has made it possible to deliver treatments more safely in terms of healthy tissue spared. As the delivery technique progresses to IMRT, the high-dose regions are concentrated in the PTV (delimited in a yellow contour). Whereas the OARs (delineated with a black contour) and the surrounding healthy tissue receives lower radiation dose (K. Bénézery, 2014).

decreases by about 40% when using IMRT (Kam et al. (2007), Nutting et al. (2011)).

2.2 Development of post-radiotherapy xerostomia

2.2.1 Parotid glands

There are three paired major salivary glands: the parotid (in front of each ear), the submandibular (under the jaw), and the sublingual glands (under the tongue) (see Figure 2.2). The parotid glands, located on either side of the mouth and in front of both ears, are the largest salivary glands, being approximately twice the size of the submandibular glands and ten times the size of the sublingual glands. The function of the major and minor glands is to produce and secrete saliva. This substance is transported from the glands to the mouth, helping with chewing, swallowing, phonation and digestion, as well as to defend against bacteria and prevent dental caries.

Every day approximately 0.5-0.6 liters of saliva are secreted. The contribution of each gland depends on the level of stimulation. For unstimulated saliva, the flow rate is between 0.3 and 0.4 ml/min, where 25% of it comes from the parotid glands. In the case of stimulated saliva, the normal flow rate increases to 1.5-2 ml/min, and the contribution of these parotids glands is about 50% (Edgar et al. (2012)). Considering the total salivary volume, the contribution of the parotid glands is approximately 30%.

The parotid glands are important OARs to consider during H&N treatment planning. They are more exposed to the radiation field because of their large size. Furthermore, studies have suggested that serous cells, mainly in the parotid glands, are more radiosen-

sitive than mucous cells, located in the sublingual and submaxillary glands (Jaguar et al. (2017), Scrimger (2011)).

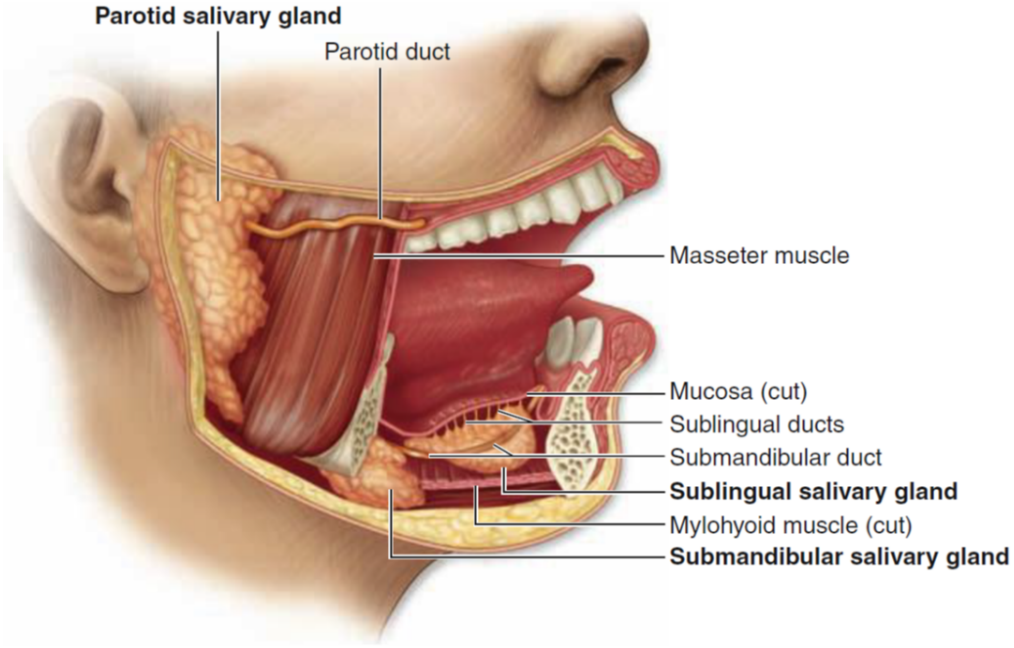


Figure 2.2: Anatomy of the right side of the mouth. Locations, relative sizes, and excretory ducts of the three major salivary glands are shown in the illustration. Image taken from Mescher (2016).

2.2.2 Xerostomia

The sensation of dry mouth, medically called xerostomia, is a condition in which the salivary glands do not function properly and either stop producing or produce insufficient saliva (Tanasiewicz et al. (2016), Hopcraft and Tan (2010)). A good production of saliva and its presence in the mouth are essential to maintaining healthy teeth, soft oral tissues, and mucosa.

Xerostomia is a critical complication for patients who suffer from it. The primary constituents of saliva are water (the main component), proteins, and electrolytes (Dirix et al. (2006)). This substance is involved in speaking, swallowing and enhance taste. It also facilitates the irrigation, lubrication, and protection of the mucous membranes of the oral cavity. Therefore, its diminution or absence can have a detrimental impact on the patient's quality of life.

The hyposalivation could have different causes, including the use of some medications, aging, autoimmune diseases, and cancer treatments such as radiotherapy and chemotherapy. In 2010, Cho et al. (2010) investigated the differences in salivary flow rates and dry

mouth symptoms in 140 patients with xerostomia caused by Sjögren's syndrome[1], post-radiation therapy in the head and neck region, antipsychotic medication, systemic diseases, or unknown cause. They found that the group of patients with a history of RT presented the most decreased values of salivary flow rate and the most severe associated symptoms.

2.2.3 Xerostomia induced by radiotherapy

Like most cancer treatments, RT can have unfortunate side effects. These effects occur as a reaction to the damage of healthy tissue near the treatment zone produced by the radiation. Symptoms often start during the second to the third week of treatment and, in some cases, may last for several weeks or months after the therapy (Jaguar et al. (2017), Deasy et al. (2010)). RT side effects are different for each person, depending on the type of cancer, location of the tumor, patient's health condition, type of RT, and the received dose, among others. Side effects can be classified into two categories: early side effects and late side effects. The first category tends to be mild and treatable, takes place during or shortly after treatment, and usually disappears within a few weeks after its end. On the other hand, late side effects take months or even years to develop and can be permanent (American Cancer Society (2019)).

Xerostomia is the most common and disabling side effect during and after RT treatment for H&N cancer (Jaguar et al. (2017), Dirix et al. (2007)). Approximately 70% of patients receiving this therapy develop hyposalivation due to a progressive loss of salivary gland function (Acauan et al. (2015)), and it may persist for six months to several years after the treatment being classified as a late effect. The sensation of dry mouth is developed when the radiation damages the salivary glands in the proximity of the tumor. The severity of this damage depends on different factors, like the dose received by the healthy tissue, the volume irradiated, and the response of each individual (Deasy et al. (2010)).

In an ideal scenario, a high dose of radiation is delivered to the tumor, while the surrounding normal tissue receives a minimal dose. However, in practice, the dose to the tumor is limited by the tolerance dose to the surrounding structures. The tolerance dose establishes a limit to the severity of the side effect and corresponds to the dose of radiation at which the normal tissue can be irradiated and keep an acceptable function. The advances in RT delivery methods have made it possible to maximize the dose to the cancerous tissue and simultaneously reduce the dose to normal organs, increasing the therapeutic window. This provides the possibility of reducing toxicity while maintaining tumor control.

In H&N treatments, some strategies have been developed through the years to spare the salivary glands in order to avoid radiation-induced xerostomia without compromising the aim of the treatment (Dirix et al. (2006), Kałużny et al. (2014)). Although IMRT

[1] Autoimmune disease, in which the body attacks the glands that produce tears and saliva, causing dry mouth and dry eyes.

has become the standard technique for this type of treatment, parts of the salivary glands still receive radiation doses that cause hyposalivation and, in consequence, xerostomia. Nevertheless, the use of IMRT has shown the possibility to reduce the xerostomia toxicity, allowing better recovery of saliva flow (Kam et al. (2007), Nutting et al. (2011)).

The Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) group suggests dose constraints with the aim of reducing the toxicity of this side effect. Specifically, it states that severe xerostomia can be avoided if at least one parotid gland receives a mean dose < 20 Gy or if both parotid glands receive a mean dose < 25 Gy (Bentzen et al. (2010)). The use of the mean dose as a threshold for the NTCP is because of the parallel functional structure of the parotid glands, where small volumes of the organ called functional subunits (FSU), are organized in parallel. In these structures, if an FSU is irradiated, the organ functionality is not entirely lost, but decreasing in time. In contrast with the serial structures, for which the irradiation of an FSU disables the entire organ function.

2.2.4 Toxicity grading of xerostomia

Measuring and reporting the severity of xerostomia is not straightforward. As xerostomia is defined as a symptom, it becomes necessary to estimate the patient's subjective appreciation of oral dryness. For such reports, some questionnaires have been created, being the most commonly used the one developed by Fox et al. (1987). There are also methods to objectively diagnose and measure this side effect, including salivary flow rate measurements and imaging techniques to evaluate salivary gland dysfunction (Kałużny et al. (2014)).

The grading system used in this work to determine the grade of xerostomia of each patient from the cohort is the Common Terminology Criteria for Adverse Events (CTCAE), which is a set of criteria standardized for the classification of side effects in cancer therapy. This system defines five general grades of toxicity for all types of treatments (National Cancer Institute (2010)). Table 2.1 shows the characteristics of the general grading and the symptoms corresponding to each grade of xerostomia.

2.3 Normal tissue complication probability models

In an RT treatment, the dose escalation to the tumor is limited by the response of the surrounding normal tissue that is inevitably irradiated during the therapy. The relationship between the probability of tumor control and the likelihood of healthy tissue corresponds to the therapeutic window (Joiner and van der Kogel (2009)). Its optimization is an important step in treatment planning for evaluating therapeutic success. The concept of normal tissue

Grade	General toxicity	Xerostomia
Grade 1	Mild. Intervention not indicated.	Symptomatic (e.g., dry or thick saliva) without significant dietary alteration; unstimulated saliva flow $>$ 0.2 ml/min
Grade 2	Moderate. Noninvasive intervention indicated.	Moderate symptoms; oral intake alterations (e.g.,copious water, other lubricants, diet limited to purees and/or soft, moist foods); unstimulated saliva 0.1 to 0.2 ml/min
Grade 3	Severe. Hospitalization indicated.	Inability to adequately aliment orally; tube feeding or TPN[1] indicated; unstimulated saliva $<$ 0.1 ml/min
Grade 4	Life-threatening. Urgent intervention indicated.	-
Grade 5	Death. Death related to side effects.	-

[1] TPN: total parental nutrition, provides fluids and essential nutrients directly into the bloodstream through an intravenous catheter.

Table 2.1: Classification of toxicity grading according to the Common Terminology Criteria for Adverse Effects v4.03 for general toxicity and xerostomia (National Cancer Institute (2010)).

complication probability (NTCP) allows evaluating the risk for healthy tissue. NTCP is the probability that a given organ will present a negative response (complication) when it is irradiated with a certain dose. Therefore, NTCP models the radiation response of healthy tissue (Palma et al. (2019), Joiner and van der Kogel (2009)).

Predicting normal tissue toxicity following RT treatment becomes a challenge because different OARs, each with their own structure, functionality and sensitivity, must be considered during irradiation. In addition, patient-related factors and treatment parameters, need to be evaluated in a toxicity model. Considering the number and complexity of factors that can potentially influence the treatment outcome, algorithms that can detect patterns from certain information available in datasets have become useful to predict toxicity after treatment. In this context, ML algorithms have been widely applied to predict the toxicity following RT for different cancer treatments (Naqa et al. (2018)).

Classically, NTCP models are obtained by fitting the dose response to a sigmoidal curve to predict the toxicity of an organ or structure. With a different approach, multivariate logistic regression is commonly used when the probability model considers different variables that can affect the toxicity, where different coefficients describe the contribution of each variable in the final model (Naqa et al. (2006), Hosmer et al. (2013)).

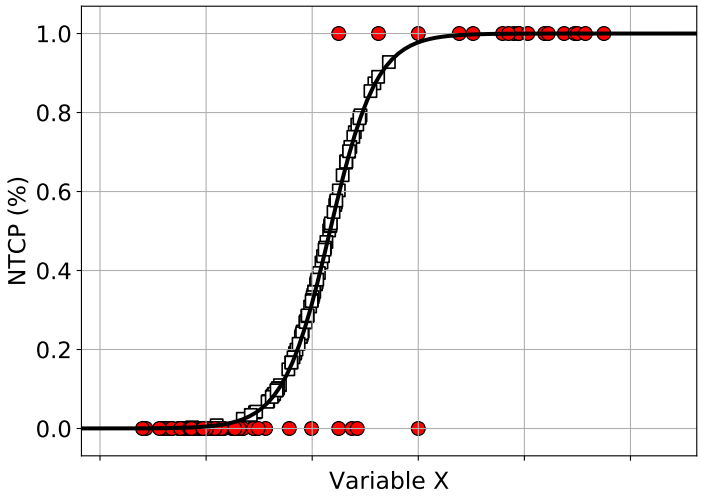


Figure 2.3: NTCP curve obtained with a univariate logistic regression model based on the predicted variable X. The red circles indicate each sample's actual probability, while the squares correspond to the probability predicted by the LR model. The black line curve indicates the NTCP model.

ML algorithms can be trained with a dataset along with the classification of each sample to obtain a model providing a probability associated to the algorithm prediction. This probability allows the development of an NTCP model for the studied toxicity. The function that models the probability of the prediction, meaning the NTCP values, corresponds to a sigmoid function, thus:

$$\begin{aligned} & \text{NTCP} = (1 - e^{-\text{S}})^{-1}, \\ \text{with} \qquad & \text{S} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \end{aligned} \tag{2.2}$$

where β_i corresponds to the coefficient of each variable x_i. Figure 2.3 shows a curve for NTCP developed using a univariate logistic regression model based on the variable X.

2.4 Machine Learning prediction model

In the Big Data era, Artificial Intelligence (AI) algorithms have been implemented in the medical area, demonstrating great potential to support clinical decisions. In the field of RT, machine learning applications have been increasing in recent years, in the fields of automatic contouring, motion management, quality control, and outcome modeling (Naqa

et al. (2018)). The last mentioned application has been widely used to predict treatment associated diseases or side effects of RT treatments. This is possible because of the ability of ML classification algorithms to detect patterns in patient information from certain datasets.

In general, the objective of ML classification algorithms is to build a model from a given training data to make predictions or decisions about unseen new data. The process of how the algorithm learns from the available data to build a model can be divided mainly into supervised and unsupervised learning. In the supervised ML technique, the algorithm learns from training data that are already labeled. Within this type of learning are the classification methods, which allow to predict the class of a given sample.

2.4.1 Classification algorithms

There are various classification algorithms available, some of them based on linear and others on non-linear classification methods. The type of ML algorithms used depends on the nature of the available dataset and the problem to be solved. The algorithms differ in the objective function or cost function that is optimized in the algorithm training step to build the desired model to correctly determine the output for input samples that were not part of the training set.

Some algorithms belonging to the category of supervised learning are presented below:

1. **Logistic regression (LR):** Logistic regression is a linear model based on the concept of probability, widely used for binary classification problems. This algorithm uses the sigmoid function or logistic function to obtain the likelihood of a given sample to belong to the positive or negative class (Tsangaratos and Ilia (2016)). A penalty term can be added to the cost function of this algorithm to make the model more robust; this approach is known as regularization. Three types of penalty terms were applied in this work: L1, L2, and elastic net (EN). The difference between these three types of regularization is the added term to the cost function. L1 regularization adds the coefficient absolute value as penalty term, whereas the L2 regularization adds the squared magnitude of the coefficient. On the other hand, EN considers a penalty term, which is a combination of L1 and L2 methods. The L1 regularization removes the less important features by weighting them as 0; however L2 allows to subtract importance to these features but does not eliminate them (Kuhn and Johnson (2013)).

2. **Support vector machines (SVM):** is a versatile algorithm capable of performing linear or non-linear classification, outlier detection, and regression. The SVM aims to find hyperplanes that segregate data into classes (Cortes and Vapnik (1995)). The strength of this classifier is that, in addition to performing linear classifications when the number of input features is two, it has the ability to perform non-linear classifications when the number of features increases. This is because SVM applies

the "kernel trick" (Yang (2019)) to find the hyperplane that better separates the data in a higher-dimensional space. However, due to this the interpretation and understanding of the model based on this algorithm become more difficult as the dimension of the feature space increases.

3. **k-Nearest neighbors (kNN):** this algorithm is a non-parametric method of classification. kNN compares the new unknown sample to be classified and the rest of the training dataset. Every sample or element is represented by feature vectors as a point in the feature space. To classify the new sample, kNN assigns it a point in this space and calculates the distance, representing the difference between the feature values, between the new sample and the nearest ones. The k elements at the shortest distance are considered for the classification. The unknown sample class is determined by the category to which the majority of the selected k elements (the k neighbors) belong (Bishop (2006)).

4. **Extra-trees (ET):** this algorithm is a type of ensemble learning technique, which uses multiple decision trees to train the algorithm with the training set and predict the classification result for a new sample. ET builds multiple trees, starting with a single node called root note. Each node represents a feature, each branch represents a decision criterion, and each leaf represents a classification result. From a random set of features, a feature selection measure is used, and the best one is selected. This feature becomes a root node and from there, the tree starts to be built using a threshold value of the feature to split the samples from the training set (Geurts et al. (2006)). The tree is built recursively using the other features for the separation of the samples. The tree is grown until the minimum sample size for splitting a node is reached. Once the algorithm is already trained, a new sample classification corresponds to the majority vote of all the trained trees classification.

5. **Gradient tree boosting (GTB):** GTB, as the algorithm mentioned above, uses an ensemble of decision trees to predict the final label. However, unlike ET, this algorithm builds one tree at a time. In each tree, misclassified samples are identified and assigned a higher weight, with the aim of having greater importance in the classification performed by the next tree. Each decision tree has a weight according to the error in its prediction. Once all the trees are completed, the final GTB prediction is calculated as the sum of the tree's weighted prediction (Friedman (2001)).

2.4.2 Model evaluation technique

To develop a prediction model based on ML classification algorithms, a dataset must be available. The algorithm will be trained with the dataset to build the predictive model.

The evaluation of the model performance is based on measuring how well it predicts the class of a new sample.

In order to avoid overfitting, the ideal is to evaluate the model on a set of data that is completely unknown to the predictive model. This means independent of the one used to train the algorithm. There are different techniques used to evaluate model performance. In the case of a large dataset, it is divided into three independent groups: a training and a test set to train the model and optimize the algorithm hyperparameters, and a validation set for evaluating the general performance of the model. In small datasets, the way the data is divided can lead to overfit and produce inaccurate results because the model was trained on a dataset that may not be representative. For those cases, sampling techniques have been developed to divide the available data, avoiding low generalization and unreliable predictions.

