Aus der Radiologische Universitätsklinik der Universität Heidelberg
(Geschäftsführender Ärztlicher Direktor: Prof. Dr. Jürgen Debus)
Abteilung für RadioOnkologie und Strahlentherapie
(Geschäftsführender Direktor: Prof. Dr. Jürgen Debus)

# Machine learning using radiomics and dosiomics for normal tissue complication probability modeling of radiation-induced xerostomia

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Hubert Szymon Gabryś

aus
Neu Sandez, Polen

2018

Dekan: Herr Prof. Dr. Wolfgang Herzog
Doktorvater: Herr Prof. Dr. Markus Alber

# Contents

CHAPTER 1

# INTRODUCTION

The main goal of radiotherapy is the eradication of cancerous cells. Unfortunately, irradiation inevitably leads to concurrent damage to healthy tissue. The reason is that the radiation beams deposit the dose before they reach and after they leave the target volume (the latter is less of an issue for proton beams). Consequently, a good radiotherapy treatment plan delivers a sufficiently high dose to kill malignant cells while minimizing the injury to the organs surrounding the tumor. This renders an accurate estimation of normal tissue complication probability (NTCP) a crucial element of radiotherapy treatment planning.

Over the years, many approaches were taken to model NTCP in radiotherapy treatments. Empirical models focused on simplicity and ability to describe experimental data without making assumptions on the mechanisms of radiation-induced tissue damage. Conversely, tissue-architecture models stemmed from theoretical underpinnings of processes managing tissue response to radiation. Recently, increasingly more attention has been given to data-driven approaches. Decreasing cost of computing power and booming research on machine learning algorithms have facilitated use of advanced pattern recognition methods in modeling of radiotherapy treatment outcomes (Kang et al. 2015).

In the course of routine diagnostics and radiotherapy treatment planning, vast amount of data is generated. This data contains information on the patient's demographics, anatomy, physiology, and various treatment parameters. Consequently, the number and complexity of recorded factors that can potentially influence treatment outcome is enormous. Holistic comprehension of all this information and its relation to NTCP exceeds human capabilities. In such settings, algorithms that are able to recognize patterns in data and provide predictions for new observations become useful. In recent years, there has been a growing interest in the adoption of machine learning methods in NTCP modeling (Gulliford 2015, El Naqa et al. 2018). Most of the studies, however, were limited to evaluation of one or two methods. An analysis comprehensively comparing various machine learning algorithms in terms of their suitability for NTCP modeling was missing, even though such studies were available in the fields of radiomics (Parmar et al. 2015) and bioinformatics (Olson et al. 2017).

In this context, a cohort of 153 head-and-neck cancer patients treated with radiotherapy was retrospectively collected. The cohort was used to investigate predictive potential of miscellaneous demographic, radiomic (organ-shape), and dosiomic (dose-shape) features. Furthermore, suitability of seven classification algorithms, six feature selection methods, and nine data cleaning/class

balancing techniques to NTCP modeling of xerostomia was examined.

## 1.1 Normal tissue complication probability models

### 1.1.1 Empirical models

Early radiobiological studies revealed that the dose-response follows a sigmoidal (s-shaped) curve. This relation suggests that the complication-inducing dose is subject to some probability distribution. The underlying uncertainty of NTCP stems from the probabilistic nature of the radiation damage mechanism and inter-individual differences in radiosensitivity, for example, efficiency of DNA repair mechanisms. Generally, when a tissue is irradiated to a dose $d$, NTCP can be described by

$$\mathrm{NTCP}(d) = \int_{-\infty}^{d} p(x)\,\mathrm{d}x\,, \tag{1.1}$$

where $p(x)$ is the probability that the complication-inducing threshold dose is equal to $x$.

The sigmoidal shape of NTCP function has been modeled using various representations (Figure 1.1). Most common choices were cumulative distribution functions of the normal distribution (probit model), logistic distribution (logit or logistic model), and log-logistic distribution (log-logit or log-logistic model). All these functions can be parameterized to have a similar shape and therefore the choice of the particular representation was typically a matter of personal preference or mathematical convenience (Bentzen & Tucker 1997).

The dose-response curves are often parameterized with the location parameter $d_x$ corresponding to a certain complication probability and the normalized dose-response gradient $\gamma$ quantifying the steepness of the dose-response function. The $d_x$ parameter is usually chosen to correspond to 50% ($d_{50}$) or 37% ($d_{37}$) complication probability. The $\gamma$ parameter was first introduced by Brahme (1984) and is defined as

$$\gamma(d) = d\frac{\mathrm{dNTCP}(d)}{\mathrm{d}d}\,. \tag{1.2}$$

The $\gamma$ parameter can be interpreted as the percentage change in the NTCP corresponding to a 1% change in the dose. The value used for model parameterization is usually chosen to correspond to the maximum steepness of the dose-response curve and is often marked with a subscript, for example, $\gamma_{50}$. It is important to distinguish that the dose for which $\mathrm{dNTCP}(d)/\,\mathrm{d}d$ is maximal is not the same as the dose for which the $\gamma$ reaches the maximum value (Bentzen & Tucker 1997).

#### Probit model

It was first observed by Holthusen (1936) that the empirical dose-response curve resembles the cumulative distribution function of the normal distribution. The name of the model comes from the probit function, which is the inverse of the cumulative distribution function of the standard normal distribution. The probit model assumes that radiosensitivity is normally distributed within a population and can by represented by

$$p(d|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\,, \tag{1.3}$$

where $\mu$ is the location parameter and $\sigma$ is the scale parameter. In the normal distribution, $\mu$ corresponds to the mean, the median, and the mode of the distribution, whereas $\sigma$ is the standard deviation. It follows from the Equation 1.1 that the NTCP can be calculated as

$$\mathrm{NTCP}(d|\mu,\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(d-\mu)/\sigma} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \tag{1.4}$$

or, equivalently,

$$\mathrm{NTCP}(d|\mu,\sigma) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \tag{1.5}$$

where erf is the error function given by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t\,. \tag{1.6}$$

Under the parameterization with $d_{50}$ and $\gamma_{50}$, the equation becomes

$$\mathrm{NTCP}(d|d_{50},\gamma_{50}) = \frac{1}{2}\left\{1 - \mathrm{erf}\left[\gamma_{50}\sqrt{\pi}\left(1 - \frac{d}{d_{50}}\right)\right]\right\}. \tag{1.7}$$

One of the advantages of the probit model is its convenience for studying effects of multiple sources of uncertainty in radiosensitivity and dose delivery (Bentzen & Tucker 1997).

The Probit model has also been defined in terms of the natural logarithm of the dose rather than the dose, which gave rise to the log-probit model (Goitein 1979). The attractiveness of such a model is that the NTCP tends to zero as the dose approaches zero. In contrast, NTCP in the probit model can be very small but can never reach zero.