Cross-validation

Cross-validation (CV) is a resampling technique that allows the evaluation of ML models in a limited dataset. This method is based on dividing the data into two groups, the training set and the test set. The model is designed with the training set, where the algorithm is trained, and its parameters and hyperparameters are determined. The performance of the designed model is evaluated in the test set. The validation set mentioned above is independent of these two groups. It is used after performing the CV method to evaluate the final performance of the predictive model determined with this technique.

There are different types of CV techniques, which differ in the way and the number of times the dataset is divided. One of those is the Monte Carlo cross-validation (MCCV). In this technique, the process of splitting the data into train and test sets is repeated a desired number of iterations. In each iteration, the samples of the dataset forming each set are randomly selected. However, when the selection is performed keeping the proportion of samples of each class, it is called Stratified MCCV. In each iteration, the model is designed with the training set, and its performance is estimated in the test set using some evaluation metrics. The model's final performance is calculated as the average of the metric score obtained in each iteration.

2.4.3 Model evaluation metrics

Confusion matrix

Considering a binary classification problem, there are four possible outcomes given by the classification model. When the model correctly classifies a sample, a true positive (TP) or true negative (TN) is obtained depending on whether the actual class of the sample is positive or negative, respectively. On the other hand, if the sample is misclassified a false

negative (FN) is defined when the real class is positive and the predicted class is negative, and false positive (FP) when the real class is negative, but the model classifies the sample as positive. This can be represented in a confusion matrix (see Figure 2.4).

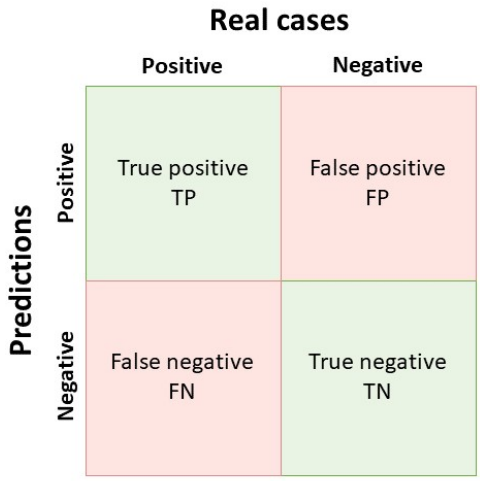


Figure 2.4: Confusion matrix, which allows visualizing the performance of an algorithm. Real cases correspond to the true condition of the samples, whereas predictions are the outcome given by the classifier.

From Figure 2.4 it is possible to observe that the total number of real positive cases (P) in the cohort corresponds to:

$$P = TP + FN \tag{2.3}$$

and the total number of real negative cases (N) corresponds to:

$$N = FP + TN \tag{2.4}$$

Two statistical metrics widely used in medicine, can be derived from the confusion matrix. These metrics indicate the performance of a binary classifier and are defined as:

- **Sensitivity:**
  Is the ability of a classifier to classify as positive the actual positive samples. This metric is also known as true positive rate (TPR) and is given by:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} \tag{2.5}$$

  where TP and P, are the true positive and the actual positive samples (eq. 2.3), respectively.

- **Specificity:**
  As opposed to sensitivity, the specificity is the ability of a classifier to classify the actual negative cases as a negative prediction. This metric corresponds to the true negative rate given by:

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} \tag{2.6}$$

  The specificity is related to the false positive rate (FPR) by:

$$\mathrm{FPR} = 1 - \mathrm{TNR} \tag{2.7}$$

Receiver Operating Characteristic curve

When working with a binary classification system, a graphic way to illustrate the performance of the classifier used is the Receiver Operating Characteristic curve (ROC curve). The ROC curve (Fig. 2.5) is a two-dimensional graph built by plotting the TPR or sensitivity on the Y-axis and the FPR or 1- specificity on the X-axis.

Each point of the ROC curve is built by using a different decision threshold to predict the samples' label. This curve not only helps to illustrate the performance of a classifier, but it is also useful for comparing two or more classifiers through a metric that represents the expected performance.

- **Area Under the ROC Curve (AUC):**

  To evaluate a classifier, it is necessary to have a metric that represents the expected performance. To compare the classification models studied in this work, the Area Under the ROC Curve or AUC (Hanley and McNeil (1982)) was used. This metric is a measure of separability that indicates how capable is a model to distinguish between two classes. An ideal classifier would have an AUC of 1.0 (upper curve in Fig. 2.5). This means that it correctly identifies and classifies all samples, while a random classifier (lower curve in Fig. 2.5) would have an AUC of 0.5. In practice, a classifier

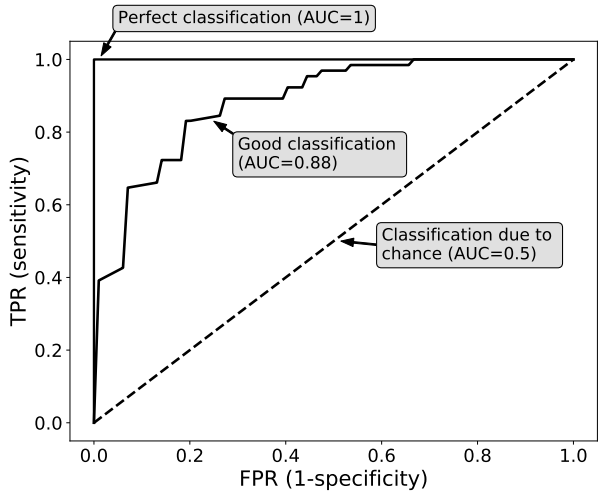


Figure 2.5: Three ROC curves representing three classifications with their respective AUC values.

generally has an AUC between 0.5 and 1.0 (middle curve in Fig. 2.5). The closer is the AUC to the latter value, the better its performance. Nevertheless, a classifier can also have an AUC between 0 and 0.5, in which case negative samples are predicted as positive class and positive samples as negative class. Thus, if a classifier achieves a score close to 0, it is almost perfectly incorrect. To obtain the samples' correct classification, the prediction obtained must be changed to the opposite class.

Chapter 3

Materials and Methods

This chapter presents the materials and methods employed in this work. The first section indicates the characteristics of the patient cohort. Likewise, the time intervals used for the categorization of the xerostomia follow-up reports are presented.

The classification of xerostomia is based on multiple dosimetric and volumetric features of and within the parotid glands. These features are extracted from RT treatment dose distribution and the patient CT image, respectively, and will be introduced in section 3.2. The methodology for building a predictive model based on ML algorithms is then described. Finally, the last two sections explain the methodology to study the uncertainties in the planned dose and the uncertainties in the toxicity grading system and its impact on the prediction model.

3.1 Patient cohort

The cohort was selected retrospectively, meaning that all patients had been already treated at Heidelberg University Hospital at the time of data collection between the years 2010-2015.

Considering that this work aims to study the development of post-radiotherapy xerostomia, the patients in the cohort should be at risk of developing this disease due to this type of treatment only. Therefore patients with xerostomia reported before therapy, replanning during the treatment, tumor in the parotid gland, second irradiation, second chemotherapy, or ion beam boost were excluded.

The final cohort consisted of 153 head-and-neck cancer patients, between 29 to 82 years old, mostly men (76%), treated with either tomotherapy or IMRT. For every patient, the tumor was located in different organs close to the parotid glands. As explained in chapter 2, xerostomia induced by radiotherapy occurs when the radiation damages the healthy parotid glands during H&N treatment. Due to this, patient selection started with

patients for whom the tumor was closer to the parotids: tumors located in the nasopharynx and oropharynx. The selection was then extended to patients with tumors located in organs further from these glands; hypopharynx, larynx, lips, and brain. For most patients (approximately 65%), the tumor was located in the oropharynx, while 24% has the tumor in the hypopharynx or larynx. Table 3.1 shows the characteristics of the final cohort.

Of the 153 patients in the cohort, 72% of them met the QUANTEC group's recommendation that both parotids should receive, on average, a mean dose < 25 Gy. As can be seen in Table 3.1, less than 25% of patients received a mean dose < 20 Gy in the ipsilateral gland, whereas approximately 50% of them fulfilled this constraint in the contralateral gland.

Total patients		153
Follow-up reports	0-6 months	134
	6-15 months	131
	15-24 months	77
Age	Median	61
	Q1-Q3	55-66
	Range	29-82
Sex	Female	37
	Male	116
Tumor site	Oropharynx	99
	Hypopharynx/Larynx	37
	Nasopharynx	12
	Other	5
Radiation modality	Intensity modulated radiotherapy	37
	Tomotherapy	116
Ipsilateral parotid gland mean dose	Median	24.3 Gy
	Q1-Q3	20.6-27.6 Gy
	Range	0.4-63.4 Gy
Contralateral parotid gland mean dose	Median	19.9 Gy
	Q1-Q3	15.4-23.1 Gy
	Range	0.3-30.9 Gy

Table 3.1: Characteristics of the cohort: 153 head and neck cancer patients. Follow up reports and their time intervals are defined in Section 3.1.1. Q1 and Q3 correspond to the first and third quartiles, respectively.

3.1.1 Follow-up reports and endpoints

One of the criteria in the process of patient selection was that at least one report confirming or denying xerostomia had to be available for every patient. Considering this, some patients had only one follow-up report, while others had more than one between 0

and 24 months after treatment. This is because the frequency of the follow-up reports was not the same for every patient. A total of 693 xerostomia toxicity follow-up reports were collected for the cohort of the 153 patients.

Given the retrospective nature of this study and the fact that the dates on which follow-up reports were recorded varied among the patients, Gabryś et al. (2018) defined three different time intervals to organize the available follow-up reports:

- **Early xerostomia:** follow-up reports between 0 and 6 months.
- **Late xerostomia:** follow-up reports between 6 and 15 months.
- **Long-term xerostomia:** follow-up reports between 15 and 24 months.

Figure 3.1 shows the number of follow-up reports recorded depending on the time at which the evaluation was performed after the RT treatment.

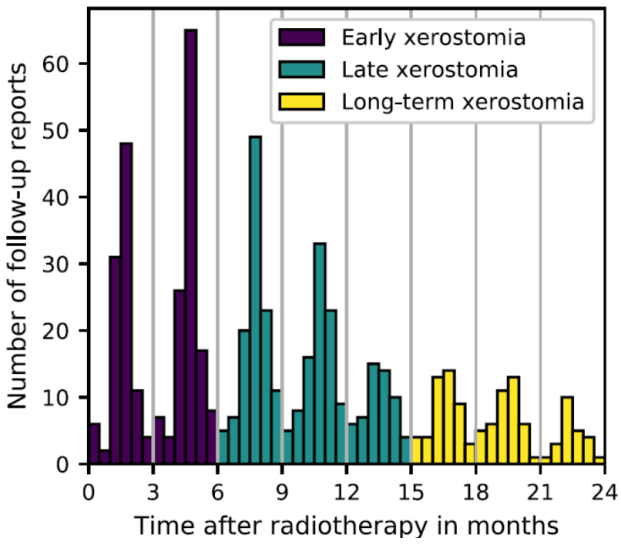


Figure 3.1: Distribution of follow-up reports in time after treatment. Figure taken from Gabryś et al. (2018).

Each follow-up report indicates the patient's grade of toxicity, which can be Grade 0, Grade 1, or Grade 2. This thesis focused on xerostomia grade 2 or higher, defined as G2+, corresponding to moderate-to-severe xerostomia, according to CTCAE (see Table 2.1). As stated above, each patient in the cohort must have at least one follow-up report between 0-24 months after RT, which means that for some patients, more than one grade of toxicity was recorded for the same time interval (early, late or long-term interval). In that case, the final score of xerostomia for the corresponding time interval is determined by the mean of toxicity grades within the interval. Depending on whether this mean value is greater or less than 1.5, the endpoint was selected accordingly.

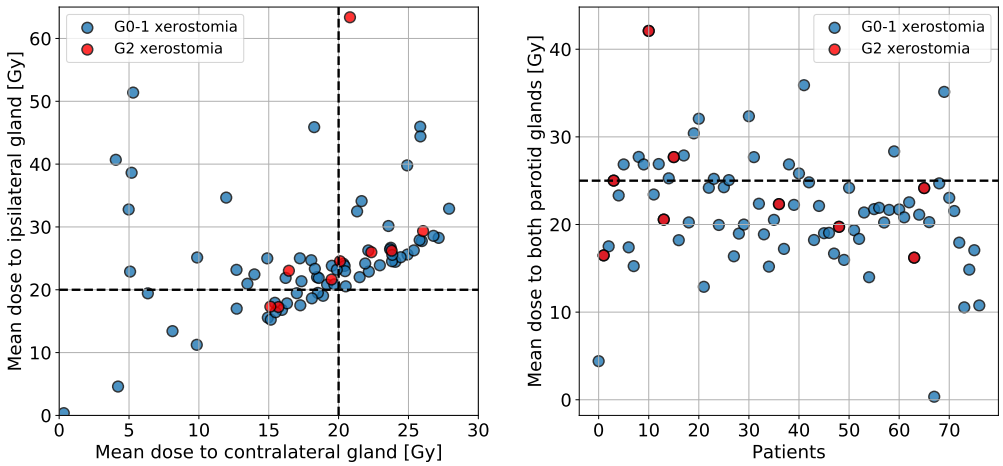
3.1.2 Patient cohort: long-term xerostomia

Gabryś et al. (2018) analyzed the prediction of xerostomia over the three time intervals mentioned above, obtaining the best result for patients with follow-up reports between 15 and 24 months. Taking this into consideration, this work focuses on long-term xerostomia, meaning the subgroup of patients from the final cohort who were evaluated between 15 and 24 months after the RT treatment.

Table 3.2 presents the characteristics of this subgroup for long-term xerostomia consisting of 77 patients, of which 9 of them presented grade G2+. Figure 3.2 shows the mean dose distribution for each parotid gland and the average of the mean dose received by both parotids. It is possible to mention that even following the recommendations of the QUANTEC group for the mean dose received by the parotids, some patients developed post-radiotherapy xerostomia of grade G2+ between 15 and 24 months after treatment. Specifically, seven of the patients whose parotids were irradiated with a dose lower than 25 Gy developed G2 + xerostomia. Whereas two patients and four patients of those who met the QUANTEC recommendation for the ipsilateral and contralateral parotid, respectively, developed G2 + xerostomia. This leads to the challenge of exploring new predictors of this side effect to develop a more precise prediction model than the NTCP model based on the mean radiation dose to the parotid glands.

Grade		**G0**	**G1**	**G2**
Total patients		15	53	9
Age	Median	61	61	61
	Q1-Q3	55–68	52-66	54-68
	Range	47-80	39-78	41-80
Sex	Female	2	9	4
	Male	13	44	5
Tumor site	Hypopharynx/larynx	3	15	0
	Nasopharynx	0	5	0
	Oropharynx	11	32	9
	Other	1	1	0
Radiation modality	IMRT	2	18	1
	Tomotherapy	13	35	8
Ipsilateral parotid gland mean dose [Gy]	Median	22.9	23.8	24.5
	Q1-Q3	18.5-31.5	20.8-26.4	21.6-26.2
	Range	0.4-51.4	4.6-46.0	17.3-63.4
Contralateral parotid gland mean dose [Gy]	Median	12.7	19.7	20.1
	Q1-Q3	5.2-17.9	16.3-23.7	16.4-22.3
	Range	0.3-27.9	4.1-27.2	15.1-26.0

Table 3.2: Patient characteristics for long-term xerostomia.



(a) (b)

Figure 3.2: Distribution of the mean dose received by each parotid (a) and by both parotids (b), for patients in the long-term time interval. The black line **(- - -)** in each graph indicates the dose threshold of the recommendations by the QUANTEC group.

3.2 Features

In an ML problem, features are measurable properties of the studied samples. These features allow the algorithms to be able to recognize patterns in a given dataset (training data), allowing them to classify samples from a new, unseen dataset (testing data). Since the features are extracted from the samples or objects studied, the types of features that will be given to the classification algorithms will depend on the nature of these samples.

By focusing on post-radiotherapy xerostomia prediction, the features used are based on the patient's characteristics, which can be found in his medical record, as well as his treatment plan and dose distribution. All this information is stored in DICOM (Digital Imaging Communication in Medicine) format for each patient in the dataset. DICOM is the standard method for storing, processing, and transferring medical imaging data (Law and Liu (2009)). The DICOM files of each patient include the organs or structures defined during the planning in the DICOM RT Structure, the distribution of the planned dose stored as a three-dimensional array in the DICOM RT Dose, and information about the patient anatomy in the DICOM CT images. Figure 3.3 illustrates the CT image and planned dose distribution of a patient from the cohort.

The DICOM files were imported to MATLAB using a publicly available toolbox[2]. With this toolbox, the CT and dose cube were extracted and interpolated to an isotropic resolution of 1 mm. For every structure defined in the DICOM RT Structure, a logical mask was generated. Each mask indicates the voxels belonging to each structure. This

[2]https://github.com/hubertgabrys/DicomToolboxMatlab

enables the extraction of radiomic and dosiomic features from the DICOM files for each patient. Radiomic features describe the parotids' shape, while dosiomic features relate to the dose distribution within the parotids.

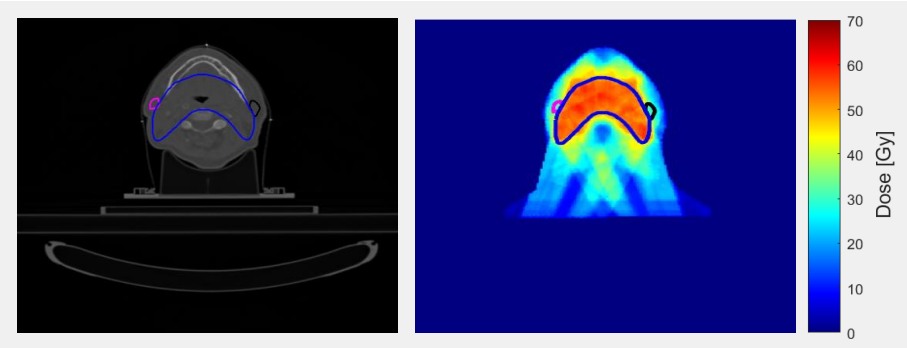


Figure 3.3: CT image (left) and 2D distribution of planned dose (right) from a patient's DICOM files from the dataset. In both images, it is possible to recognize the PTV delimited in blue and both parotid glands: right and left delimited in black and pink, respectively.

3.2.1 Feature definitions

Gabryś et al. (2018) initially investigated a space of 61 features from which a large subset exhibited a substantial correlation with each other. After performing the Kendall rank correlation coefficient, they condensed the feature space to a final set of 28 features. In this thesis we studied the final features set as potential predictors of post-radiotherapy xerostomia. These 28 indicators can be categorized into five groups listed in Table 3.3.

The features indicated in Table 3.3 are described below.

Demographics

The demographics features group comprises two features extracted from the medical report of every patient. These features are age and sex.