### Logistic model

The logistic model is based on the logistic distribution

$$p(d|\mu,s) = \frac{\exp[-(d-\mu)/s]}{s\{1+\exp[-(d-\mu)/s]\}^2}, \tag{1.8}$$

where $\mu$ is the location parameter and $s$ is the scale parameter. In the logistic distribution, $\mu$ corresponds to the mean, the median, and the mode of the distribution, whereas $s$ is quantifies the spread of the distribution and is a linear function of the standard deviation ($\sigma = s\pi/\sqrt{3}$). The logistic distribution has a shape similar to the normal distribution with slightly heavier tails. The related NTCP function is given by

$$\mathrm{NTCP}(d|\mu,s) = \frac{1}{1+\exp[-(d-\mu)/s]} \tag{1.9}$$

or, equivalently,

$$\mathrm{NTCP}(d|d_{50},\gamma_{50}) = \frac{1}{1+\exp[4\gamma_{50}(1-d/d_{50})]}\,. \tag{1.10}$$

In real-life data sets, probit and logistic models yield similar results. The advantage of the

logistic model is that its coefficients can be interpreted in terms of the odds of complication, that is

$$\frac{\mathrm{NTCP}(d)}{1-\mathrm{NTCP}(d)} = \exp\left(-\frac{\mu}{s}+\frac{1}{s}d\right) = \exp\left(-4\gamma_{50}+\frac{4\gamma_{50}}{d_{50}}d\right) . \tag{1.11}$$

It follows that the odds of complication with no irradiation are $\exp(-\mu/s)$ or $\exp(-4\gamma_{50})$. Each additional one-unit increase in radiation dose changes the odds of complication by a factor of $\exp(1/s)$ or $\exp(4\gamma_{50}/d_{50})$.

### Log-logistic model

Another commonly used NTCP model stems from the log-logistic probability distribution. This model is equivalent to the logistic model with the natural logarithm of the dose as the covariate. Such description of the dose-response relationship was proposed by Suit et al. (1965). The probability density function of the log-logistic distribution is given by

$$p(d|\mu,k) = \frac{(k/\mu)(d/\mu)^{k-1}}{[1+(d/\mu)^k]^2} , \tag{1.12}$$

where $\mu$ is the scale parameter and $k$ is the shape parameter. The scale parameter $\mu$ corresponds to the median of the distribution and together with the shape parameter $k$ affects the spread of the distribution. The NTCP function related with the log-logistic probability distribution is given by

$$\mathrm{NTCP}(d|\mu,k) = \frac{1}{1+(\mu/d)^k} . \tag{1.13}$$

Under the parameterization with $d_{50}$ and $\gamma_{50}$, the equation becomes

$$\mathrm{NTCP}(d|d_{50},\gamma_{50}) = \frac{1}{1+(d_{50}/d)^{4\gamma_{50}}} . \tag{1.14}$$

The support of the log-logistic model is limited to positive values, which makes physical sense since the dose cannot be negative. Similarly to the log-probit model, the NTCP predicted by the log-logistic model tends to zero as the dose approaches zero. In contrast to probit or logistic model, the log-logistic model is asymmetric and exhibits positive skewness for $k > 1$.

### The Lyman model

The previously described NTCP models are suitable for description of a dose response of an organ under condition of homogeneous irradiation of the whole organ volume. They require, however, modifications to accommodate them for treatment scenarios when only a part of the organ is irradiated. One of the first empirical models taking this into account was the Lyman model (Lyman 1985).

Lyman based his model on the probit function. As mentioned before (and also stated by Lyman) the choice of the probit model was arbitrary and any other function of sigmoidal shape could have been chosen as well. Lyman introduced two modifications to the probit model. First, based on the previous studies (Schultheiss et al. 1983), he assumed that the partial volume tolerance dose $\mu(v)$

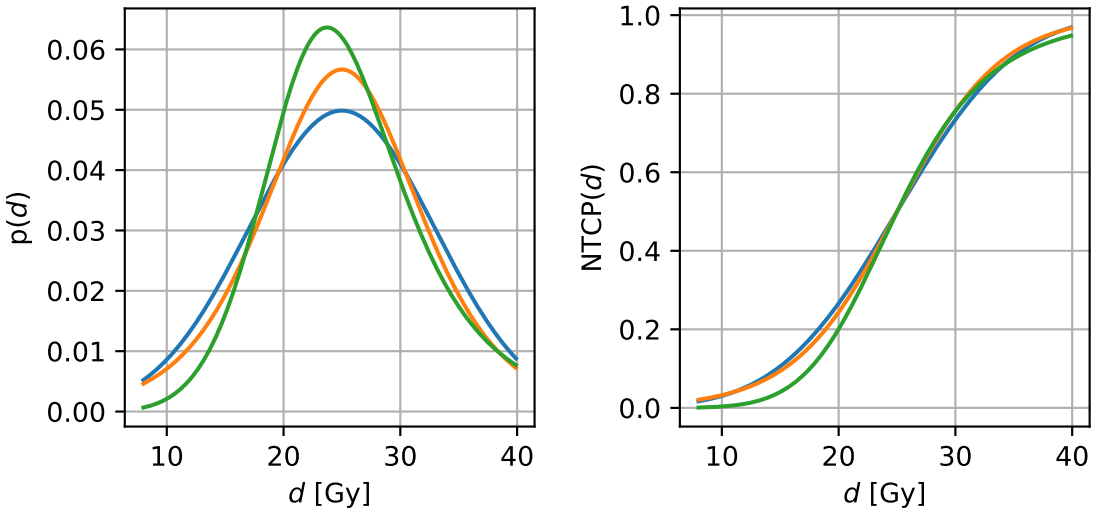



**Figure 1.1:** A comparison of three probability distribution functions: the normal distribution (—), logistic distribution (—), and log-logistic distribution (—). The cumulative distribution functions of these distributions were often used to model sigmoidal dose response of NTCP models. All distributions were parameterized to have 25 Gy mean and 8 Gy standard deviation.

can be described by the power law

$$\mu(v) = \mu_0 v^{-n}\,, \tag{1.15}$$

where $v$ is a partial volume and $\mu_0$ is the tolerance dose for the whole organ irradiation, that is $v = 1$. Second, he approximated the standard deviation $\sigma$ as

$$\sigma = m\mu(v)\,, \tag{1.16}$$

where $m$ is a free parameter. As a result, the Lyman model is a function of the dose $d$ to the partial volume $v$, given by

$$\mathrm{NTCP}(d, v|\mu_0, m, n) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(d-\mu_0 v^{-n})/(m\mu_0 v^{-n})} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x\,, \tag{1.17}$$

where $\mu_0 v^{-n}$ is the mean and $m\mu_0 v^{-n}$ is the standard deviation of the normal distribution. The mean is parameterized by the partial volume $v$ and the $n$ parameter, which controls the strength of the volume effect. For whole organ irradiation, $v = 1$ and the mean is simply $\mu_0$. The $n$ parameter is typically bounded by the interval $[0, 1]$. Values of $n$ close to zero indicate a weak volume effect, whereas values of $n$ close to one signify a strong volume effect. A weak volume effect corresponds to the situation where the NTCP is driven mainly by the dose and depends weakly on the volume of the irradiation. Conversely, strong dependence of the NTCP on the irradiated volume indicates a strong volume effect.