Etiological studies revealed that xerostomia mostly affects patients older than 65 years, due to physiological aging processes (such as the reduction in the number of secretory cells in the salivary glands) and the high probability of consumption of different medications (Cho et al. (2010)). On the other hand, menopausal women are vulnerable to xerostomia because of the deficiency of some hormones (Tanasiewicz et al. (2016)). Although patients with basal xerostomia were not considered in the cohort, these factors could be good predictors.

Feature group	Features
Demographics	age
	sex
Parotid shape	volume (V)
	sphericity (Ψ)
	eccentricity (ϵ)
Dose-volume histogram	mean ($\bar{\mathrm{D}}$)
	minimum ($\mathrm{D}_{\mathrm{min}}$)
	maximum ($\mathrm{D}_{\mathrm{max}}$)
	spread (D_{σ})
	skewness (D_{γ})
Dose gradient	gradient x (∇_x)
	gradient y (∇_y)
	gradient z (∇_z)
Three-dimensional dose moments	dose variance ($\eta_{200}, \eta_{020}, \eta_{002}$)
	dose covariance ($\eta_{110}, \eta_{101}, \eta_{011}$)
	dose skewness ($\eta_{300}, \eta_{030}, \eta_{003}$)
	dose coskewness ($\eta_{210}, \eta_{201}, \eta_{120}, \eta_{021}, \eta_{012}, \eta_{102}$)

Table 3.3: Features analyzed in this work as predictors of xerostomia.

Parotid shape

The parotid shape features characterize the size and shape of the parotid glands. This group consists of three features described below:

1. **Volume:**
   The initial volume corresponds to the number of voxels occupied by the parotid. Mathematically it can be expressed as:

$$\mathrm{V} = \sum_{x,y,z} I(x,y,z) \tag{3.1}$$

   where $I(x,y,z)$ is a three-dimensional mask with values 0 and 1. The value is 0 if the voxel (x,y,z) is not part of the parotid and is 1 if the voxel (x,y,z) is occupy by the parotid. Since all the patient's data has the same resolution, the volume of the gland is proportional to the number of voxels included in the parotid.

2. **Sphericity:**
   Is defined as the ratio of the surface area of a sphere with the same volume (V) as the parotid to the actual surface area (A) of the parotid, expressed as:

$$\Psi = \frac{\sqrt[3]{\pi}(6\mathrm{V})^{\frac{2}{3}}}{\mathrm{A}} \tag{3.2}$$

the surface area, A, is also a feature extracted with the toolbox and corresponds to the voxels of the parotid contour in the three dimensions.

3. **Eccentricity:**
   The eccentricity indicates the degree of deviation of the parotid gland from a circle. Mathematically it is given by:

$$\epsilon = 1 - \sqrt{\frac{\lambda_{min}}{\lambda_{max}}} \tag{3.3}$$

   where λ_{min} and λ_{max} are the eigenvalues of the parotid shape covariance matrix corresponding to the dimensions of the parotid gland along the principal axes defined by the eigenvectors.

Dose-volume histogram

The dosimetric features extracted from the differential DVH of the parotid gland are the following features:

1. **Mean:**
   This feature corresponds to the mean dose within the parotid and is defined as:

$$\bar{\mathrm{D}} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{3.4}$$

   where N corresponds to the total number of parotid gland voxels and d_i represents the dose in the voxel i.

2. **Minimum:**
   Corresponds to the minimum dose within the volume of the parotid and is defined as:

$$\mathrm{D_{min}} = min\left\{d_1, d_2, ..., d_i, ..., d_N\right\} \tag{3.5}$$

   where N corresponds to the total number of parotid gland voxels.

3. **Maximum:**
   Corresponds to the maximum dose within the volume of the parotid and is defined as:

$$\mathrm{D_{max}} = max\left\{d_1, d_2, ..., d_i, ..., d_N\right\} \tag{3.6}$$

4. **Spread:**
   The spread feature is calculated as the standard deviation of the dose in the parotid gland, mathematically:

$$\mathrm{D}_\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|d_i - \bar{D}|^2} \tag{3.7}$$

   where $\bar{\mathrm{D}}$ indicates the mean dose given by equation 3.4.

5. **Skewness:**
   The skewness describes the asymmetry of the DVH and is given by:

$$\mathrm{D}_\gamma = \frac{\frac{1}{N}\sum_{i=1}^{N}(d_i - \bar{D})^3}{(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \bar{D})^2})^3} \tag{3.8}$$

   If the skewness is negative, the DVH is skewed toward a low-dose region, whereas if the skewness is positive, the DVH is skewed toward a high-dose region.

Dose gradients

This group of features represents the average change of the dose along the three patient axes: *x, y* and *z*. The gradient for each axis is defined as:

1. **Gradient along *x* axis:**

$$\nabla_x = \frac{\sum_{x,y,z} D(x+1,y,z)I(x+1,y,z) - D(x-1,y,z)I(x-1,y,z)}{2\sum_{x,y,z} I(x,y,z)} \tag{3.9}$$

2. **Gradient along *y* axis:**

$$\nabla_y = \frac{\sum_{x,y,z} D(x,y+1,z)I(x,y+1,z) - D(x,y-1,z)I(x,y-1,z)}{2\sum_{x,y,z} I(x,y,z)} \tag{3.10}$$

3. **Gradient along *z* axis:**

$$\nabla_z = \frac{\sum_{x,y,z} D(x,y,z+1)I(x,y,z+1) - D(x,y,z-1)I(x,y,z-1)}{2\sum_{x,y,z} I(x,y,z)} \tag{3.11}$$

   in eq. 3.9, 3.10 and 3.11 *x, y* and *z* correspond to the voxel coordinates, *D(x,y,z)* to the dose received by the voxel, and *I(x,y,z)* was previously defined in eq. 3.1.

Three-dimensional dose moments

This group of morphological features proposed by Buettner et al. (2012) quantifies the shape of the dose distribution in all directions of the parotids, anterior-posterior, medial-lateral, and cranio-caudal.

The dose moments are defined as:

$$\eta_{pqr} = \frac{\sum_{x,y,z}(x-\bar{x})^p(y-\bar{y})^q(z-\bar{z})^r D(x,y,z)I(x,y,z)}{(\sum_{x,y,z} D(x,y,z)I(x,y,z))^{\frac{p+q+r}{3}+1}} \tag{3.12}$$

where:

$$\bar{x} = \frac{\sum_{x,y,z} xI(x,y,z)D(x,y,z)}{\sum_{x,y,z} I(x,y,z)D(x,y,z)} \tag{3.13}$$

$$\bar{y} = \frac{\sum_{x,y,z} yI(x,y,z)D(x,y,z)}{\sum_{x,y,z} I(x,y,z)D(x,y,z)} \tag{3.14}$$

$$\bar{z} = \frac{\sum_{x,y,z} zI(x,y,z)D(x,y,z)}{\sum_{x,y,z} I(x,y,z)D(x,y,z)} \tag{3.15}$$

Four subgroups of features in 3D belong to this category:

1. **Dose variance:**
   To quantify the spread of the dose in different directions, along the x, y and z axis.

   $$\eta_{200},\ \eta_{020},\ \eta_{002}$$

2. **Dose covariance**
   To quantify how the dose deposition varies along two axes, this mean along xy direction, xz direction and yz direction.

   $$\eta_{110},\ \eta_{101},\ \eta_{011}$$

3. **Dose skewness**
   To quantify the skewness of the dose distribution in a given direction. The asymmetry of the dose distribution along the x, y and z direction are given by:

   $$\eta_{300},\ \eta_{030},\ \eta_{003}$$

4. **Dose coskewness**
   To quantify how the variance of the dose distribution along one direction covaries with another direction.

   $$\eta_{210},\ \eta_{201},\ \eta_{120},\ \eta_{021},\ \eta_{012},\ \eta_{102}$$

3.2.2 Feature extraction

The demographic features were obtained from the patient's clinical record, whereas the radiomic and dosimetric features from each patient's DICOM files. The radiomics features were extracted from the CT and the dosiomics features from the dose stored in the DICOM files. The extraction of these features was performed with the same toolbox used for preprocessing the DICOM files, which allows the extraction of the different features of both the ipsilateral and the contralateral parotid gland.

3.3 Machine learning prediction model

In this thesis, we worked with the supervised learning technique with the aim of predicting the development of post-radiotherapy xerostomia through the use of classification algorithms.

After data preprocessing and feature extraction, features that will be considered in the ML model building must be selected. Section 3.4 indicates the different combinations of features considered as predictors.

Working with features of different nature implies different ranges in their values. For example, for the ipsilateral parotid gland, the mean dose has values between 0.4-63.4 Gy, while the volume ranges from 3390 mm^3 to 50867 mm^3. Since an ML algorithm works on numbers without considering its meaning or units, it is important to transform the features to a common scale. This step allows the features to be in a comparable order of magnitude before building the model, avoiding that one predictor would have more importance than another just because of the magnitude. The feature scaling is a common requirement for many classification algorithms, especially those that calculate the distance between two features or algorithms based on decision trees. The technique applied in this thesis is the standardization or Z-score normalization, through which the standardization of an x-sample is given by:

$$z = \frac{x - \mu}{\sigma} \tag{3.16}$$

where z and x are the feature value of the standardized and unstandardized sample, respectively. μ and σ correspond to the mean and standard deviation of the feature value considering all samples.

3.3.1 Model building

To build the predictive model of post-radiotherapy xerostomia, the code developed in Python by Gabryś (2018) was modified and adapted for this work. The code uses the

Scikit-learn[3] library, an open-source library for ML in Python. The library allows working with classification algorithms that are already implemented and to apply different techniques to evaluate their performance for the dataset. Additional open-source libraries were also used for visualization and data handling such as Pandas, NumPy, and Matplotlib.

In this thesis, seven classification methods were studied, using the five algorithms presented in Section 2.4.1:

- Logistic regression with L1 penalty (LR-L1)
- Logistic regression with L2 penalty (LR-L2)
- Logistic regression with elastic net penalty (LR-EN)
- k-Nearest neighbors (kNN)
- Extra-trees (ET)
- Support vector machines (SVM)
- Gradient tree boosting (GTB)

Due to the fact that in this thesis we worked with different classification algorithms, for a certain set of features, seven prediction models were built.

To build a prediction model, the ML algorithm uses its hyperparameters to train a model. The hyperparameters determine various characteristics of the models, such as the number of neighbors (k) in the kNN, and help in the learning process (Kuhn and Johnson (2013)). Knowing which values to use for each algorithm's hyperparameters is not an easy task because it depends on the chosen algorithm and the dataset used to work. Table 3.4 shows the hyperparameter values considered for every algorithm.

The first step to build the predictive model is to split the dataset into a training set and a test set. Since we worked with a small cohort, the nested-cross-validation method was applied to the dataset to train the model and then evaluate its performance.

Figure 3.4 shows the diagram of the nested-cross-validation method. In simple words, it consists in a cross-validation within a cross-validation (Cawley and Talbot (2010)), which corresponds to MCCV. In the inner loop (with 70 iterations), a process called model tuning is performed in order to train the model, and in the outer loop (with 100 iterations) the model testing is performed to evaluate the trained model. These two processes are explained below:

- **Model tuning:**

  This process aims to find a set of optimal hyperparameters that maximize the model performance, for each algorithm. There are different techniques to carry out this process. In this work, the random search optimization was used. This technique

[3]https://scikit-learn.org/

Algorithm	Hyperparameter values
LR-L1	class_weight: {None, balanced}
	C: $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, ..., 2^{10}\}$
LR-L2	class_weight: {None, balanced}
	C: $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, ..., 2^{10}\}$
LR-EN	class_weight: {None, balanced}
	alpha: $\{2^{-10}, 2^{-9.985}, 2^{9.97}, ..., 2^{5}\}$
	l1_ratio: {0,1}
kNN	n_neighbors: {1, 2, 3, ..., 9}
	p: {1, 2, inf}
SVM	class_weight: {None, balanced}
	C: $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, ..., 2^{10}\}$
	gamma: $\{2^{-15}, 2^{-14.982}, 2^{-14.964}, ..., 2^{3}\}$
ET	n_estimators: [90, 230]
	class_weight: {None, balanced}
	criterion: {gini, entropy}
	max_features: {0.05, 0.10, 0.15, ..., 1.0}
	min_samples_split: {2, 3, 4, ..., 20}
	min_samples_leaf: {1, 2, 3, ..., 20}
GTB	n_estimators: [200, 2000]
	learning_rate: $\{2^{-7}, 2^{-6.994}, 2^{-6.988}, ..., 2^{-1}\}$
	max_depth: {1 2, 3, ..., 6}
	gamma: {0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0}
	min_child_weight: {1, 3, 5, 7}
	subsample: {0.6, 0.65, 0.70, ..., 1.0}
	reg_lambda: [0, 1]
	reg_alpha: [0, 1]

Table 3.4: Hyperparameter values used to tune the classification algorithm. These values correspond to the same values used in the original code for the model building.

consists of selecting 50 random samples from the space of hyperparameter values for the corresponding algorithm (Table 3.4). With each of these samples, the classifier is trained with the train set, and the model performance is evaluated in the test set by the AUC value. For the AUC, the 95% confidence interval (CI) was calculated with bias-corrected and accelerated (BCa) bootstrap. Then, for the corresponding CV, the optimal hyperparameters will be the set of the model with the highest AUC calculated.

- **Model testing:**

The aim of the model testing process is to evaluate the model performance with the set of hyperparameters found in model tuning. This process is performed in the external nested-cross-validation loop. In each iteration, the model determined in the

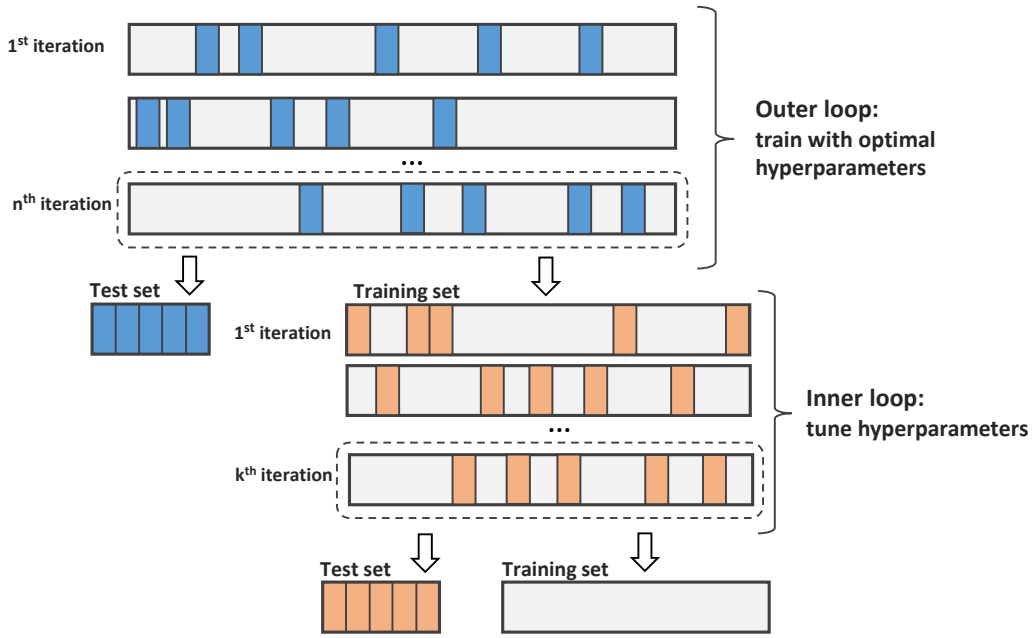


Figure 3.4: Nested-cross-validation diagram. The split of the database into a test and a training sets in both loops is carried out by MCCV in each iteration. The hyperparameters are optimized in the inner loop through the model tuning process to find the optimal set. In the outer loop, the model's performance with the set of hyperparameters found in model tuning is estimated.

corresponding model tuning is trained in the training set and evaluated in the test set, calculating the value of the AUC and its CI.

Finally, the selection of the final predictive model corresponds to the model with the set of hyperparameters based on the algorithm that maximized the AUC value in the model testing.

3.4 Uncertainties in planned dose

One of the main objectives of this thesis is to investigate the impact of planned dose uncertainties on predicting xerostomia development post-radiotherapy.

Dose uncertainties refer to the relative difference between the planned dose and the estimated dose received in reality by the patient. In this work, the dosimetric uncertainties due to setup errors were considered. These errors correspond to the difference between the expected position (used for the treatment plan) and the actual position of the part of the patient to be irradiated, i.e., the PTV, concerning the treatment beam (Hurkmans et al. (2001)). The setup errors can be of two types: random or systematic. Random errors describe setup variations between different treatment fractions that lead to a dose coverage difference to the CTV. The systematic errors correspond to deviations between the

planned position and the average position of the patient throughout the whole treatment, causing a displacement of the dose distribution.

In the literature, it has been reported that for the H&N region, the standard deviation of the setup errors varies between 1.1 mm and 4.6 mm (Hurkmans et al. (2001), Astreinidou et al. (2005)). A tolerance of 2-3 mm is accepted for H&N cancer patients treated with IMRT (Delana et al. (2009), Astreinidou et al. (2005)). In this context, three different scenarios were studied in this work for setup errors between 1 mm and 3 mm.

As the dose received by the patient in each treatment fraction is not exactly the same as the planned dose, the features extracted from it do not completely correspond to the features of the planned dose. Performing xerostomia predictions using the features from the planned dose may lead to unrealistic results, but the actual dose distribution received by the patient is unknown for this cohort and cannot be exactly reconstructed. Nevertheless, predictive models can be studied based on the expected value of these features or their variability, with the aim of developing a predictive model of xerostomia that approximates to a real treatment scenario.

Since the idea of this first part of the thesis is to study the impact of dose uncertainties in the classification of xerostomia, we studied four scenarios: one static scenario based on the planned dose and three uncertainty scenarios based on the shifted dose for three displacements of the dose: 1 mm, 2 mm and 3 mm. We performed analysis for each of them for univariate models, classical models, and multivariate models. These scenarios are explained below:

- **Static scenario:**

  The first scenario analyzed corresponds to the case without dose uncertainties. This means that the features evaluated as potential predictors of post-radiotherapy xerostomia development were extracted directly from the planned dose stored in the dose cube of each patient's DICOM file, as explained in section 3.2.2.

- **Uncertainty scenario:**

  Since random setup errors consists of many small displacements with an arbitrary distribution, it can be assumed that the setup error tends to have a normal distribution (Mileusnic (2005)). To study the impact of dosimetric uncertainties in the prediction of xerostomia, we developed a function in Matlab to simulate random displacements of each patient along the three directions. Theses displacements were based on a normal distribution of 100 samples, for which the standard deviation corresponds to

the displacement to be studied. With this, the setup error is given by:

$$\Delta_i \sim N(\mu_i, \sigma_i^2) \tag{3.17}$$

where i corresponds to the direction x, y or z, $\mu = 0$ to the mean of the distribution, and σ to the standard deviation with three different values, one for each scenario studied: 1 mm, 2 mm and 3 mm.