The advantages of the Lyman model are simplicity and applicability to partial irradiations. Its main limitation, however, is the assumption of a uniform dose distribution within the irradiated partial volume. In other words, the Lyman model was only suitable for single-step dose-volume histograms. In order to extend applicability of the model to non-uniform irradiations, dose-volume

histogram reduction schemes had to be designed.

#### The Lyman-Kutcher-Burman model

Kutcher & Burman (1989) proposed a dose-volume histogram reduction scheme called the effective volume method. This approach transforms a non-homogeneous histogram to a homogeneous single-step histogram with the volume $v_{eff}$ and the dose equal to the maximum dose to the organ ($d_{max}$). The effective volume is calculated based on the assumption that the power-law relationship applies independently to each volume element of the histogram. The Kutcher-Burman DVH reduction scheme is given by

$$v_{eff} = \sum_{i=1}^{N} \left( \frac{d_i}{d_{max}} \right)^{1/n} \Delta v_i \,, \tag{1.18}$$

where $N$ is the number of histogram bins, $d_i$ is the dose to the $i$th bin, $d_{max}$ is the maximum dose, and $v_i$ is the height of the histogram step. The Lyman model together with the Kutcher-Burman DVH reduction scheme was later called the Lyman-Kutcher-Burman (LKB) model.

In 1992, Mohan et al. proposed a different formulation of the LKB model. Instead of calculating a partial volume receiving the maximum dose, they calculated an equivalent uniform dose $d_{eud}$ to the whole organ volume that would result in the same complication probability. $d_{eud}$ is given by

$$d_{eud}(\{d_i\}; n) = \left( \sum_{i=i}^{N} v_i d_i^{1/n} \right)^n \,, \tag{1.19}$$

where $N$ is the number of voxels, $\{d_i\}$ is the set of doses to all voxels, $d_i$ is the dose to the $i$th voxel, $v_i$ is the partial volume of the $i$th voxel, and $n$ is a parameter which controls the strength of the volume effect. The NTCP is then given by

$$\mathrm{NTCP}(d_{eud}|d_{50}, m) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(d_{eud}-d_{50})/(md_{50})} \exp\left( -\frac{x^2}{2} \right) \mathrm{d}x \,, \tag{1.20}$$

where $d_{50}$ is the equivalent uniform dose corresponding to 50% complication probability and $md_{50}$ controls the slope of the NTCP curve. Mohan's formulation allows to bypass the use of DVHs as an intermediate step and was widely adopted afterwards.

### 1.1.2 Tissue-architecture models

A completely different approach to NTCP modeling was taken in tissue architecture models. The idea was that theoretical assumptions on the structure of a tissue and its response to radiation can lead to successful NTCP models. Some of the most well-known tissue-architecture models are the Poisson model and the relative-seriality model.

#### Poisson model

The Poisson model of radiation response was first introduced by Munro & Gilbert (1961) with an application to predicting tumor control probability (TCP). The model assumes log-linear cell-

survival curve. This means a single hit on a critical target in the cell is enough to induce cell death. Under this assumption, the survival fraction $S$ of cells irradiated with a dose $d$ is given by

$$S(d) = e^{-d/d_0}\,, \tag{1.21}$$

where $d_0$ is a dose which results in the survival fraction of $e^{-1}$ (approximately 0.37). Therefore, the number $N$ of cells surviving irradiation can be calculated as

$$N = N_0 e^{-d/d_0}\,, \tag{1.22}$$

where $N_0$ is the initial number of cells. Naturally, $N$ given by Equation 1.22 is the expected number of cells surviving irradiation. The exact number of cells surviving irradiation with a dose $d$ is a random variable subject to some probability distribution. Assuming that the survival of a single cell is independent from the state of any other cell, the number of surviving cells at a given dose $d$ can be approximated with the Poisson probability distribution $N \sim \text{Pois}(N_0 S(d))$ parameterized by the product of the initial number of cells and the surviving fraction. The distribution is given by

$$p(N|N_0 S(d)) = \frac{(N_0 S(d))^N e^{-N_0 S(d)}}{N!}\,. \tag{1.23}$$

If we fix the number of surviving cells to zero, meaning that all cells are killed by the radiation dose $d$, the dose-response curve is then represented by

$$p(N = 0|N_0 S(d)) = e^{-N_0 S(d)}\,, \tag{1.24}$$

which after after substituting $S(d)$ from Equation 1.21 gives

$$\text{NTCP}(d|d_0) = e^{-N_0 \exp(-d/d_0)}\,. \tag{1.25}$$

The Poisson model is often parameterized with the normalized dose-response gradient $\gamma$ and a dose $d_x$ corresponding to $x/100$ complication probability, typically $d_{37}$ or $d_{50}$. The dose-response gradient of the dose-response curve, given by Equation 1.25, has the greatest value for the dose corresponding to $e^{-1}$ complication probability. Therefore, $\gamma$ for the Poisson model is given by

$$\gamma = d_{37}\frac{\text{dNTCP}(d)}{\text{d}d} = \frac{\ln N_0}{e} \tag{1.26}$$

and $d_x$ by

$$d_x = d_0[e\gamma - \ln(\ln 100 - \ln x)]\,. \tag{1.27}$$

With this parameterization, the NTCP for the Poisson model is given by

$$\text{NTCP}(d|d_x, \gamma) = e^{-\exp\{e\gamma - [e\gamma - \ln(\ln 100 - \ln x)]d/d_x\}}\,. \tag{1.28}$$

For $d_x = d_{50}$, the NTCP is then given by

$$\text{NTCP}(d|d_{50}, \gamma) = e^{-\exp\{e\gamma - [e\gamma - \ln(\ln 2)]d/d_x\}}\,, \tag{1.29}$$

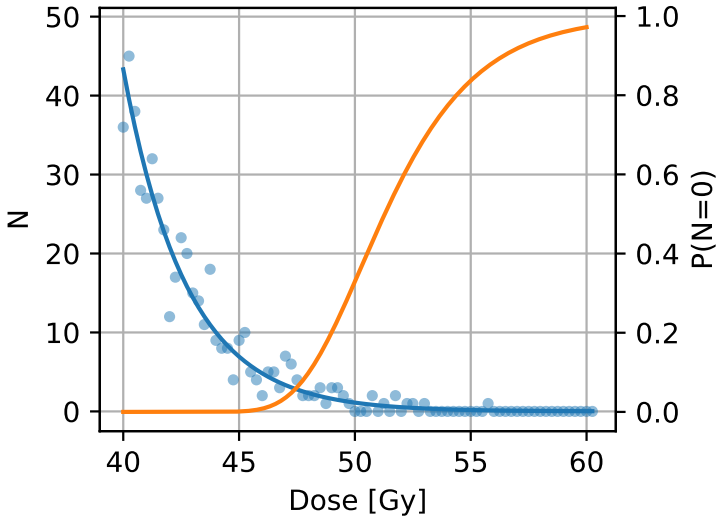



**Figure 1.2:** Illustration of a Poisson distribution for number $N$ of cells surviving irradiation with dose $d$ (•) (Equation 1.23), the expected number of cells surviving irradiation with dose $d$ (—) (Equation 1.22), and corresponding probability that no cells will survive (—) (Equation 1.25). The parameters used to plot the curves are $N_0 = 10^8$ and $d_0 = 2.73$ Gy and are the same as the parameters used to plot one of the TCP curves in Figure 2 in Brahme (1984).

whereas for $d_x = d_{100/e} \approx d_{37}$ it simplifies to

$$\mathrm{NTCP}(d|d_{37},\gamma) = e^{-\exp[e\gamma(1-d/d_{37})]}\,. \tag{1.30}$$

Another formulation of the Poisson model was proposed by Källman et al. (1992). In this formulation the model is based on the number 2 instead of $e$. Consequently, parameterization with $d_{50}$ results in much simpler formula, namely

$$\mathrm{NTCP}(d|d_{50},\gamma) = 2^{-\exp[e\gamma(1-d/d_{50})]}\,. \tag{1.31}$$

In fact, this formulation is only an approximation of the Poisson model, albeit it is often referred to in the literature as the Poisson model. Even though the two formulations are mathematically different and lead to slightly different values of $\gamma$ parameter, the practical implications of this approximation are negligible (Bentzen & Tucker 1997).