Once the shifted dose distributions were obtained, the features of each sampled scenario were extracted in the same manner to the features extraction from the planned dose, yielding 100 realizations for each feature. This was carried out for every patient in the cohort.

In order to build predictive models based on the expected value and the variability of the features, we considered their mean and their standard deviation as predictors of xerostomia. For example, the final features representing the dose gradient along the x-axis, ∇_{x}, is given by the mean of the ∇_{x} features, and by the standard deviation of the ∇_{x} features, extracted from the shifted dose cubes simulated with each of the 100 samples.

3.4.1 Classical models

As previously mentioned, the mean dose has been widely used as a predictor of post-radiotherapy xerostomia development. However, it has failed to predict this disease in some cohorts. Gabryś et al. (2018) studied the predictive performance of LR models based on mean dose to the parotids for this cohort, obtaining the best performance for the model based on mean dose to the contralateral parotid gland, which corresponds to an AUC of 0.58 (0.55-0.61). In addition to these models based on the mean dose, an alternative LR model proposed by Buettner et al. (2012) was also studied. This is a multivariate LR model, called morphological, that considers the three-dimensional dose moments: ${\eta_{111}}^{\mathrm{i}}, {\eta_{002}}^{\mathrm{c}}, {\eta_{300}}^{\mathrm{c}}$ and ${\eta_{110}}^{\mathrm{i}}{\eta_{110}}^{\mathrm{c}}$ as predictors of xerostomia.

In this thesis, we studied the models mentioned above based on the nominal features (extracted from the planned dose corresponding to static scenario) and the features that consider the shifted dose (corresponding to uncertainty scenarios). We tested five different LR models based on the following features:

1. Mean dose to ipisilateral parotid gland: $\bar{\mathrm{D}}^{\mathrm{i}}$
2. Mean dose to contralateral parotid gland: $\bar{\mathrm{D}}^{\mathrm{c}}$
3. Mean dose to both parotid glands: $\bar{\mathrm{D}}^{\mathrm{b}}$
4. Mean dose to ipsilateral and to contralateral parotid gland: $\bar{\mathrm{D}}^{\mathrm{i}}$, $\bar{\mathrm{D}}^{\mathrm{c}}$
5. Three-dimensional dose moments: morphological

3.4.2 Univariate analysis

The aim of the univariate analysis is to determine the predictive power of each feature defined in section 3.2. To analyze whether considering dose uncertainties results in better predictors of xerostomia, the predictive power of the dosimetric features were studied considering the different scenarios. Therefore, the features considered were: the nominals (extracted from the planned dose), the mean and the standard deviation of the dosimetric features (both extracted from the shifted dose).

The predictive power of each feature was calculated using the Mann-Whitney U-Statistic, which is a non-parametric statistical test applied to two independent samples. This test provides the value of the statistic called U and the p-value. The U-statistic is related to the AUC value by the following expression (Bamber (1975)):

$$\mathrm{AUC} = \frac{\mathrm{U}}{n_- n_+} \tag{3.18}$$

where U corresponds to the test statistic, n_- and n_+ to the size of the negative and positive class respectively. The 95% confidence intervals of the AUC value were estimated by bias-corrected and accelerated (BCa) bootstrap.

3.4.3 Multivariate analysis

The multivariate analysis performed in this work aimed to study the predictive performance, applying ML algorithms, of features groups considering the different scenarios.

We studied four multivariate models, where the selection of features was based on previously reported results for this cohort. The predictive performance of each model was determined by considering the value of AUC obtained in the model testing.

The first group of features corresponds to the mean dose received by the ipsilateral and contralateral parotid glands. These features were explored because they are currently used as predictors of xerostomia (Bentzen et al. (2010), Beetz et al. (2012)). The difference between the study of classical models, and this analysis is that this analysis considers different algorithms to develop a predictive model.

For this cohort, Gabryś et al. (2018) reported that the volume, in particular, the ipsilateral volume was the best predictor of xerostomia and the dosimetric feature with the best performance was the contralateral gradient in the right-left direction. Their multivariate analysis found that the best model was an ET model based on the contralateral spread and the ipsilateral volume. They also presented the features selected by the predictive models based on each of the algorithms used.

Considering the reported results stated above, the second group of features consists of

the ipsilateral parotid gland volume and the gradient along the x-axis of the contralateral gland. We investigated if better results can be obtained by considering a more realistic prediction model, i.e., with dosimetric uncertainties. Since the volume is not a dosimetric feature is does not change its value when the dose is shifted. Therefore the gradient was the only feature in this model that considered uncertainties in the dose.

The third features group analyzed in this work comprises two predictors: the contralateral gland's spread and the ipsilateral volume. This group was studied with the purpose of investigating whether it is possible to obtain better results with a simpler prediction model; without features selection and sampling methods, considering not only the planned dose but also the shifted dose.

The reported features selected by the predictive model built by each algorithm allows to study the predictive power of a single model based on these features, which are: the patient's sex, eccentricity, standard deviation, volume and dose gradient along the x-axis of the contralateral gland, as well as the volume and one of the dose skewness features of the three-dimensional dose moments of the ipsilateral parotid gland.

Summarizing, the set of features studied in this thesis are:

1. Mean dose to ipsilateral and contralateral parotid gland: $\bar{\mathrm{D}}^{\mathrm{i}}, \bar{\mathrm{D}}^{\mathrm{c}}$
2. Ipsilateral volume and contralateral right-left dose gradient: $\mathrm{V}^{\mathrm{i}}, \nabla_{\mathrm{x}}^{\mathrm{c}}$
3. Ipsilateral volume and spread of the contralateral gland: $\mathrm{V}^{\mathrm{i}}, {\mathrm{D}_{\sigma}}^{\mathrm{c}}$
4. Sex, contralateral eccentricity, spread of the contralateral gland, contralateral volume and right-left dose gradient, ipsilateral volume and ipsilateral spatial dose skewness along the z-axis: $\mathrm{sex}, \epsilon^{\mathrm{c}}, {\mathrm{D}_{\sigma}}^{\mathrm{c}}, \mathrm{V}^{\mathrm{c}}, \nabla_{\mathrm{x}}^{\mathrm{c}}, \mathrm{V}^{\mathrm{i}}, {\eta_{300}}^{\mathrm{i}}$

3.5 Uncertainties in toxicity grading

When a patient presents more than one grade of xerostomia in the follow-up reports for a given time interval, the final toxicity grade is considered as the arithmetic mean of the grades reported. A patient who presents grades: 1, 1, and 2 will be labeled as negative (G1), despite having a positive G2 grade. However, we cannot be certain that this patient will not develop post-radiotherapy xerostomia, but we can consider a 67% probability that he will be negative. The variability in the grades can influence the learning process of the classification model because when it is not considered, the algorithm assumes that all patients have the same importance in the classification. Nonetheless, one may argue that patients with consistent xerostomia grade in a time interval should weigh more in the classification than patients with, for example, part-time lower grading.

In this second part of the work, we studied the impact of uncertainties in toxicity grading for the LR classical models defined in section 3.4.1. The different xerostomia grades of each patient recorded in the follow-up reports for a specific time interval were

considered. Keeping the original label of each patient, we calculated the cohort samples' weights as the probability of each patient or sample to belong to his class, as can be seen in Table 3.5.

Patient	Grades	Label	Weight
1	1, 1, 2	negative	0.67
2	1, 2	positive	0.5
3	1, 0	negative	1.0
4	2, 1, 1 ,1	negative	0.75
5	2, 2	positive	1.0

Table 3.5: Grades of xerostomia, classification according to the final toxicity grade and the weight assigned to some patients in the cohort.

To implement this approach, a vector with the weights that will be assigned to each sample is delivered as a parameter ("sample_weight") to the LR algorithm. This parameter will be considered when fitting the model to the training data. In this process, each sample will have a different contribution to the algorithm's cost function, depending on its weight. Thus, the classifier will put more emphasis on correctly classifying the samples with more weight.

Chapter 4

Results

This chapter presents the results of this work, divided into two sections. The first, describes the prediction of xerostomia considering uncertainties related to the planned dose, and the second section examines the uncertainties related to the grading. The methodology used to obtain these results was presented in Chapter 3, and their analysis and interpretation are given in the next chapter.

4.1 Prediction of xerostomia considering uncertainties in planned dose

4.1.1 Univariate models

For dosimetric features, the univariate study evaluated the performance not only of the nominal features values, but also of the mean and standard deviation of these features obtained under different dose uncertainty scenarios (see Section 3.4).

The comparison between nominal and mean features performance is presented in Figure 4.2. In general, there is a small difference between the AUC values. There is no systematic increase or decrease in the predictive performance when considering dose uncertainties. The most significant difference is in the right-left dose gradient (∇_x) for both parotids. The predictive power of the ipsilateral gradient decreases from AUC = 0.78 (0.58-0.92) with the nominal feature to AUC = 0.63 (0.51-0.75) when the mean of this feature is used. And from AUC = 0.84 (0.71-0.93) to AUC = 0.56 (0.50-0.68) for the contralateral gland, considering the scenario of 3 mm for both parotids.

Better performance is obtained with dose uncertainties in the ipsilateral η_{101}, improving from an AUC = 0.56 (0.50-0.77) calculated with the nominal feature, to an AUC = 0.62 (0.54-0.81) with the mean of the η_{101} feature for the model of 2 mm. Also for the ipsilateral η_{102}, for which an AUC = 0.57 (0.50-0.67) is obtained with the nominal feature and AUC = 0.67 (0.53-0.78) with the mean of η_{102} for 3 mm. However, none of these four models

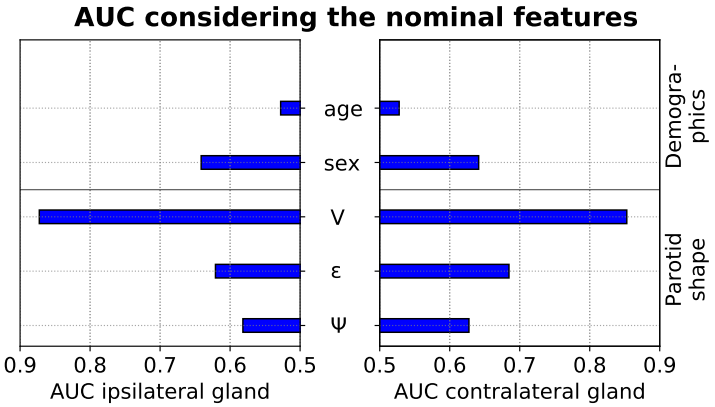


Figure 4.1: AUC for both parotid glands calculated with the Mann-Whitey U statistic for the non-dosimetric features. The central axis lists the analyzed features, and the right-hand side vertical axis lists the groups of features.

is statistically significant, with p-values greater than 0.05. For most of the features, the differences between the AUC values obtained with the different scenarios are within their uncertainty ranges, except for the contralateral ∇_x mentioned above.

Parotid volume is a strong indicator for predicting patients with G2+ xerostomia. Additionally, patients with small parotids and a higher dose gradient in the right-left direction present a higher risk of xerostomia (see Figure 4.3). Nevertheless, when considering dosimetric uncertainty scenarios, the ∇_x decreases in value and becomes less predictive as the dose shift increases. This is reflected by observing positive patients under different scenarios. In the static scenario, these patients are located in the region with the highest gradient, but as the uncertainty scenario increases, these patients' gradient value approaches the value corresponding to negative patients.

Figure 4.4 presents the comparison of the performance between the nominal features and their standard deviation. It shows a better predictive power for most of them when considering the dose uncertainties, resulting in a higher value of AUC.

The largest difference is observed for the mean dose feature ($\bar{D}$) to both parotids. The standard deviation of the $\bar{D}$ presents a better performance prediction with AUC between 0.73 and 0.83 (p-values$<$0.05), compared to the nominal feature with AUC between 0.53 and 0.58. For the ipsilateral parotid gland, the standard deviation of the $\bar{D}$ considering a 2 mm displacement predicts xerostomia with AUC = 0.77 (0.56-0.90) with p-value = 0.009, and for the contralateral gland, the standard deviation considering 3 mm presents AUC = 0.83 (0.68-0-92) with p-value = 0.001.

Although the predictive power presents a great improvement when considering the uncertainties in some features, only some models with AUC over 0.7 are obtained. The model based on the standard deviation for 3 mm of the contralateral skewness ($D_\gamma{}^c$) presents an AUC = 0.78 (0.62-0.89) and p-value=0.007, the standard deviation for 3 mm

of the contralateral η_{002} an AUC = 0.72 (0.53-0.87) and p-value=0.030, and the model based on the standard deviation for 1 mm of contralateral η_{201} give an AUC = 0.76 (0.64-0.86) and p-value=0.009.

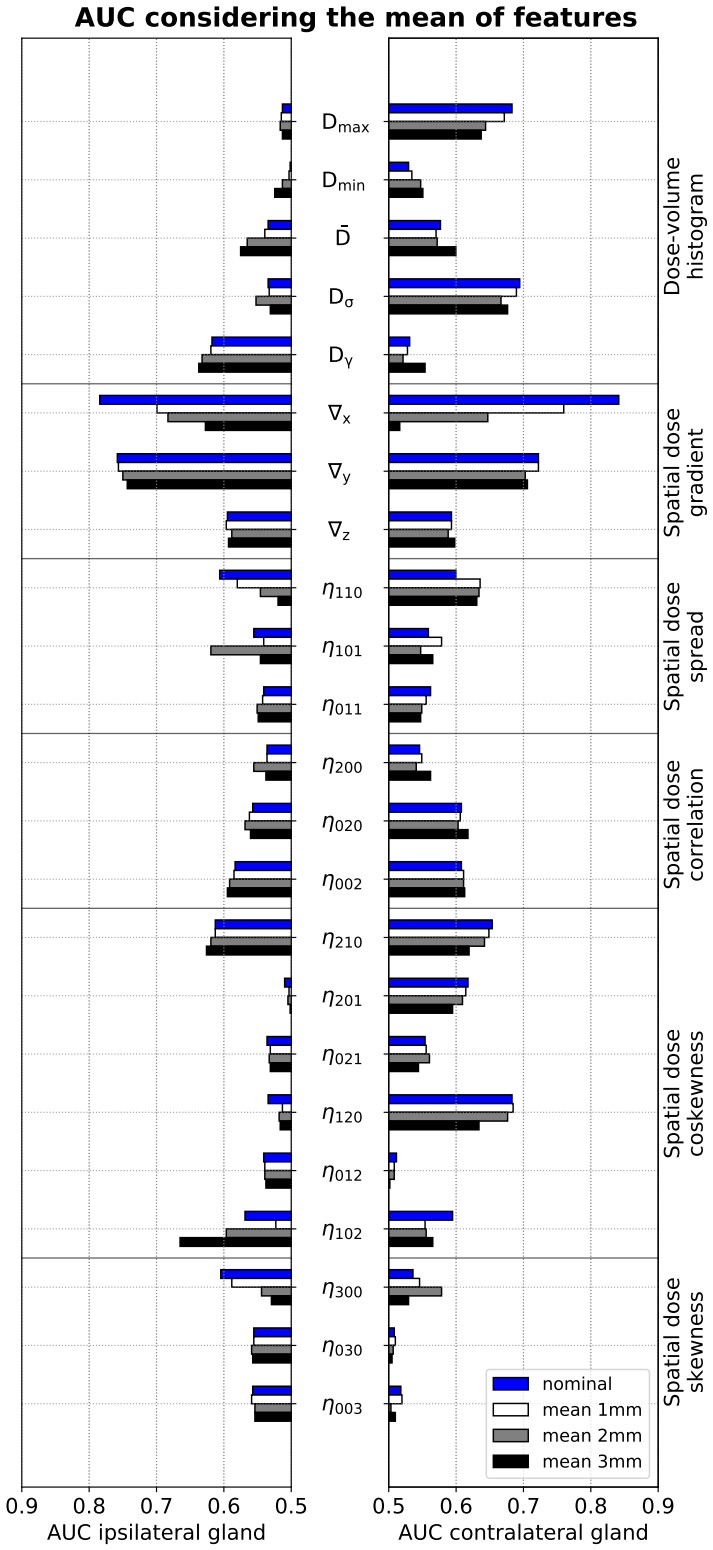


Figure 4.2: AUC for both parotid glands calculated with the Mann-Whitey U statistic for univariate models. The models consider the nominal features and the mean of the features for three uncertainty scenarios. The central axis lists the analyzed features and the right-hand side vertical axis lists the groups of features.

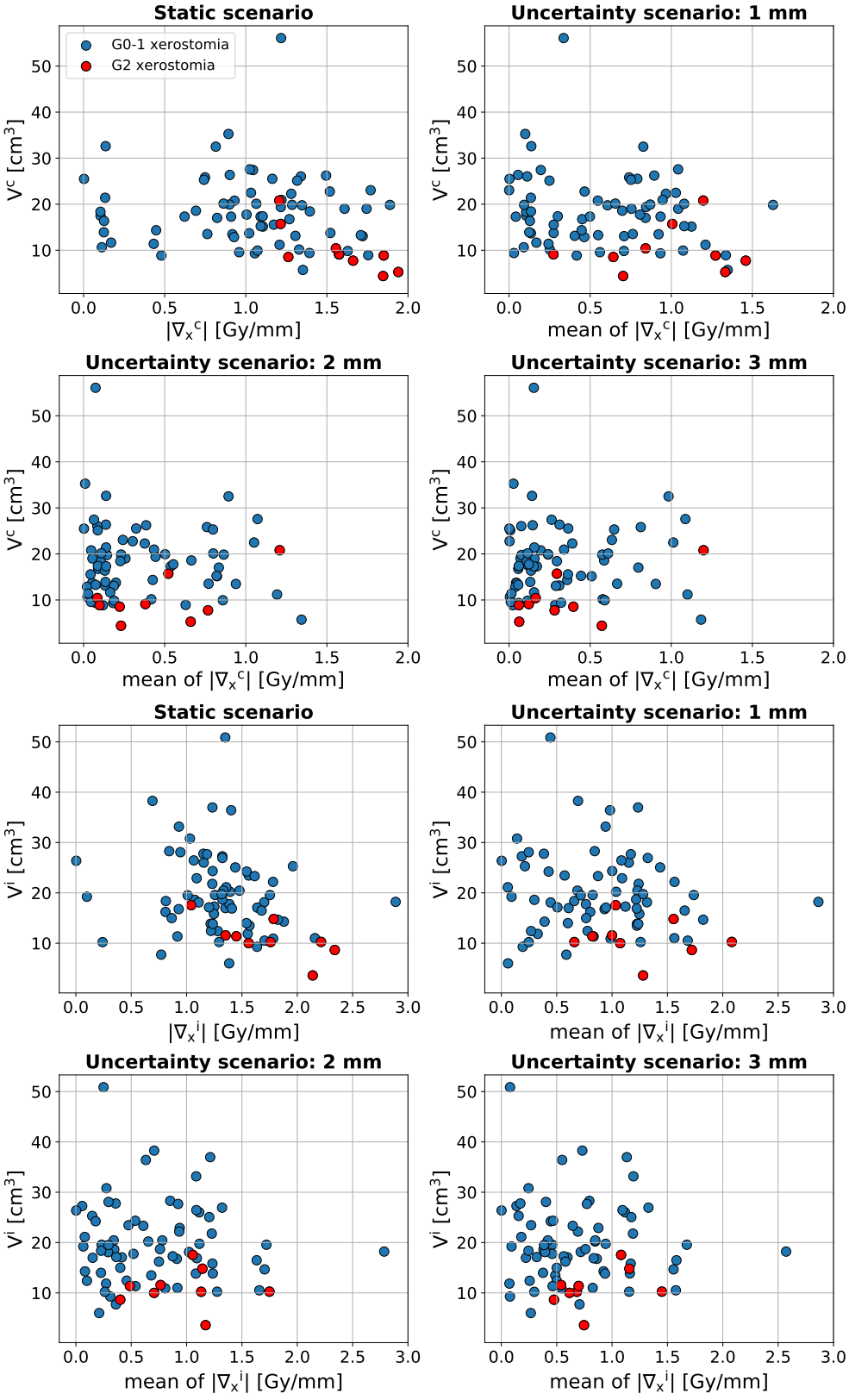


Figure 4.3: The volume and the absolute right-left dose gradient distribution for both parotids. The scenarios consider the nominal and the mean of the features for the three displacements.

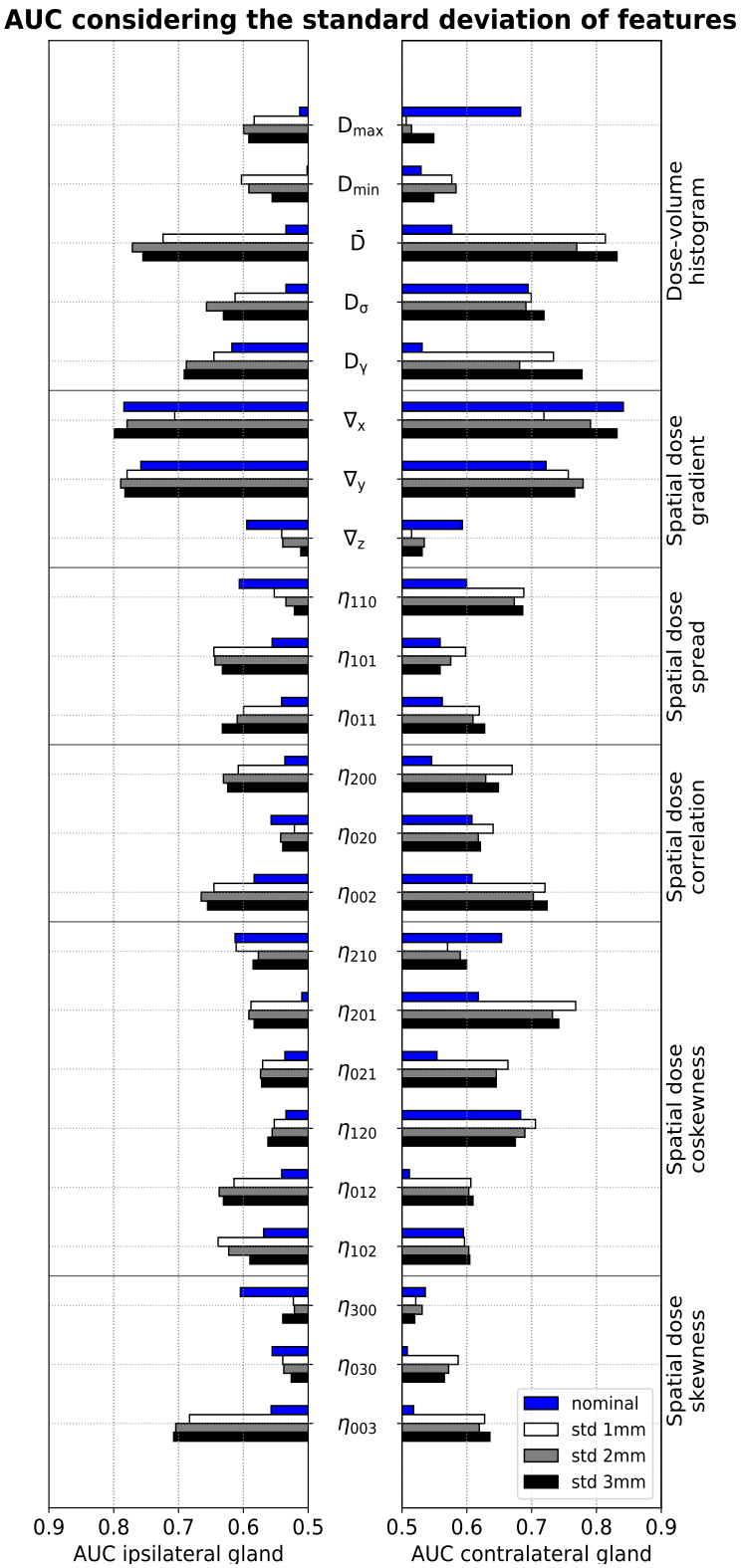


Figure 4.4: AUC for both parotid glands calculated with the Mann-Whitey U statistic for univariate models. The models consider the nominal features and the standard deviation of the features for three uncertainty scenarios. The central axis lists the analyzed features and the right-hand side vertical axis lists the groups of features.

4.1.2 Classical models

This analysis considers classic univariate LR models based on the ipsilateral mean dose ($\bar{D}^i$), contralateral mean dose ($\bar{D}^c$), and the mean dose to both parotids ($\bar{D}^b$). It also includes a bivariate model based on the ipsilateral and contralateral mean dose ($\bar{D}^i$, $\bar{D}^c$), and a multivariate morphological model (Buettner et al. (2012)).

Table 4.1 shows the AUC values with 95% confidence intervals representing the predictive power of the classical and morphological models studied. Classification models based on $\bar{D}$ features extracted directly from the planned dose fail to predict xerostomia with AUC values lower than 0.58. Using mean and standard deviation of the feature, better AUC values are obtained for both mean dose models and morphological models. This occurs especially when considering the scenario for the 3 mm displacement of the planned dose. Although the improvement in predictive power is mainly observed in models based on $\bar{D}$, it is also possible to observe it, to a lower degree, for the morphological model.

Figure 4.5 illustrates the ROC curves generated by the model based on $\bar{D}$, considering the mean and the standard deviation for the three displacements. Models that consider the mean of the $\bar{D}$ features show ROC curves similar to the performance of a random classifier. In the case of the standard deviation of the $\bar{D}$ features, these models show better results in comparison with the nominal mean doses.

Model	$\bar{D}^i$	$\bar{D}^c$	$\bar{D}^b$	$\bar{D}^i$, $\bar{D}^c$	Morphological
Nominal	0.4 (0.37-0.44)	0.58 (0.55-0.61)	0.56 (0.52-0.59)	0.47 (0.44-0.50)	0.64 (0.60-0.67)
Mean 1 mm	0.41 (0.37-0.44)	0.57 (0.54-0.60)	0.56 (0.52-0.60)	0.48 (0.44-0.51)	0.62 (0.59-0.65)
Mean 2 mm	0.43 (0.40-0.47)	0.58 (0.55-0.61)	0.58 (0.55-0.62)	0.47 (0.44-0.51)	0.56 (0.53-0.60)
Mean 3 mm	0.45 (0.41-0.48)	0.60 (0.57-0.63)	0.60 (0.57-0.64)	0.52 (0.49-0.55)	0.60 (0.56-0.63)
Std 1 mm	0.73 (0.69-0.76)	0.82 (0.80-0.84)	0.80 (0.77-0.83)	0.80 (0.77-0.83)	0.67 (0.64-0.70)
Std 2 mm	0.77 (0.74-0.80)	0.76 (0.73-0.78)	0.81 (0.78-0.83)	0.80 (0.76-0.82)	0.66 (0.63-0.70)
Std 3 mm	0.77 (0.73-0.80)	0.83 (0.81-0.85)	0.83 (0.80-0.86)	0.83 (0.80-0.85)	0.72 (0.68-0.75)

Table 4.1: AUC values with 95% confidence intervals of the predictive performance for classical and morphological LR models.

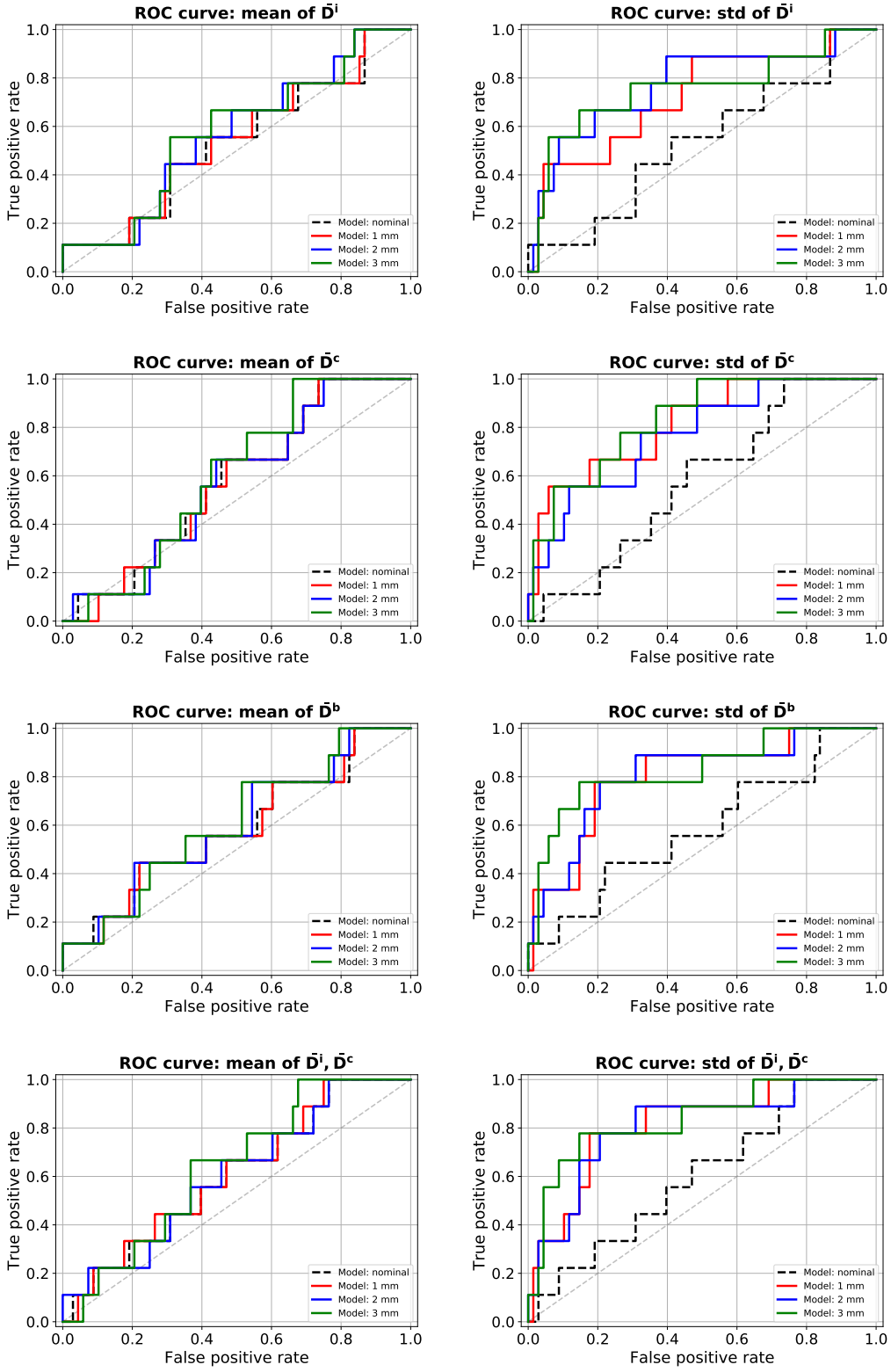


Figure 4.5: ROC curves for long-term G2+ xerostomia prediction corresponding to mean dose ($\bar{D}$) LR models, based on the nominal and the mean of the features (left) and based on the standard deviation of the features (right). The grey dotted line represents a random classifier.

With morphological models, similar performances were obtained when considering the mean and the standard deviation of the features. This can be seen in the ROC curves generated by this model in Figure 4.6.

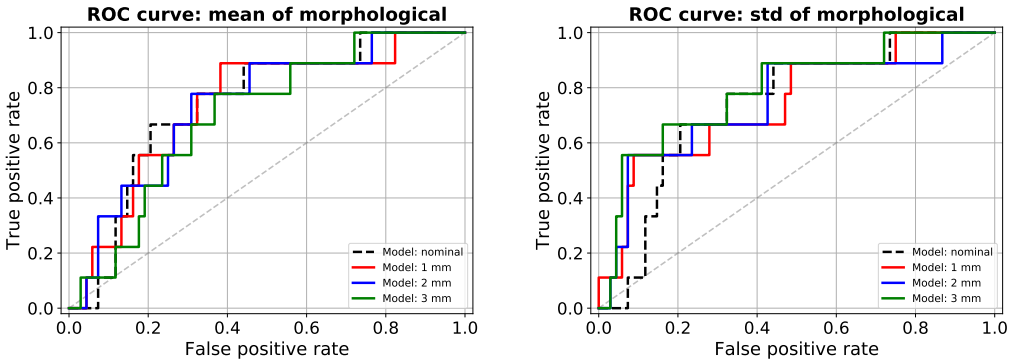


Figure 4.6: ROC curves for long-term G2+ xerostomia prediction corresponding to morphological LR models, based on the nominal and the mean of the features (left) and based on the standard deviation of the feature (right). The grey dotted line represents a random classifier.

A comparison of the NTCP curves generated by univariate LR models based on mean dose to ipsilateral, contralateral, and both parotids, is presented in Figure 4.7. It can be seen that these models give approximately linear probabilities of developing G2+ xerostomia. On the other hand, the NTCP curves of the models based on the standard deviation of the mean dose features represent more clearly the logistic approximation of the probability.

For the bivariate model based on the mean dose to ipsilateral and contralateral parotids, the AUC (see Table 4.1) increases by 76% when considering the standard deviation of the features for the 3 mm uncertainty scenario compared to the nominal features of the static scenario. Figure 4.8 illustrates the classification probability of this model for the three displacements. The separation of patients with and without G2+ xerostomia, is visually more clear in Figure 4.8.d), which corresponds to the model considering uncertainties in the planned dose for the scenario of 3 mm.

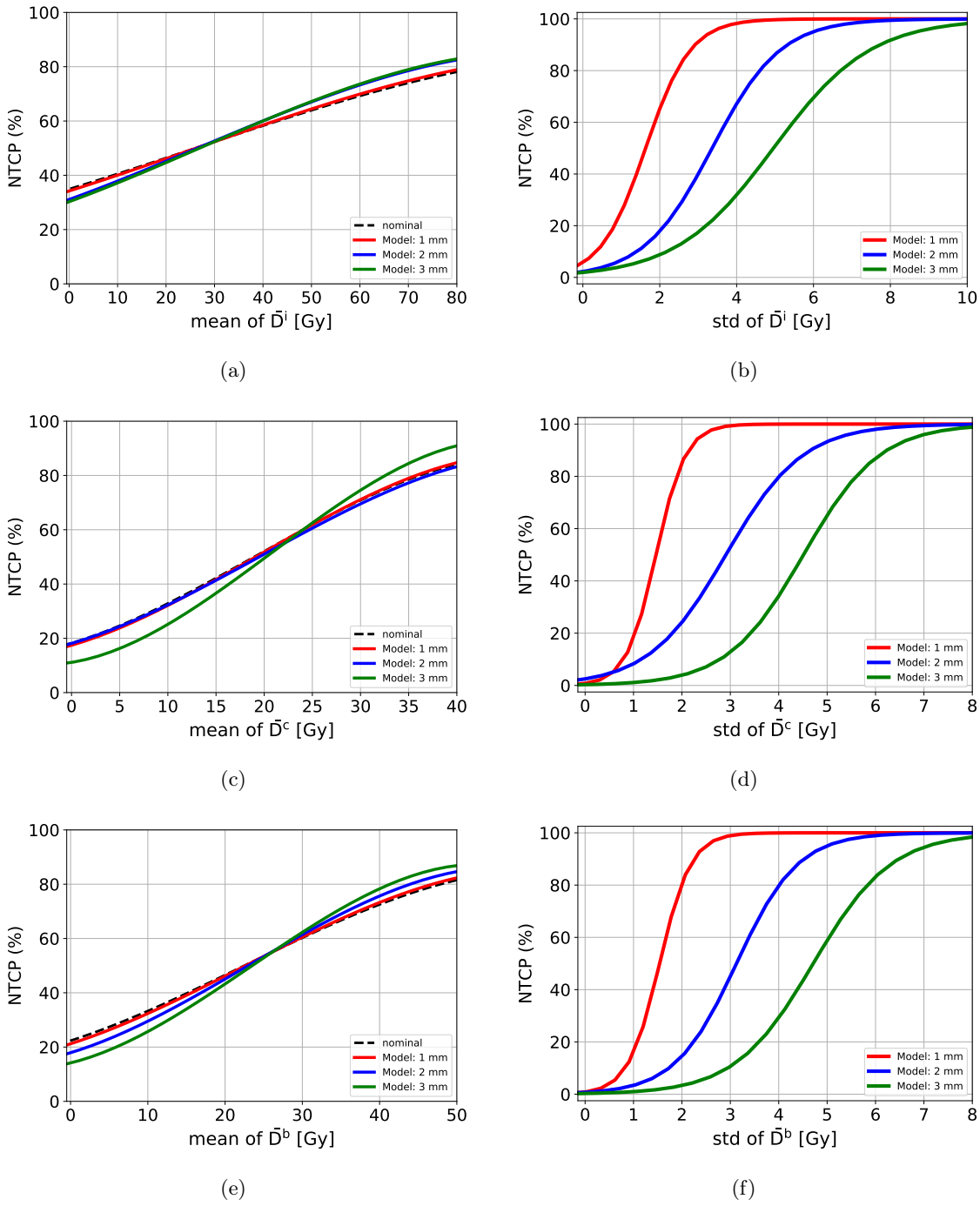
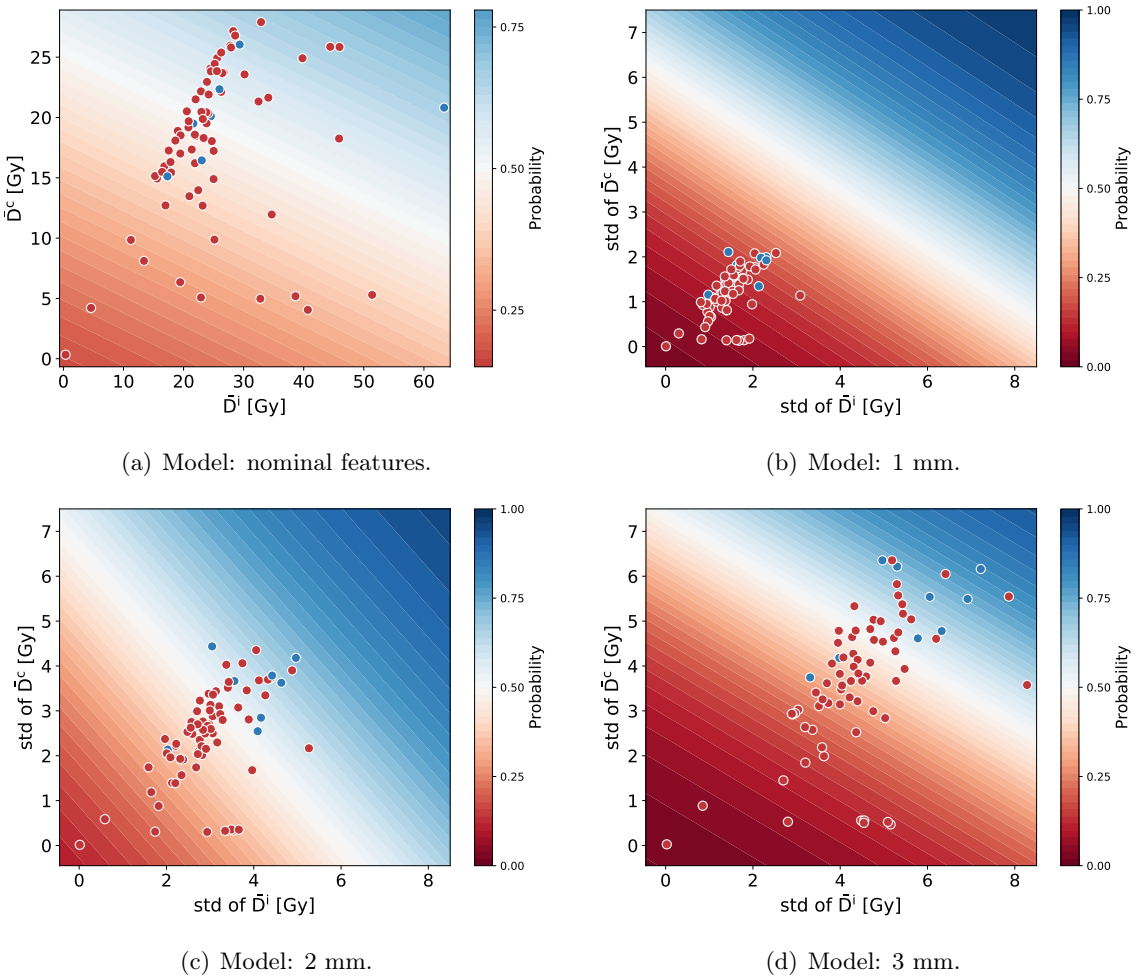