### Relative seriality model

The relative seriality model proposed by Källman et al. (1992) assumes that an organ can be divided into functional subunits (FSUs). The concept of FSUs was first introduced by Withers et al. (1988) and is based on the idea that cells in a given tissue are organized into subunits responsible for a certain function, hence the name functional subunits. The FSUs are interconnected and the overall response of an organ depends not only on the number of FSUs surviving irradiation but also the organization of the subunits (Figure 1.3). For example, if the subunits are connected in series, destruction of a single FSU leads to impairment of the whole organ. The NTCP for such a structure

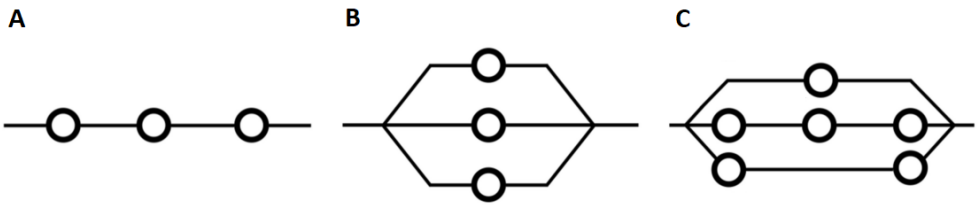



**Figure 1.3:** Schematics of different kinds of tissue architectures: A) serial structure, B) parallel structure, C) mixed (complex) structure. Adapted from Coates et al. (2016).

can be represented by

$$\mathrm{NTCP} = 1 - \prod_{i=1}^{M}(1 - p_i(d_i))\,, \tag{1.32}$$

where $p_i$ is a probability of complication of the $i$th subunit when receiving dose $d_i$ and $M$ is the total number of FSUs. Conversely, if an organ exhibits parallel organization of FSUs the NTCP is given by

$$\mathrm{NTCP} = \prod_{j=1}^{N} p_j(d_j)\,, \tag{1.33}$$

where $p_i$ is a probability of complication of the $j$th subunit when receiving dose $d_i$ and $N$ is the total number of FSUs. In practical applications, the radiation-response of an organ resembles a mix of serial and parallel response and can by described by

$$\mathrm{NTCP} = \prod_{j=1}^{N}\left[1 - \prod_{i=1}^{M}(1 - p_{ij}(d_{ij}))\right]\,, \tag{1.34}$$

where $M$ is the number of FSUs connected in series and $N$ is the number of parallel connections. Källman et al. showed that an arbitrary combinations of serial and parallel FSU arrangements can by conveniently modeled when instead of a number of serial and parallel FSUs one calculates a relative seriality coefficient $s$

$$s = \frac{M}{M \cdot N} = \frac{1}{N}\,. \tag{1.35}$$

If we assume that each subunit has the same complication probability $p$, then

$$p_{ij}(d_{ij}) = p(d_{ij}) \tag{1.36}$$

and the NTCP given by Equation 1.34 becomes

$$\mathrm{NTCP}(d|s,v) = \left[1 - \prod_{i=1}^{N}(1 - p(d_i)^s)^{v_i}\right]^{1/s}\,, \tag{1.37}$$

where $N$ is the number of subunits, $d_i$ dose to the $i$th subunit, $s$ relative seriality coefficient, $v_i$ fractional volume of $i$th subunit. This formulation is called the relative seriality model. The complication

probability $p$ is typically represented with the Poisson model

$$p(d) = 2^{-\exp\left[e\gamma(1-d/d_{50})\right]}. \tag{1.38}$$

In practical applications, it is usually assumed that a FSU is a voxel of the dose cube.

### 1.1.3 Data-driven approaches

Data-driven models are in a way a step-up from empirical models. Instead of relying on a few, preselected features, they are designed to recognize patterns in high-dimensional data sets, extract informative features, and make predictions. In other words, they are able to learn from data. Such models are suitable for applications where dimensionality and complexity of the available data renders hand-crafting the models impractical. The problem of learning from data is a focus of the field of machine learning. Machine learning has gained a lot of interest in recent years and has been successfully used in many applications, such as speech recognition, face recognition, targeted advertising, autonomous driving, just to name a few.

Machine learning problems can be categorized into three general classes: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, both input data and data labels are available. Its goal is to capture the input-output relation in the data. In clinical context this could be predicting treatment outcome based on various predictors. Usually large amounts of input data are relatively easily accessible but assignment of data labels may be time-consuming and expensive. For this reason, there is a subclass of supervised learning called semi-supervised learning. Semi-supervised learning algorithms make use of data where apart from input-output data samples, input-only observation are available. In unsupervised learning one is given only input data. The goal of unsupervised learning is to find patterns and relations in the available data. An example could be observations clustering, for example, finding groups of patients characterized by similar properties. In reinforcement learning, the model is not presented with data labels but rather with a reward for a correct decision. In the area of NTCP modeling, the vast majority of models belong to the supervised learning category.

Machine learning models are often designed to form a pipeline consisting of a number of layers. For example, a model can have a data cleaning layer followed by, a sampling layer, a feature selection layer, and finally a classification layer. The data cleaning layer scans the data looking for outliers, noisy observations, and observations on the border between the class clusters. Removal of such observations improves class cluster definitions, often facilitating model learning. The sampling layer can reduce imbalance in the number of observations between the classes. It can be beneficial because class imbalance together with low size of the minority class can hinder the performance of predictive models. The feature-selection layer allows to reduce the dimensionality of the feature space. A simpler model often translates to better generalizalibity and facilitates interpretation of the features underlying the model. The latter aspect is especially important in medical applications. Finally, the classification layer is responsible for generating predictions based on the observations curated by the first two layers and features selected by the third layer. The selection of the classifier is a critical part of model building, which directly determines the model's flexibility to describe the underlying data. Furthermore, the interpretability of the model depends strongly on the type of the

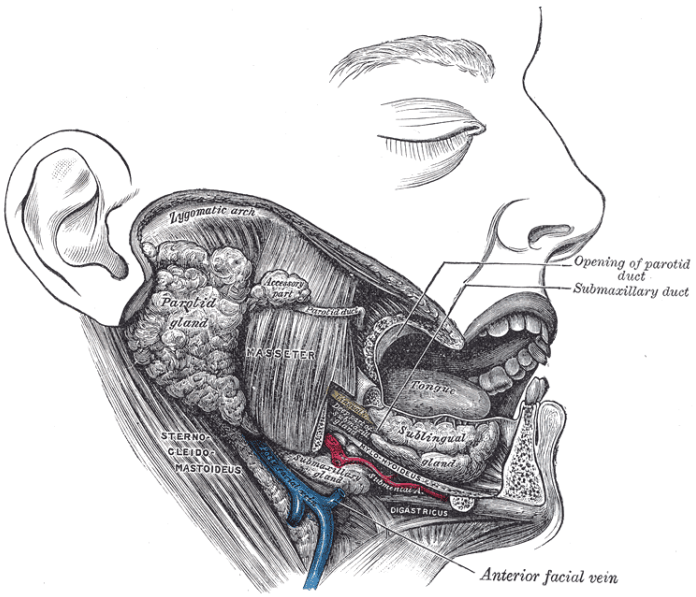



**Figure 1.4:** Historic graphic depicting the salivary glands of right side (Gray 1918). The submandibular gland is marked as submaxillary gland (historical name).

chosen algorithm.

In recent years, there has been a growing interest in the adoption of machine learning methods in NTCP modeling (Gulliford 2015, Kang et al. 2015, El Naqa et al. 2018). Support vector machines were employed to model radiation-induced pneumonitis in Chen et al. (2007). Buettner et al. (2012) showed that the use of Bayesian logistic regression together with features describing shape of the dose within parotid glands allow to predict xerostomia better than NTCP models based on the mean dose. Ospina et al. (2014) built models predicting rectal toxicity following prostate cancer radiotherapy using random forests. Dean et al. (2016) compared the suitability of three machine learning classifiers, namely support vector machines, random forest, and logistic regression, to model acute oral mucositis with spatial dose metrics. Recently, Tseng et al. (2017) presented an application of reinforcement learning with deep neural networks for automated treatment plan adaptation in lung cancer patients.

## 1.2 Radiation-induced xerostomia

One of common side effects of head-and-neck radiotherapy is xerostomia. Xerostomia originates from incidental irradiation of salivary glands during treatment. Radiation impairs salivary gland function, often leading to reduced saliva flow and change in saliva composition. Patients suffering from xerostomia experience difficulties in speaking, chewing, swallowing, dental health, and sleeping. For these reasons, xerostomia often severely reduces patient quality of life after treatment (Wijers et al. 2002).

Three major kinds of salivary glands can be usually distinguished: parotid glands, submandibular glands, and sublingual glands (Figure 1.4). These glands exist in pairs, that is, one on each side of the body. Additionally, many minor salivary glands are distributed throughout the oral cavity (Nanci 2017). The majority of saliva is produced in submandibular glands. These glands are responsible

mainly for unstimulated saliva flow lubricating the mucosal surfaces in the mouth. Their secretion is mostly thick and mucous. Proper functioning of the submandibular glands has a significant impact on the patient's subjective sensation of moisture (Kałużny et al. 2014). The largest salivary glands, that is the parotid glands, contribute 30% of the total salivary volume. These glands are located in front and below of the ear canals on either side of the mouth. The parotid glands are built mainly from serous cells, hence their secretion is watery and rich in proteins. They are especially active on stimulation and their contribution accounts for 50% of the stimulated saliva flow. Sublingual and minor salivary glands provide only a few percent of the total salivary volume.

Most of clinical guidelines focus only on sparing parotid glands during radiotherapy. However, patients with spared parotid glands but damaged submandibular glands may still suffer from xerostomia, especially at night, when the majority of saliva production is unstimulated (Dijkema et al. 2012). Moreover, they may experience constant sensation of dryness in the mouth. It seems that sparing submandibular glands could be beneficial due to the high volume of unstimulated saliva secretion. Unfortunately, submandibular glands are often located close to or within the target volume and receive high doses of radiation. For many patients, a reduction of a dose to submandibular glands could compromise local control.

## 1.3 Thesis overview

Chapter 2 supplies the information on the clinical data used for the study. This includes details on the patient cohort, the follow-up, and the radiotherapy treatment planning data. Chapter 3 is dedicated to an evaluation of the mean dose to parotid glands as a predictor of xerostomia in the collected cohort. The analysis is complemented with the comparison to the results of the PARSPORT clinical trial (Nutting et al. 2011). The focus of chapter 4 is the examination of univariate associations between xerostomia and various features describing parotid shape (radiomics), dose shape (dosiomics), and demographic characteristics. It was investigated whether dosiomic, radiomic, and demographic predictors could provide more precise xerostomia risk assessment than the mean radiation dose to parotid glands. Chapter 5 is dedicated to a comprehensive evaluation of the suitability of seven machine learning classifiers, six feature selection methods, and nine data cleaning/class balancing algorithms for multivariate NTCP modeling of xerostomia. In chapter 6, the results of the external validation on the PARSPORT data are presented. These include comparison of individual predictive power of the analyzed features, performance of the multivariate models, analysis of features underlying well-generalizing models, and examination of differences in the distribution of covariates between the two cohorts. Chapter 7 provides final conclusions of this thesis.

Parts of the work presented in this thesis have been published in peer-reviewed scientific journals. A list of publications is provided on page 117.

CHAPTER 2

# CLINICAL DATA RETRIEVAL AND PREPROCESSING

## 2.1 Patient cohort

In order to study NTCP of radiation-induced xerostomia, a cohort of radiotherapy patients that were at risk of developing this side effect had to be assembled. Consent of the Ethics Committee of Heidelberg University has been acquired to allow patient data extraction from the databases of Heidelberg University Hospital[1]. Patient cases for this study were selected retrospectively, that is at the time of data collection all patients had already been treated according to the protocols of Heidelberg University Hospital.

Radiation-induced xerostomia develops in patients whose salivary glands were damaged by ionizing radiation during the course of radiotherapy. Thus patients with tumors situated close to parotid glands are at the greatest risk of developing this side effect as sparing organs in the vicinity of the tumor is often unfeasible. Accordingly, patient data acquisition started from cases with tumors in oro- and nasopharynx. Later, the search for suitable cases was extended to patients with malignancies situated in locations further from parotid glands, such as hypopharynx, larynx, lips, and brain.

In order to be included in the cohort, patients had to meet a number of criteria. First and foremost, radiotherapy treatment planning data and follow-up documentation with at least one report explicitly confirming or denying xerostomia had to be available. Due to the retrospective character of the data acquisition process, available patients varied in terms of treatment protocol, treatment modalities, and baseline xerostomia. Consequently, further patient inclusion criteria were defined to improve cohort homogeneity and reduce influence of potentially confounding variables on the analysis. Therefore, patients with baseline xerostomia, replanning during the treatment, tumor in the parotid gland, second irradiation, second chemotherapy, or ion beam boost were considered unsuitable.