Figure 4.7: NTCP curves obtained with the classical LR models based on the mean dose ($\bar{D}$) to ipsilateral: (a) and (b); contralateral: (c) and (d); and both parotids: (e) and (f). Left graphs consider the nominal $\bar{D}$ and the mean of this feature. And the right graphs consider the standard deviation of the $\bar{D}$.



(a) Model: nominal features.

(b) Model: 1 mm.

(c) Model: 2 mm.

(d) Model: 3 mm.

Figure 4.8: Classification probability for bivariate LR model based on mean dose to the ipsilateral and contralateral parotid gland. Patients with G0-G1 xerostomia and G2+ xerostomia are represented in red and blue circles, respectively.

4.1.3 Multivariate models

The AUCs of the multivariate analysis for the models explained in Section 3.4.3 are presented in Figure 4.9. They correspond to the best AUC obtained for each model after the seven algorithms were evaluated. As explained in Section 3.3, the AUC was estimated by the nested-cross-validation for 70 iterations in the inner loop and 100 iterations in the outer loop.

Similar to the analysis of the classical models (Section 4.1.2), Figure 4.9.a) indicates that the model performance is close to a random classifier (AUC = 0.5) when considering the nominal or the mean of the $\bar{D}$ features, even when different classification algorithms are used. In the case of the standard deviation of the features, AUC values from 0.76 to 0.79 are obtained. The model with the best performance (AUC = 0.79 (0.74-0.83)) is the

one based on the standard deviation of the $\bar{D}$ features for the scenario of 3 mm, with the ET classifier.

For the second model based on ipsilateral volume (V^i) and contralateral gradient in the right-left direction ($\nabla_x{}^c$), a value of AUC = 0.92 (0.9-0.94) is obtained with the ET classifier. This value is higher than the individual predictive power of the V^i: AUC = 0.87 (0.75-0.95). This suggests that adding the dosimetric feature $\nabla_x{}^c$, the model can discriminate better between patients with and without G2+ xerostomia. It is possible to observe that when considering uncertainties in this dosimetric feature, the model performance decreases, obtaining lower values of AUC.

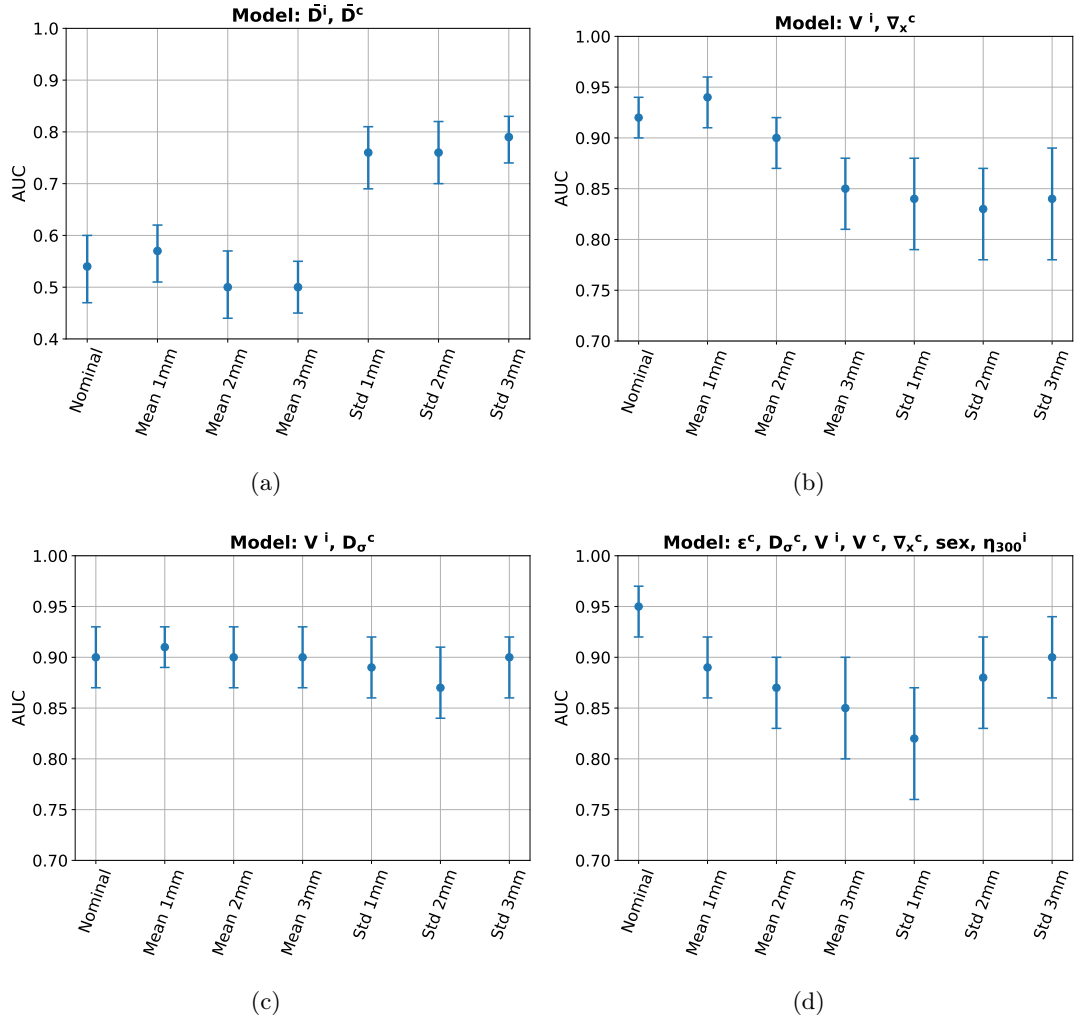


Figure 4.9: AUC values with 95% confidence intervals of the performance for the multivariate models based on the features indicated in each graph. Three scenarios are considered: nominal features, mean of the features, and standard deviation of the features.

When considering as predictors of G2+ xerostomia the V^i and the contralateral spread ($D_\sigma{}^c$) of the DVH, similar performances are obtained with or without uncertainties in the planned dose. This is illustrated in Figure 4.9.c), where the AUC values range from 0.87 to 0.91. In five of these models, the AUC corresponds to the performance obtained with the ET classifier. This classification algorithm is the same one reported by Gabryś et al. (2018) for their best predictive model.

Among all the models studied the best performance was obtained with the model comprised the following features: eccentricity of the contralateral parotid (ϵ^c), spread of the DVH for the contralateral parotid ($D_\sigma{}^c$), ipsilateral and contralateral volume (V^i, V^c), patient's sex, contralateral gradient in the right-left direction ($\nabla_x{}^c$) and ipsilateral η_{300}. This group of features predicts G2+ xerostomia with an AUC = 0.95 (0.92-0.97) with the LR-L2 classifier for the nominal scenario. When including uncertainties in the planned dose, lower AUC values were observed compared to the performance of the nominal feature.

Figure 4.10 shows the analysis of the multivariate models that consider the mean plus the standard deviation of each feature.

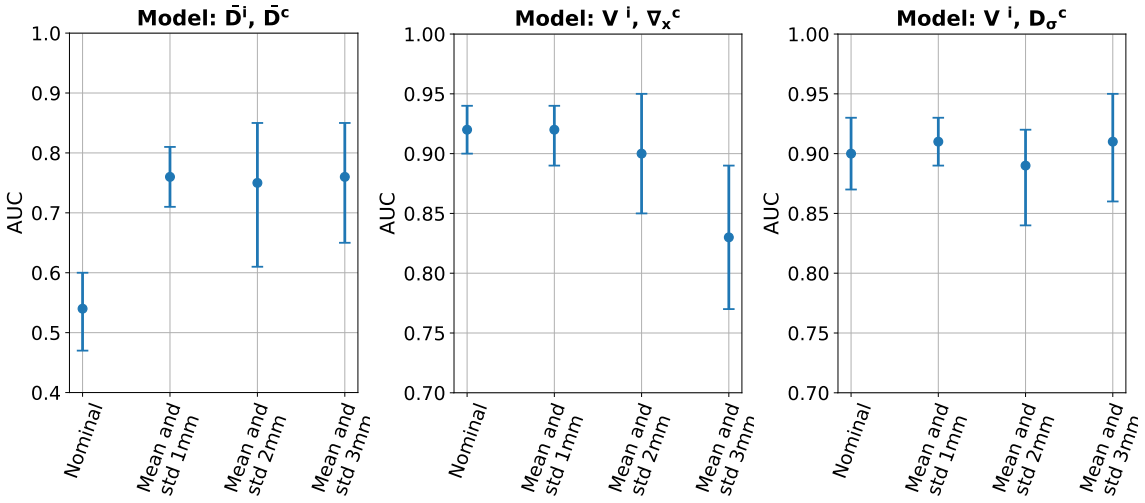


Figure 4.10: AUC values and 95% confidence intervals of the performance for the bivariate models based on the features indicated in each graph. The performance was calculated considering the nominal features, and the mean plus the standard deviation of the features.

Figure 4.10.a) indicates the value of AUC for the model based on the ipsilateral and contralateral mean doses. The mean together with the standard deviation of the features results in a better classification performance than the nominal features: AUC = 0.54 (0.47, 0.60). Considering the dose uncertainties, the models for the three displacements studied give similar predictive powers between them.

Figure 4.10.b) shows the AUC values for the case of the model based on the ipsilate-

ral volume and the right-left dose gradient of the contralateral parotid. One can note that the performance gets worse as the displacement of the planned dose increases when uncertainties are included, especially when using the largest scenario of 3 mm.

Considering the ipsilateral volume and the spread of the contralateral DVH as predictors of G2+ xerostomia, similar AUC values are obtained with the four scenarios, as shown in Figure 4.10.c).

4.2 Prediction of xerostomia considering uncertainties in toxicity grading

The results of the second part of this work (Section 3.5) are presented in this section.

The impact of uncertainties on the grade of toxicity was studied for the classical predictive models of xerostomia and a bivariate LR model based on ipsilateral volume (V^i) and spread of contralateral DVH ($D_\sigma{}^c$).

The investigation of the xerostomia grading uncertainties does not provide better prediction models. Nevertheless, there are variations in the probability of classification between the two scenarios. Figure 4.11 shows a comparison between the classification probabilities for both scenarios with the best predictive model. Although the AUC for the model represented in Figure 4.11.a) is 0.90 (0.88-0.91) in both cases, the decision boundary for G2+ xerostomia is affected when the different xerostomia grades of the patients are considered.

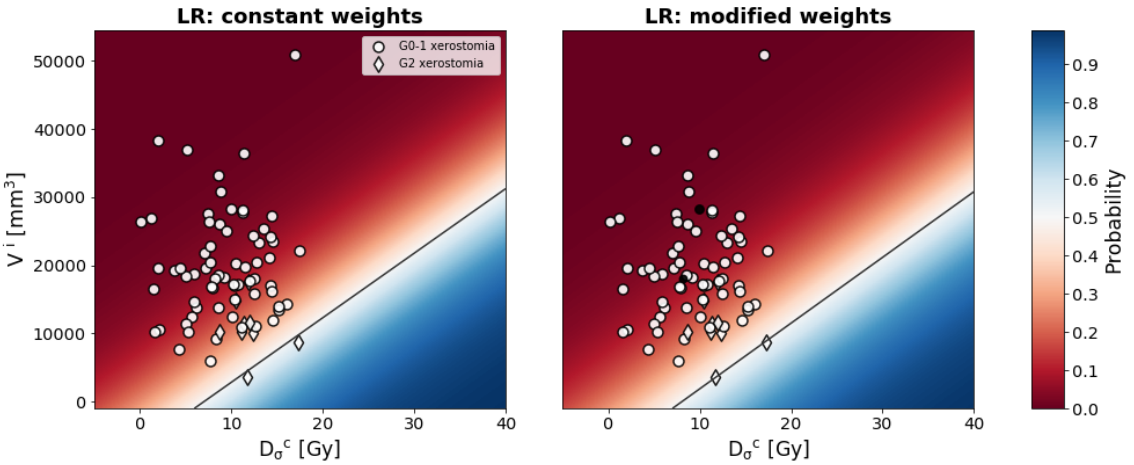
Sample weights	**Long-term xerostomia**	**Late xerostomia**
1.0	75	113
0.8	0	1
0.75	0	2
0.67	1	8
0.60	0	1
0.5	1	6

Table 4.2: Number of patients according to sample weights calculated for the two time intervals studied.

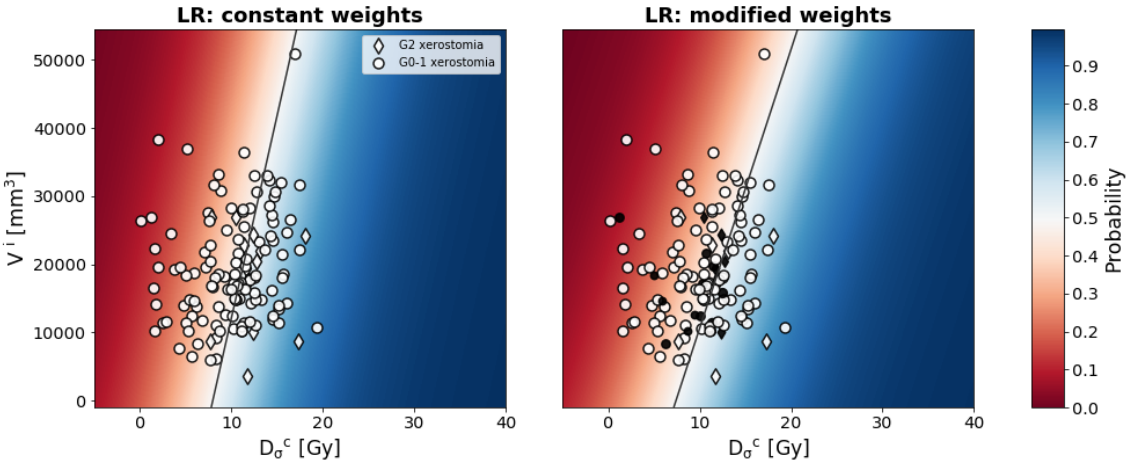
In the studied cohort, only two patients present different grades of xerostomia toxicity during the long-term interval. For this reason, the model analyzed above was also studied for the late G2+ xerostomia. In this interval, 18 patients show inconsistency in their grade of toxicity. Table 4.2 presents the number of patients who presented different weights for the classification according to the time interval studied. Like the long-term interval, the AUC value of the performance is the same in both scenarios: 0.64 (0.62-0.67). Figure 4.11.b)

shows the classification probabilities obtained with the model studied when the different xerostomia grades of each patient in the cohort are considered and not considered. It also shows the decision boundary for both cases, which presents a greater difference between the two scenarios compared to that obtained for long-term G2+ xerostomia.

The modified weights of the samples that were affected in both time intervals are presented in Table 4.3. The table also shows the actual sample class and the sample predictions obtained with the classification model. These samples correspond to the points indicated in black in Figure 4.11. It is possible to visualize that for some of these samples, their classification changes when considering in the model the modified sample weights.



(a) Long-term xerostomia.



(b) Late xerostomia.

Figure 4.11: Classification probability for long-term(a) and late G2+ xerostomia (b) of the LR model based on ipsilateral parotid volume and contralateral parotid spread with and without consideration of uncertainties in xerostomia grading, right and left images respectively. The black line corresponds to the decision boundary for each model. Patients with uncertainties in their toxicity grading are indicated in black.

Time interval	Sample class	Modified sample weights	Original class prediction	Modified class prediction
Long-term	Negative	0.67	Negative	Negative
	Negative	0.50	Negative	Negative
Late	Negative	0.75	Positive	Positive
	Positive	0.67	Positive	Negative
	Negative	0.80	Positive	Positive
	Positive	0.67	Positive	Positive
	Negative	0.50	Negative	Negative
	Negative	0.67	Negative	Negative
	Positive	0.50	Negative	Negative
	Negative	0.67	Positive	Negative
	Negative	0.75	Positive	Positive
	Positive	0.50	Positive	Positive
	Negative	0.50	Negative	Negative
	Negative	0.67	Negative	Negative
	Negative	0.67	Positive	Negative
	Negative	0.67	Negative	Negative
	Negative	0.50	Negative	Negative
	Positive	0.60	Positive	Negative
	Negative	0.50	Negative	Negative
	Negative	0.67	Negative	Negative

Table 4.3: Classification of the samples that modified their weights when considering the uncertainty in the toxicity grading. For the two time intervals studied, these predictions for G2+ xerostomia were obtained with the LR model based on the ipsilateral volume and the contralateral spread of the DVH. The sample class column corresponds to the original class, the modified sample weights to the weights considering the uncertainties, the modified and original class prediction correspond to the model predictions with and without considering grading uncertainties, respectively.

Chapter 5

Discussion

In the previous chapter, the results obtained by incorporating the planned dose uncertainties were presented for the different predictive models studied. It was also analyzed the impact when considering the uncertainties in the xerostomia toxicity grading system. The analysis and interpretation of the results set out in Chapter 4 are discussed in this chapter.

5.1 Prediction of xerostomia considering uncertainties in planned dose

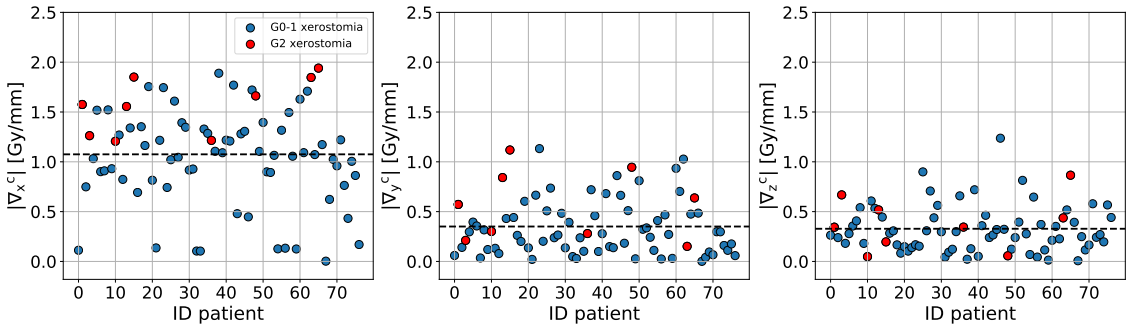
5.1.1 Univariate analysis

The study of the individual features predictive power showed that the best predictors were the volume and the gradient in the left-right direction for both parotids. Nevertheless, by incorporating dosimetric uncertainties, the expected value of this gradient decreased its ability to recognize xerostomia, even though the other features, in general, were not affected under this scenario. It is important to notice that the fact of computing the mean among the 100 samples of shifted doses for each patient generates a "blurring" effect in the dose. This implies that the gradient of the total dose decreases (Korevaar et al. (2019)). The dose gradient values in the medial direction are high compared to the gradients in the other directions, as can be seen in Figure 5.1. Therefore, by moving the dose randomly along this direction, the parotid may be receiving higher or lower doses than in the nominal scenario. This fact results in a lower mean gradient value to both parotids, as shown in Figure 5.2, making patients with and without xerostomia less recognizable.

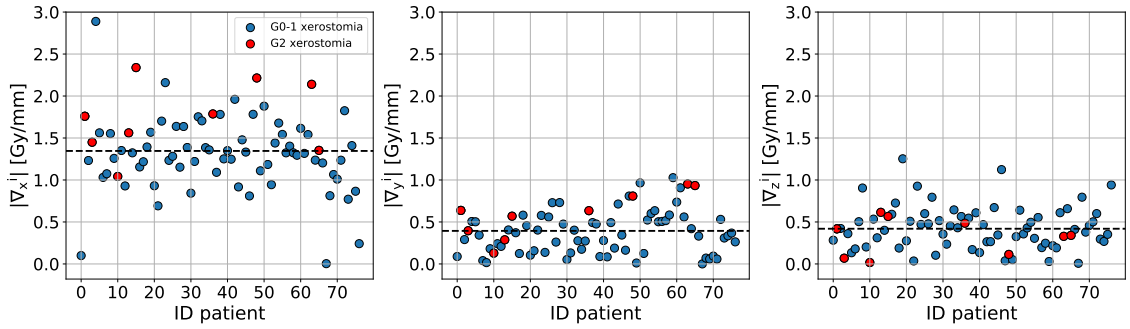
We found that, in general, the standard deviation of the individual features has better predictive power of xerostomia than the nominal feature values. Consequently, post-radiotherapy xerostomia development correlates with the variability of the dosimetric features during the treatment. Different studies have reported that during H&N RT treat-

ment, the parotids decrease in volume and generally shift in a medial direction (Barker et al. (2004), Fiorentino et al. (2012), Yao et al. (2015)). These two events can lead to the parotids moving towards the radiation field, increasing the patient's probability of developing xerostomia. This may explain the variation in the features between the fractions, especially the change in the mean dose to both parotids. Figure 5.3 shows the distribution of the mean dose to both parotids considering the different scenarios. It is posible to visualize that the separation between the positive (G2+ xerostomia) and negative (G0-1 xerostomia) patients is clearer for the uncertainty scenarios. Astaburuaga et al. (2019) studied for this same cohort the parotid gland migration in the medial direction. They found that this indicator was correlated with the gradient in the medial direction, improving the xerostomia prediction.

Considering the standard deviation of the features provides us with more information about what can occur during the treatment. Therefore, it is beneficial to have one value with which we can capture information related to the gradient and dose shape, and not only to the dose values.



(a) Contralateral parotid gland.



(b) Ipsilateral parotid gland.

Figure 5.1: Distribution of the absolute values of the gradients for the patients in the cohort. The dotted black line indicates the average value of each gradient.

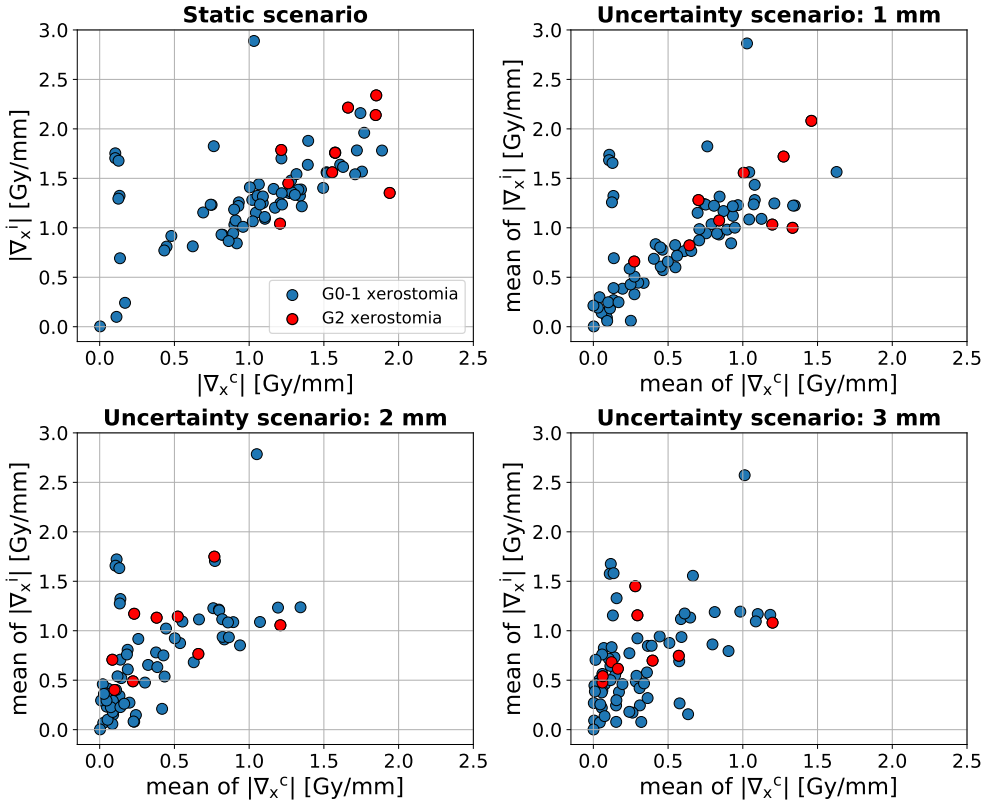


Figure 5.2: The right-left dose gradient to both parotids. The scenarios consider the nominal and the mean of the features for the three displacements.

5.1.2 Classical models

Currently, NTCP models for xerostomia development are based on the mean dose to parotid glands (Bentzen et al. (2010)). However, these classical models failed to predict this side effect for this cohort. In accordance with univariate analysis, the models that consider the mean doses expected value showed a slightly higher performance than the model based on nominal values. Nevertheless, we were able to build predictive models with a better capacity to recognize patients with G2+ xerostomia when considering the standard deviation of this feature for the three scenarios analyzed. This can be explained by the fact that due to the variation suffered by the parotids during the treatment, either in the volume or in the displacement of these, a variation in the mean dose has been reported (ODaniel et al. (2007), Robar et al. (2007), Yao et al. (2015)). Therefore, models based on the mean dose extracted directly from the planned dose become unrealistic for a H&N treatment. In the univariate NTCP models, it is possible to observe that the greater the displacement applied to the planned dose, the greater the variation of the dose between fractions necessary to damage the parotids leading to xerostomia. This is because having a high standard deviation means that the mean dose between fractions

varies widely. Therefore, the final dose received by the patient will be more blurred leads to a less conformal dose distribution.

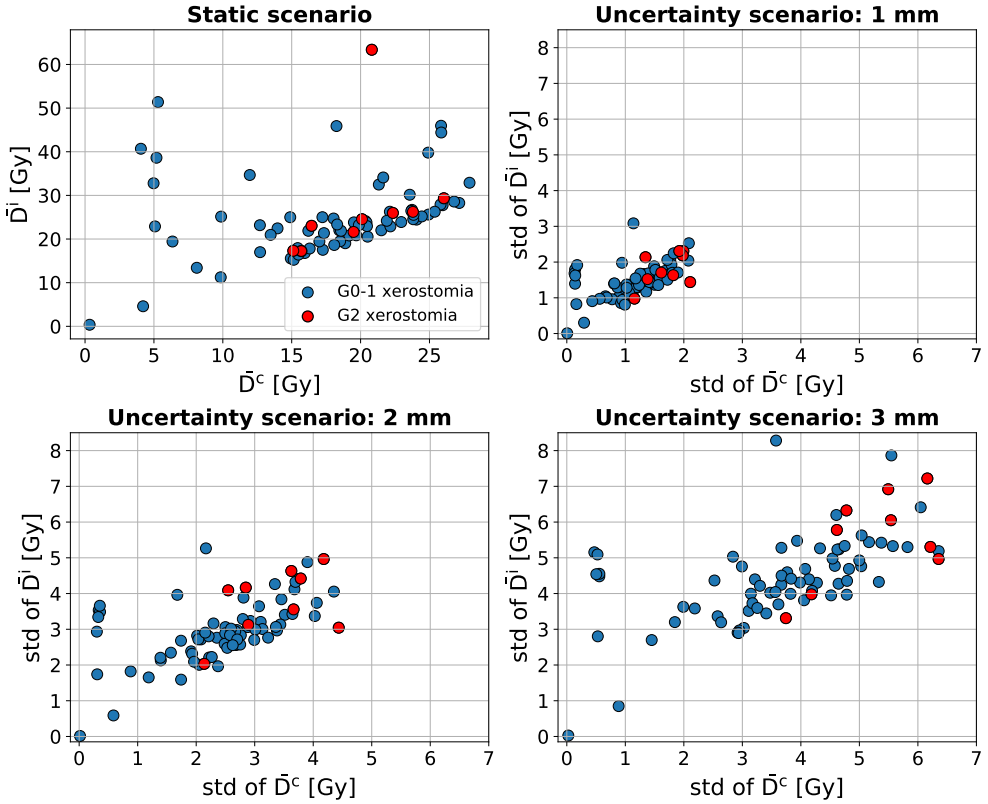


Figure 5.3: The distribution of the mean doses to both parotids. The scenarios consider the nominal and the standard deviation of the features for the three displacements.

We found the most remarkable improvement by including uncertainties for the model based on the mean dose to the ipsilateral parotid and contralateral parotid gland, for which the AUC increases from 0.47 to 0.83 (see Table 4.1). The bivariate model based on the standard deviation of the mean doses for the 3 mm displacement showed better identification of patients at risk (see Figure 4.8). Patients who present high variations in the mean dose for both ipsilateral and contralateral parotids during RT treatment are more likely to develop xerostomia. One explanation for this finding could be that QUANTEC recommendations for preventing this side effect are based on sparing the parotids with a certain dosage amount. However, if the mean dose varies during the treatment, it becomes more challenging to follow this recommendation.

5.1.3 Multivariate analysis

In an ML application, different factors can affect the performance of the model studied. The algorithm, the model evaluation metric, and the technique to estimate this metric

determine the model's performance. In this work, we tested seven ML algorithms with different hyperparameter options for each one. This methodology gives more possibilities of developing a model with a good performance.

In accordance with the results of the classical LR models based on the mean dose to both parotids, the multivariate analysis also found higher values of AUC when considering the standard deviation of these features as predictor of G2+ xerostomia. However, despite trying the different algorithms with variations in their hyperparameters, we did not obtain better predictive power than the classic LR models. Although we used the cross-validation technique in both studies, it is important to consider that the number of iterations in both cases was not the same. Because of the methodology used in the multivariate study, the number of iterations for the cross-validation was lower (70 iteration for the inned loop and 100 iterations for the outer loop) in comparison with the cross-validation for the study of the classical models (300 iterations).

In the univariate study, it was found that the volume of the ipsilateral parotid was a strong predictor of G2+ xerostomia, along with the contralateral gradient in the left-right direction. We found that the performance decreases when considering dosimetric uncertainties in the model based on these two features. This is in accordance with the results observed for the gradient feature discussed in the univariate study (Section 5.1.1) because the volume is not affected by uncertainties as it is not a dosimetric feature. However, a model with great predictive power: AUC = 0.92 (0.90-0.94) was developed with the ET classifier based on the nominal features of ipsilateral volume and contralateral gradient in the x-direction.

Gabryś et al. (2018) reported that the features selected for his best model were the ipsilateral parotid volume and the spread of the contralateral dose-volume histogram, with an AUC = 0.88 (0.84 - 0.91). In the present study, we obtained an AUC = 0.90 (0.87 - 0.93) for the model based on these features in the static scenario. The diference between these two results is due to the iteration number used in the nested-cross-validation. We found that the incorporation of the uncertainties almost does not influence the performance of the predictive model based on these features. This result is because the contralateral gland is less affected by the dose uncertainties than the ipsilateral gland. Thus, the AUC value of the contralateral spread feature practically does not change when considering the uncertainties, which is possible to notice in the univariate analysis.

The volume of the parotids can strongly predict the development of post-radiotherapy xerostomia by itself. Patients with smaller parotids are more likely to suffer from this complication because receiving a specific amount of dose more FSU can be irradiated at the same time, disabling faster the function of this gland. On the other hand, if a big parotid receives the same dose distribution, a lower number of FSU will be irradiated. With the last two models studied, it is possible to observe that by considering, in addition to the

volume, dosimetric characteristics of the parotid, such as the x-gradient and the spread, the xerostomia predictive power increases. For the static scenario, the performance obtained with the model based on the ipsilateral volume and the contralateral right-left dose gradient is slightly better compared to the performance of the model based on the ipsilateral volume and the contralateral spread (Figure 4.9.b) and c)). Although the gradient and the spread feature describe the dose variation within the parotids, they show different behaviors when considering the planned dose uncertainties. The first feature represents the dose variation only in the right-left direction, the direction in which the parotid glands usually migrate during the treatment, as mentioned above. In contrast, the spread of the dose-volume histogram indicates the dose variation in the three directions.

The fourth multivariate model studied considers dosimetric, radiomics, and a demographic feature corresponding to the patient's sex. We obtained a model with high predictive power: AUC = 0.95 (0.92 - 0.97) based on the nominal features. Nonetheless, we found that the incorporation of dosimetric uncertainties into the model decreases its predictive performance. An explanation could be that the three dosimetric features considered in this model present a different response under the uncertainty scenarios. From the univariate analysis and the bivariate model previously explained, it was found that the spread within the contralateral parotid is not mainly affected by the uncertainty scenarios. Therefore the variation in the AUC value for this multivariate model is explained by individually response of the contralateral gradient in the medial direction and the ipsilateral η_{300}. The individual performance of these features decreases when considering the expected value of these features. However, when considering the standard deviation of the features values, better performances are obtained.

It would be expected to have a better performance for the predictive models based on the mean together with the standard deviation of the features for the uncertainty scenarios because more information about these features is considered. However, we found that the AUC values of these models are similar to the performance obtained with the models explained above.

5.2 Prediction of xerostomia considering uncertainties in toxicity grading

Since post-radiotherapy xerostomia is a side effect that develops during and after treatment, its symptoms can appear and change during this period. For this reason, it is important to determine time intervals to be considered when building predictive models. The severity of side effects is generally evaluated using a standardized toxicity scale. In this work, CTCAE was used. This system considers objective measures and subjective symptoms (Trotti et al. (2007)), which can influence the inconsistency of the grade of xerostomia

assigned to a patient when recorded at different times.

We noticed that inconsistency in the xerostomia grades is more present in the time intervals closer to the end of the RT treatment (late xerostomia) because this side effect and its symptoms may still be developing. Nevertheless, as it was expected, the more time elapses since the RT treatment (long-term xerostomia), the more stable the grade recorded for each patient (see Table 4.2).

We found that incorporating a weight assigned to each sample (patient) by considering the grading variations in the predictive model affects the probability space generated by each model for both time intervals. Despite this, these differences were not significant, obtaining the same AUC values for the constant and modified weights corresponding to each time interval. This could be due to the small number of positive patients (G2+ xerostomia), corresponding to both times: 10% and 11% for late and long-term intervals, respectively.

In the probability space corresponding to the predictive model of xerostomia in the late interval, it is possible to observe a clearer difference than in the long-term interval. This is explained by the number of patients who present a non-constant weight. In an unweighted analysis, all samples are considered to have weight equal to 1, contributing equally to the cost function optimized by the corresponding algorithm to build the predictive model. In that case, all patients will have the same importance in the training process of the algorithm. On the other hand, when considering the grading uncertainties in a weighted analysis, each sample or patient influences the learning process of the algorithm according to its assigned weight.

Although in the long-term interval, only two patients present weights different from 1, it is possible to notice that the decision boundary moves slightly towards the region of greater spread of the contralateral gland that corresponds to the region of a higher probability of developing G2+ xerostomia. This is because there are two patients with G0-1 xerostomia weighing less than 1. We obtained observable differences in the probability space generated by the model for both time intervals; nevertheless, they are not reflected in the AUC value of the model performance.