The final cohort comprised 153 head-and-neck cancer patients who were treated in years 2010–2015 at Heidelberg University Hospital. Patient and tumor characteristics are listed in Table 2.1. The patients' age ranged from 29 to 82 years with median of 61 years. Males constituted 76% of patients. Most of the tumors were located in oropharynx (65%) and hypopharynx/larynx (24%). All patients were treated with intensity-modulated radiotherapy (IMRT) (24%) or helical tomotherapy

[1] Nr. S-392/2016 *Validation and development of probabilistic prediction models for radiation-induced xerostomia*.

**Table 2.1:** Patients and tumor characteristics. Q1 and Q3 correspond to first and third quartiles, respectively.

| | |
|---|---|
| **Total patients** | 153 |
| **Age** | |
| Median | 61 |
| Q1-Q3 | 55–66 |
| Range | 29–82 |
| **Sex** | |
| Female | 37 |
| Male | 116 |
| **Tumor site** | |
| Hypopharynx/Larynx | 37 |
| Nasopharynx | 12 |
| Oropharynx | 99 |
| Other | 5 |
| **Radiation modality** | |
| Intensity-modulated radiotherapy | 37 |
| Tomotherapy | 116 |
| **Ipsilateral parotid mean dose [Gy]** | |
| Median | 24.3 |
| Q1-Q3 | 20.6–27.6 |
| Range | 0.4–63.4 |
| **Contralateral parotid mean dose [Gy]** | |
| Median | 19.9 |
| Q1-Q3 | 15.4–23.1 |
| Range | 0.3–30.9 |

(76%)[2].

## 2.2 Follow-up reports

The extraction of the follow-up data from the clinical databases was a time-consuming and challenging process. Remarks on occurring side effects were usually distributed among many documents and were often mixed with information on patient medication and post-therapy examinations. Furthermore, a lot of follow-up reports had to be rejected due to ambiguous and imprecise statements, such as *xerostomia present*, *xerostomia as usual*, *xerostomia is better/worse*. In total, 693 informative follow-up reports of 153 patients were retrieved.

The frequency of the follow-up reports collection was visualized by plotting the number of follow-up reports against the time after treatment at which the report was written (Figure 2.1). The patients were interviewed, on average, at three-month intervals. The number of evaluations varied among patients; for some patients there was only one evaluation available, while for others there were more than ten. Most often three follow-up reports were available per patient.

### 2.2.1 Toxicity grading systems

Side effects in oncology are usually divided into early and late effects. Early effects, also often called acute effects, become evident during and soon after treatment. They are usually reversible.

[2]To be precise, helical tomotherapy is a type of IMRT (Khan & Gibbons 2014).

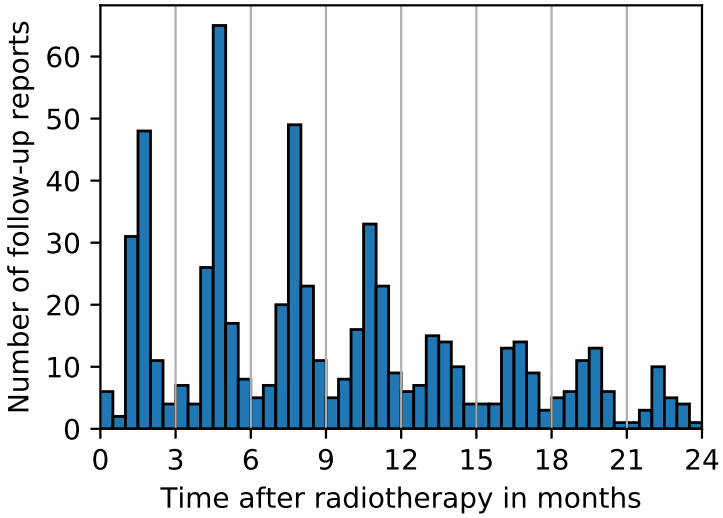



**Figure 2.1:** Frequency of the follow-up reports collection.

Late effects appear several months after therapy, often have a permanent character, and can progress causing severe organ dysfunction or even death (Pavy et al. 1995).

There is no general consensus regarding a cutoff point separating early and late effects. Commonly, it is either 90 days from the onset or from the end of the therapy (Bentzen et al. 2003). Nonetheless, complex and multimodal nature of modern oncology treatments may also motivate alternative definitions (Trotti et al. 2003). Late effects are usually evaluated at 6, 12, and 24 months after therapy.

The severity of side effects in oncological practice is usually assessed using standardized toxicity grading systems. A toxicity grading system provides a list of grades matched to symptoms for specific indications, such as xerostomia, dysphagia, and pneumonitis. Thus each indication can be recognized and graded according to the provided criteria.

Bentzen et al. (2003) distinguished four domains of toxicity measures that are subject to a trade-off between specificity and patient recognition of the side effect (Figure 2.2).

1. The analytic domain covers laboratory tests and quantitative imaging. On the one hand, it is observer-independent and provides the highest specificity of the scores. On the other hand, it often poorly correlates with patient subjective assessment of the adverse effect severity.

2. Objective measures come from physical examinations and simple imaging techniques. They include signs such as edema, weight loss, and swelling.

3. Subjective symptoms are symptoms described by the patient, such as pain, and often translated to medical terms and grades by clinicians.

4. Patient-reported toxicity is collected through standardized quality of life questionnaires evaluating various physical and social functions. Although such measures represent low specificity, they usually correlate best with the patient's perception of quality of life and severity of adverse effects.

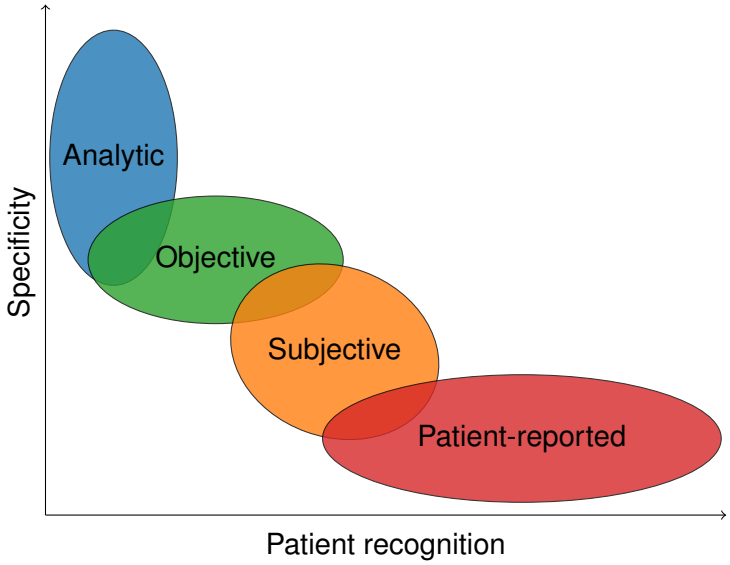



**Figure 2.2:** The trade-off between specificity and patient recognition of adverse effects by domain of toxicity measure. Adapted from Bentzen et al. (2003).

The first toxicity grading systems for oncological treatments were the WHO Handbook for Reporting Results of Cancer Treatment, introduced in 1979 by World Health Organization (World Health Organization and others 1979, Miller et al. 1981), and the Common Toxicity Criteria (CTC), presented in 1983 by National Cancer Institute. Both systems were suitable only for description of acute effects of chemotherapy. The first approach to aid recognition and grading of adverse effects in radiotherapy was realized in 1984 by the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). The RTOG/EORTC introduced two systems: one for assessment of acute effects and one for assessment of late effects of radiotherapy (Cox & Stetz 1995). In 1995, however, due to the emerging shortcomings of the RTOG/EORTC Late Effects System, Late Effects of Normal Tissue (LENT-SOMA) scales were presented (Rubin et al. 1995, Pavy et al. 1995). The LENT-SOMA scales have been found more reliable than RTOG/EORTC grading system for assessment of late effects (Livsey et al. 2002). In 1998, the National Cancer Institute revised the CTC grading system introducing CTC v2.0 (Trotti et al. 2000). This version was expanded to describe adverse effects related to both chemotherapy and radiotherapy. It remained, however, applicable only to acute effects of treatment. In 2003, the third revision of the CTC system took place. The Common Toxicity Criteria were renamed to Common Terminology Criteria for Adverse Events (CTCAE) (Trotti et al. 2003). Many of LENT-SOMA items were incorporated in this version. The two major cornerstones of this revision were a unification of early and late effects criteria onto a single time-independent system and applicability for radiotherapy, chemotherapy, and surgery. Therefore, since the third revision, the CTCAE has been the first comprehensive toxicity grading system applicable for recognition and assessment of both early and late adverse effects of multimodal oncology treatments. More information on the current and previous CTC and CTCAE versions can be found on the National Cancer Institute website[3].

The CTCAE grading system takes into account analytic measures, objective signs, and subjective clinician-graded symptoms (Trotti et al. 2007). The grades are represented by numbers ranging from

[3] https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm

**Table 2.2:** Common Terminology Criteria for Adverse Effects v4.03 for dry mouth (xerostomia) (National Cancer Institute 2010). Grades 4 and 5 are not defined for dry mouth.

| Grade 1 | Grade 2 | Grade 3 |
|---|---|---|
| - symptomatic (e.g., dry or thick saliva)<br>- no significant dietary alterations<br>- unstimulated saliva flow > 0.2 ml/min | - moderate symptoms<br>- oral intake alterations (e.g., copious water, lubricants, diet limited to purees and soft, moist foods)<br>- unstimulated saliva flow 0.1–0.2 ml/min | - inability to adequately aliment orally<br>- tube feeding or total parenteral nutrition<br>- unstimulated saliva flow < 0.1 ml/min |

one to five with increasing severity of the side effect

- **Grade 1** Mild complications not requiring medical intervention.
- **Grade 2** Moderate complications with minimal noninvasive medical intervention.
- **Grade 3** Severe side effects that often require hospitalization but are not life-threatening.
- **Grade 4** Life-threatening complications.
- **Grade 5** Death related to side effects.

### 2.2.2 Toxicity grading in our cohort

In 74% of the follow-up reports either CTCAE v3.0 or CTCAE v4.03 were used to assess severity of radiation-induced xerostomia. Dry mouth (xerostomia) grading definitions are word-for-word identical in both versions, so no inconsistency was introduced by using both versions. In case no score was provided in the report but descriptive toxicity information was available, appropriate grade was assigned after consulting Heidelberg University Hospital clinicians. In order to minimize intra- and inter-observer variability in this process, a set of rules in a form of a dictionary was introduced. The definitions of CTCAE grades for xerostomia are presented in Table 2.2.

## 2.3 Radiotherapy treatment planning data

The treatment planning data contains information about patient anatomy, structures defined during treatment planning, radiation beam orientations, and the planned dose distribution. It was necessary to extract this data in order to match the treatment plan parameters with treatment outcomes available in the follow-up reports.

### 2.3.1 The DICOM standard

Radiotherapy treatment planning data is typically stored in the Digital Imaging and Communications in Medicine (DICOM) files. The DICOM file format is a part of a widely adopted standard for the management of medical imaging information and related data. Apart from the file format, the DICOM standard specifies syntax, semantics, and a set of protocols that facilitate communication and interoperability of medical equipment.

Every DICOM file contains information that allows to match files to a specific patient, study, series, and equipment. The data is organized in modules that consist of attributes, for example, the patient module contains attributes that describe and identify the patient. These attributes could be the patient's name, ID, and date of birth. Similarly, the equipment module comprise attributes that describe the device that produced the data, such as the manufacturer and the device serial number.

The DICOM standard specifies many information objects suitable for handling various kinds of data. There are objects dedicated to managing data generated by imaging devices, such as computed radiography (CR), computed tomography (CT), and magnetic resonance (MR). Additionally, the DICOM standard defines information objects for handling tasks specific to radiotherapy, for example, portal images (RT Image) and treatment plan realization (RT Plan).

A comprehensive description of all DICOM information objects, modules, and attributes is beyond the scope of this thesis. In this work, only the elements of the DICOM standard that are most relevant for this study are discussed. For a complete reference, the DICOM standard documentation is available at http://medical.nema.org.

### 2.3.2 Data extraction

In order to extract treatment plan features, the planned dose distribution and the definitions of organs at risk were needed. The information about the dose distribution calculated by radiotherapy treatment planning systems was stored in DICOM RT Dose files, typically as a three-dimensional array. Apart from the dose distribution itself, the RT Dose files provided information about patient position, patient orientation, pixel spacing, position of slices, and dose units. The dose distribution, however, was of little use without definitions of the patient's anatomical and treatment-related structures. These were stored in DICOM RT Structure Set files as coordinates of polyhedra vertices, which represented the defined structures. The CT images stored in DICOM CT Image files were collected whenever possible. They were used, however, only for a visual check of the correct alignment between the structure definitions and dose distribution as they were not required for the extraction of the treatment plan features.

The DICOM files were exported either from the hospital picture archiving and communication system (PACS) or directly from the treatment planning stations. The process of data retrieval was slow and time consuming. The DICOM files of a single patient were often distributed among several databases. The location and availability of the data depended on the time of the patient's treatment and used modalities. Furthermore, the access to the treatment planning stations from which some patients had to be manually exported was limited due to routine clinical work.

### 2.3.3 Data preprocessing

After the retrieval, the DICOM files were preprocessed to accommodate them for the subsequent analysis. First, all files were pseudonymized, that is each patient was assigned a new identification number and all sensitive data was removed. Next, for each patient, DICOM RT Dose, DICOM RT Structure Set, and DICOM CT Image files were imported to MATLAB. The CT and dose cubes were extracted and interpolated to an isotropic resolution of 1 mm. A logical mask was generated for every structure defined in the DICOM RT Structure Set files. The masks were of the same size and

spacing as the dose cubes and indicated which voxels belong to a structure. In the last preprocessing step, the CT and dose cubes as well as the logical masks were saved as binary file, one per each patient. This way, the data could be easily read to MATLAB for visualization or feature extraction without the need for reimporting the DICOM files. The implementation of DICOM import and data preprocessing is presented in Appendix A.