It was not possible to find reported studies that consider the effect of uncertainties in the grading system. However, it is expected that performing this analysis in a cohort with more patients and follow-up reports the results to be statistically relevant when considering the grading uncertainties.

Chapter 6

Conclusions

Due to the sharp dose gradients of the dose distribution in an IMRT treatment, small changes in either volume or parotid displacement can lead to significant differences between the planned dose and the dose received by the patient, especially in regions such as the head and neck area. Therefore it is important to consider the dosimetric uncertainties that occur during the radiotherapy treatment as these may be responsible for the detriment of the organs at risk.

This thesis investigated how the incorporation of dosimetric uncertainties impacts on xerostomia prediction. The effect of grading uncertainties on the prediction of xerostomia was also studied. We examined a cohort of 77 H&N cancer patients, of which 9 of them presented G2+ xerostomia for the long-term time interval.

In order to simulate random positioning shifts between treatment fractions, dose uncertainties were simulated by shifting three-dimensionally the planned dose. This was performed considering a Gaussian distribution for the dose displacements. Three scenarios were studied using a normal distribution with σ equal to 1 mm, 2 mm, and 3 mm. It is important to notice that a static dose distribution ("dose cloud" approximation) was assumed for this simulation. Additional sources of dosimetric uncertainties, such as parotid volume change or migration, were not included in this study.

Different features were evaluated as predictors of xerostomia. It was possible to study the effect of dosimetric uncertainties in univariate, bivariate, and multivariate feature models. Predictive models based on the standard deviation of the dosimetric features presented a higher predictive power than those based on the mean of the features. This may be because the standard deviation, representing the variation of the feature between fractions, provides us more information about possible change in the parotids dose distribution during the treatment.

The most analyzed features were the mean dose to both parotids because it is currently used to predict xerostomia. The univariate analysis showed that when considering dosi-

metric uncertainties, the variation of this feature during treatment fractions is correlated with the development of this side effect. The expected value of the mean dose features, represented by their mean values, did not present a significant change in the predicted AUC value. This was also observed in the study of the classic LR models of xerostomia based on the mean dose to both parotids.

Our best xerostomia prediction model corresponds to a multivariate model based on some features that describe the dose distribution, the parotid shape, and also consider the patient's sex. An AUC value of 0.95 (0.92 - 0.97) was obtained for this model.

With the ML algorithms, it was possible to build predictive models of post-radiotherapy xerostomia with good performance considering dose uncertainties. The performance was estimated using the value of AUC calculated with the cross-validation technique. Nevertheless, it is necessary to consider that the available parameters for each algorithm and the estimation of the chosen metrics influence the development of a model. Therefore, ideally, the models should be validated in the future, evaluating their performance in an independent cohort.

Different studies that have made simulations of uncertainties in the planned dose have implemented re-planning of the treatment for each displacement. Performing this with our methodology could give us an even more realistic scenario.

To achieve a good predictive model in a supervised ML problem, it is important to have an accurate labeling of the samples from the database. The analysis of the grading uncertanities showed that different clasification probabilities were obtained when the weight of each sample is considered. These differences were not significant in this study and the AUC values obtained were equal to the calculated without the grading uncertainties. This is because the size of the samples used in this thesis was quite small (77 patients). Aditionally, the cohort was inbalanced: only 11% of patients developed G2+ xerostomia. Thus, it would be interesting to continue this analysis in the future, with a cohort of more patients. In particular, the study of the grading uncertainties could be improved with this consideration.

Many options are still available for combinations of features to be analyzed in the multivariate models. It would also be interesting to compare the performance of multivariate study for all the algorithms used and not only for the best algorithm selected in each model. Our study about dosimetric and grading uncertainties is a good starting point for investigating the impact of these uncertainties on previously validated prediction models in order to analyze their performance under more realistic RT treatment scenarios.

Bibliography

Acauan, M. D., Figueiredo, M. A. Z., Cherubini, K., Gomes, A. P. N., and Salum, F. G. (2015). Radiotherapy-induced salivary dysfunction: Structural changes, pathogenetic mechanisms and therapies. *Archives of Oral Biology*, 60(12):1802–1810.

American Cancer Society (2019). Radiation therapy side effects. `http://www.cancer.org/content/cancer/en/treatment/treatments-and-side-effects/treatment-types/radiation/effects-on-different-parts-of-body.html`.

Argirion, I., Zarins, K. R., Defever, K., Suwanrungruang, K., Chang, J. T., Pongnikorn, D., Chitapanarux, I., Sriplung, H., Vatanasapt, P., and Rozek, L. S. (2019). Temporal changes in head and neck cancer incidence in thailand suggest changing oropharyngeal epidemiology in the region. *Journal of Global Oncology*, (5):1–11.

Astaburuaga, R., Gabryś, H. S., Sánchez-Nieto, B., Floca, R. O., Klüter, S., Schubert, K., Hauswald, H., and Bangert, M. (2019). Incorporation of dosimetric gradients and parotid gland migration into xerostomia prediction. *Frontiers in Oncology*, 9.

Astreinidou, E., Bel, A., Raaijmakers, C. P., Terhaard, C. H., and Lagendijk, J. J. (2005). Adequate margins for random setup uncertainties in head-and-neck IMRT. *International Journal of Radiation Oncology, Biology, Physics*, 61(3):938–944.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Barker, J. L., Garden, A. S., Ang, K., ODaniel, J. C., Wang, H., Court, L. E., Morrison, W. H., Rosenthal, D. I., Chao, K., Tucker, S. L., Mohan, R., and Dong, L. (2004). Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system. *International Journal of Radiation Oncology, Biology, Physics*, 59(4):960–970.

Beetz, I., Schilstra, C., Burlage, F. R., Koken, P. W., Doornaert, P., Bijl, H. P., Chouvalova, O., Leemans, C. R., de Bock, G. H., Christianen, M. E., van der Laan, B. F., Vissink,

A., Steenbakkers, R. J., and Langendijk, J. A. (2012). 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. *Radiotherapy and Oncology*, 105(1):86–93.

Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Haken, R. K. T., and Yorke, E. D. (2010). Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *International Journal of Radiation Oncology, Biology, Physics*, 76(3):S3–S9.

Bernier, J., Hall, E. J., and Giaccia, A. (2004). Radiation oncology: a century of achievements. *Nature Reviews Cancer*, 4(9):737–747.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Budgell, G. (2002). Intensity modulated radiotherapy (IMRT)—an introduction. *Radiography*, 8(4):241–249.

Buettner, F., Miah, A. B., Gulliford, S. L., Hall, E., Harrington, K. J., Webb, S., Partridge, M., and Nutting, C. M. (2012). Novel approaches to improve the therapeutic index of head and neck radiotherapy: An analysis of data from the PARSPORT randomised phase III trial. *Radiotherapy and Oncology*, 103(1):82–87.

Cawley, G. and Talbot, N. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

Cho, M.-A., Ko, J.-Y., Kim, Y.-K., and H.-S. Kho, title = Salivary flow rate and clinical characteristics of patients with xerostomia according to its aetiology, j. . J. (2010). 37(3):185–193.

Chou, W. W., Puri, D. R., and Lee, N. Y. (2005). Intensity-modulated radiation therapy for head and neck cancer. *Expert Review of Anticancer Therapy*, 5(3):515–521.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dawes, C. and Wood, C. (1973). The contribution of oral minor mucous gland secretions to the volume of whole saliva in man. *Archives of Oral Biology*, 18(3):337–342.

Deasy, J. O., Moiseenko, V., Marks, L., Chao, K. C., Nam, J., and Eisbruch, A. (2010). Radiotherapy dose–volume effects on salivary gland function. *International Journal of Radiation Oncology, Biology, Physics*, 76(3):S58–S63.

Delana, A., Menegotti, L., Bolner, A., Tomio, L., Valentini, A., Lohr, F., and Vanoni, V. (2009). Impact of residual setup error on parotid gland dose in intensity-modulated radiation herapy with or without planning organ-at-risk margin. *Strahlentherapie und Onkologie*, 185(7):453–459.

Delaney, G., Jacob, S., Featherstone, C., and Barton, M. (2005). The role of radiotherapy in cancer treatment. *Cancer*, 104(6):1129–1137.

Dirix, P., Nuyts, S., and den Bogaert, W. V. (2006). Radiation-induced xerostomia in patients with head and neck cancer. *Cancer*, 107(11):2525–2534.

Dirix, P., Nuyts, S., Poorten, V. V., Delaere, P., and den Bogaert, W. V. (2007). The influence of xerostomia after radiotherapy on quality of life. *Supportive Care in Cancer*, 16(2):171–179.

Edgar, M., Dawes, C., and O'Mullane, D. (2012). *Saliva and oral health: an essential overview for the health professional*. Stephen Hancocks Limited, London, fourth edition.

Fiorentino, A., Caivano, R., Metallo, V., Chiumento, C., Cozzolino, M., Califano, G., Clemente, S., Pedicini, P., and Fusco, V. (2012). Parotid gland volumetric changes during intensity-modulated radiotherapy in head and neck cancer. *The British Journal of Radiology*, 85(1018):1415–1419.

Fiorino, C., Rancati, T., and Valdagni, R. (2009). Predictive models of toxicity in external radiotherapy. *Cancer*, 115(S13):3135–3140.

Fox, P. C., Busch, K. A., and Baum, B. J. (1987). Subjective reports of xerostomia and objective measures of salivary gland performance. *The Journal of the American Dental Association*, 115(4):581–584.

Friedman, J. H. (2001). Greddy function approximation: a gradient goosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Gabryś, H. S. (2018). *Machine learning using radiomics and dosiomics for normal tissue complication probability modeling of radiation-induced xerostomia*. PhD thesis, Ruprecht-Karls-Universität.

Gabryś, H. S., Buettner, F., Sterzing, F., Hauswald, H., and Bangert, M. (2018). Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia. *Frontiers in Oncology*, 8.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Heukelom, J., Kantor, M. E., Mohamed, A. S., Elhalawani, H., Kocak-Uzel, E., Lin, T., Yang, J., Aristophanous, M., Rasch, C. R., Fuller, C. D., and Sonke, J.-J. (2020). Differences between planned and delivered dose for head and neck cancer, and their consequences for normal tissue complication probability and treatment adaptation. *Radiotherapy and Oncology*, 142:100–106.

Hopcraft, M. and Tan, C. (2010). Xerostomia: an update for clinicians. *Australian Dental Journal*, 55(3):238–244.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Inc.

Houweling, A. C., Philippens, M. E., Dijkema, T., Roesink, J. M., Terhaard, C. H., Schilstra, C., Haken, R. K. T., Eisbruch, A., and Raaijmakers, C. P. (2010). A comparison of dose–response models for the parotid gland in a large group of head-and-neck cancer patients. *International Journal of Radiation Oncology, Biology, Physics*, 76(4):1259–1265.

Humphrey, S. P. and Williamson, R. T. (2001). A review of saliva: Normal composition, flow, and function. *The Journal of Prosthetic Dentistry*, 85(2):162–169.

Hurkmans, C. W., Remeijer, P., Lebesque, J. V., and Mijnheer, B. J. (2001). Set-up verification using portal imaging review of current clinical practice. *Radiotherapy and Oncology*, 58(2):105–120.

Jaguar, G. C., Prado, J. D., Campanhã, D., and Alves, F. A. (2017). Clinical features and preventive therapies of radiation-induced xerostomia in head and neck cancer patient: a literature review. *Applied Cancer Research*, 37(1).

Joiner, M. and van der Kogel, A. (2009). *Basic Clinical Radiobiology Fourth Edition*. Taylor & Francis.

Kałużny, J., Wierzbicka, M., Nogala, H., Milecki, P., and Kopeć, T. (2014). Radiotherapy induced xerostomia: Mechanisms, diagnostics, prevention and treatment – evidence based up to 2013. *Otolaryngologia Polska*, 68(1):1–14.

Kam, M. K., Leung, S.-F., Zee, B., Chau, R. M., Suen, J. J., Mo, F., Lai, M., Ho, R., yin Cheung, K., Yu, B. K., Chiu, S. K., Choi, P. H., Teo, P. M., hong Kwan, W., and Chan, A. T. (2007). Prospective randomized study of intensity-modulated radiotherapy on salivary gland function in early-stage nasopharyngeal carcinoma patients. *Journal of Clinical Oncology*, 25(31):4873–4879.

Kang, J., Schwartz, R., Flickinger, J., and Beriwal, S. (2015). Machine learning approaches for predicting radiation therapy outcomes: A clinicians perspective. *International Journal of Radiation Oncology, Biology, Physics*, 93(5):1127–1135.

Korevaar, E. W., Habraken, S. J., Scandurra, D., Kierkels, R. G., Unipan, M., Eenink, M. G., Steenbakkers, R. J., Peeters, S. G., Zindler, J. D., Hoogeman, M., and Langendijk, J. A. (2019). Practical robustness evaluation in radiotherapy – a photon and proton-proof alternative to PTV-based plan evaluation. *Radiotherapy and Oncology*, 141:267–274.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York.

Law, M. Y. Y. and Liu, B. (2009). DICOM-RT and its utilization in radiation therapy. *RadioGraphics*, 29(3):655–667.

Lee, N., Puri, D. R., Blanco, A. I., and Chao, K. S. C. (2007). Intensity-modulated radiation therapy in head and neck cancers: An update. *Head & Neck*, 29(4):387–400.

Lee, T.-F., Liou, M.-H., Ting, H.-M., Chang, L., Lee, H.-Y., Leung, S. W., Huang, C.-J., and Chao, P.-J. (2015). Patient- and therapy-related factors associated with the incidence of xerostomia in nasopharyngeal carcinoma patients receiving parotid-sparing helical tomotherapy. *Scientific Reports*, 5(1).

Mescher, A. L. (2016). *Junqueira's Basic Histology: text and atlas*. McGraw-Hill Education, fourteenth edition.

Mileusnic, D. (2005). Verification and correction of geometrical uncertainties in conformal radiotherapy. *Archive of oncology*, 13(3-4):140–144.

Naqa, I. E., Bradley, J., Blanco, A. I., Lindsay, P. E., Vicic, M., Hope, A., and Deasy, J. O. (2006). Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *International Journal of Radiation Oncology, Biology, Physics*, 64(4):1275–1286.

Naqa, I. E., Ruan, D., Valdes, G., Dekker, A., McNutt, T., Ge, Y., Wu, Q. J., Oh, J. H., Thor, M., Smith, W., Rao, A., Fuller, C., Xiao, Y., Manion, F., Schipper, M., Mayo, C., Moran, J. M., and Haken, R. T. (2018). Machine learning and modeling: Data, validation, communication challenges. *Medical Physics*, 45(10):e834–e840.

National Cancer Institute (2010). Common terminology criteria for adverse events v4.03.

Nishimura, Y., Nakamatsu, K., Kanamori, S., and Okumura, M. (2004). Importance of mean dose and initial volume of parotid glands in xerostomia of patients with head and neck cancers receiving imrt. *International Journal of Radiation Oncology, Biology, Physics*, 60(1):S522.

Nutting, C. M., Morden, J. P., Harrington, K. J., Urbano, T. G., Bhide, S. A., Clark, C., Miles, E. A., Miah, A. B., Newbold, K., Tanay, M., Adab, F., Jefferies, S. J., Scrase, C., Yap, B. K., AHern, R. P., Sydenham, M. A., Emson, M., and Hall, E. (2011). Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. *The Lancet Oncology*, 12(2):127–136.

ODaniel, J. C., Garden, A. S., Schwartz, D. L., Wang, H., Ang, K. K., Ahamad, A., Rosenthal, D. I., Morrison, W. H., Asper, J. A., Zhang, L., Tung, S.-M., Mohan, R., and Dong, L. (2007). Parotid gland dose in intensity-modulated radiotherapy for head and neck cancer: Is what you plan what you get? *International Journal of Radiation Oncology, Biology, Physics*, 69(4):1290–1296.

Owosho, A. A., Thor, M., Oh, J. H., Riaz, N., Tsai, C. J., Rosenberg, H., Varthis, S., Yom, S. H. K., Huryn, J. M., Lee, N. Y., Deasy, J. O., and Estilo, C. L. (2017). The role of parotid gland irradiation in the development of severe hyposalivation (xerostomia) after intensity-modulated radiation therapy for head and neck cancer: Temporal patterns, risk factors, and testing the QUANTEC guidelines. *Journal of Cranio-Maxillofacial Surgery*, 45(4):595–600.

Palma, G., Monti, S., Conson, M., Pacelli, R., and Cella, L. (2019). Normal tissue complication probability (NTCP) models for modern radiation therapy. *Seminars in Oncology*, 46(3):210–218.

Podgorsak, E. B. (2010). *Radiation Physics for Medical Physicists*. Springer Berlin Heidelberg.

Robar, J. L., Day, A., Clancey, J., Kelly, R., Yewondwossen, M., Hollenhorst, H., Rajaraman, M., and Wilke, D. (2007). Spatial and dosimetric variability of organs at risk in head-and-neck intensity-modulated radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 68(4):1121–1130.

Scrimger, R. (2011). Salivary gland sparing in the treatment of head and neck cancer. *Expert Review of Anticancer Therapy*, 11(9):1437–1448.

Tanasiewicz, M., Hildebrandt, T., and Obersztyn, I. (2016). Xerostomia of various etiologies: A review of the literature. *Advances in Clinical and Experimental Medicine*, 25(1):199–206.

Taylor, A. (2004). Intensity-modulated radiotherapy - what is it? *Cancer Imaging*, 4(2):68–73.

Teng, F., Fan, W., Luo, Y., Ju, Z., Gong, H., Ge, R., Tong, F., Zhang, X., and Ma, L. (2019). Reducing xerostomia by comprehensive protection of salivary glands in intensity-modulated radiation therapy with helical tomotherapy technique for head-and-neck cancer patients: A prospective observational study. *BioMed Research International*, 2019:1–9.

Trotti, A., Colevas, A. D., Setser, A., and Basch, E. (2007). Patient-reported outcomes and the evolution of adverse event reporting in oncology. *Journal of Clinical Oncology*, 25(32):5121–5127.

Tsangaratos, P. and Ilia, I. (2016). Comparison of a logistic regression and naïve bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena*, 145:164–179.

Yang, X.-S. (2019). Support vector machine and regression. In *Introduction to Algorithms for Data Mining and Machine Learning*, pages 129–138. Elsevier.

Yao, W.-R., Xu, S.-P., Liu, B., Cao, X.-T., Ren, G., Du, L., Zhou, F.-G., Feng, L.-C., Qu, B.-L., Xie, C.-B., and Ma, L. (2015). Replanning criteria and timing definition for parotid protection-based adaptive radiation therapy in nasopharyngeal carcinoma. *BioMed Research International*, 2015:1–8.