For the purpose of loading, preprocessing, visualization, and feature extraction from the DICOM files, a collection of scripts was written in MATLAB. Together with the scripts, a graphical user interface (GUI) was developed which facilitates use of the toolbox. The toolbox is publicly available on GitHub[4].

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

CHAPTER 3

# MEAN DOSE TO PAROTID GLANDS AND XEROSTOMIA

*The results presented in this chapter were published in Gabryś et al. (2017). Part of section 3.3 has been quoted verbatim from Gabryś et al. (2017).*

Current clinical guidelines limiting dosage to parotid glands were formulated in 2010 by the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) group. The QUANTEC initiative aimed to provide simple yet reliable dose-volume constraints to reduce risk of side effects in radiotherapy treatments (Bentzen et al. 2010). After meta-analysis of available studies, Deasy et al. (2010) concluded that sparing at least one parotid gland to a mean dose less than 20 Gy or both parotid glands to a mean dose less than 25 Gy *usually* allows to avoid severe xerostomia. After publication of QUANTEC guidelines, a number of large-cohort studies confirmed that the mean dose well describes the dose-response relationship of the parotid gland (Houweling et al. 2010, Beetz et al. 2012).

In the following years, it has been observed that the mean-dose constraints indeed allowed to reduce prevalence of xerostomia in cohorts where the majority of patients had met the QUANTEC guidelines. At the same time, however, the mean dose was often not predictive of xerostomia any more (Buettner et al. 2012, Lee et al. 2015). Since the introduction of the QUANTEC recommendations, radiotherapy techniques improved on conformity of target dose coverage and reduction of the overall dose to organs at risk. Consequently, current treatment plans often manage to spare both parotid glands to a mean dose of 20–25 Gy (Leung & Lee 2013, Lee et al. 2015), whereas a typical mean dose to parotid glands of a head-and-neck cancer patient in 1996–2007 was usually in the range of 30–40 Gy (Braam et al. 2006, Dijkema et al. 2008).

One of prospective studies investigating the relation between parotid sparing and incidence of xerostomia was the PARSPORT clinical trial (Nutting et al. 2011). PARSPORT was designed to demonstrate the difference in the proportion of patients treated with parotid-sparing IMRT and conventional radiotherapy. The study showed reduced prevalence of late xerostomia in the IMRT group, however the mean dose to parotid glands was not predictive of patients who developed this side effect (Buettner et al. 2012).

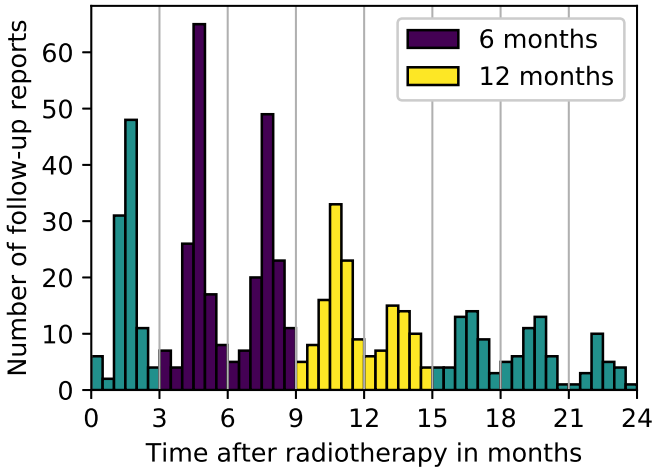



**Figure 3.1:** Frequency of the follow-up reports collection.

This chapter is dedicated to an analysis of the relationship between the mean dose to the parotid glands and xerostomia in patients treated at Heidelberg University Hospital. The analysis was carried out in context of the results of the PARSPORT trial, therefore the cohort of patients treated in Heidelberg will be referred to form now on as the HD cohort. The PARSPORT trial was chosen as the reference study for a number of reasons. First, it was a prospective clinical trial. Data from such studies usually represents higher quality than data from retrospective studies. Second, the trial was a multi-center study which allowed to reduce center-specific bias stemming from personnel or equipment factors. Last, one of the grading scales used in the study was LENT-SOMA which is similar to CTCAE grading scale available in the HD cohort. It was examined how dosage to parotid glands differ between PARSPORT and HD cohort and whether a mean-dose model built on PARSPORT data could predict prevalence of xerostomia in the HD cohort. Subsequently, the reliability of the mean dose to parotid glands as a predictor of late xerostomia in the HD cohort was investigated. For this purpose, the LKB model was used.

## 3.1 Material and Methods

### 3.1.1 Endpoints

The analysis focused on xerostomia evaluated at 6 and 12 months after radiotherapy. These time points are often used in studies of late xerostomia. Due to the time characteristic of the follow-up report creation, six-month-wide windows around each of these time points were applied and the reports within them were pooled (Figure 3.1). The toxicity score was calculated as the arithmetic mean rounded to the nearest integer with x.5 being rounded up. Furthermore, variability of the scores given to a certain patients was analyzed. It was measured as an average percentage of reports with xerostomia grades different than the final grade. At each of the time points, two xerostomia endpoints were analyzed:

1. Mild-to-severe xerostomia quantified by CTCAE grade 1 and higher (G1+),

2. Moderate-to-severe xerostomia quantified by CTCAE grade 2 and higher (G2+).

Patient and tumor characteristics stratified by endpoint are summarized in Table 3.1.

### 3.1.2 Prevalence of xerostomia

#### Tomotherapy versus IMRT

In order to investigate whether there was a statistically significant difference in xerostomia rates between the patients treated with either of these modalities, Barnard's test was used (Barnard 1945, 1947). Barnard's test examines association of two categorical variables and is uniformly more powerful than Fisher's exact test (Fisher 1925) for 2x2 contingency tables. A straightforward comparison of the two tests is provided in Mehta & Senchaudhuri (2003). The following comparisons were examined: G0 versus G1, G0 versus G2, and G0–1 versus G2 at 6 and 12 months after treatment. This resulted in six tests in total.

There are two kinds of errors associated with every hypothesis test: type I and type II error. A type I error happens when true hypothesis is rejected. It is also often called false positive or false discovery. A type II error, on the other hand, occurs when a false hypothesis is not rejected. It is also known as false negative.

It is important to realize that performing multiple comparisons increases the probability of reporting falsely positive tests. This is because each test has a probability of generating a type I error and performing any additional test increases the probability that at least one of the true hypotheses will be rejected.

Two most common approaches to control the number of type I errors in multiple testing are the family-wise error rate (FWER) and the false discovery rate (FDR). Here, the Holm-Bonferroni method was used to control the FWER of the Barnard's tests. The significance level $\alpha$ for the FWER was set to 0.05.

**Family-wise error rate** Say $m$ hypotheses $H_1, H_2, \ldots, H_m$ are tested and of these $m_0$ are true. Also, with each hypothesis there is an associated p-value $p_1, \ldots, p_m$. The FWER is then defined as the probability of reporting at least one false discovery, that is

$$\mathrm{FWER} = P(V > 0)\,, \tag{3.1}$$

where $V$ is the number of falsely rejected hypotheses.

The most well-known method to control the FWER is the Bonferroni method. In order to apply it, one compares each p-value not to the significance level $\alpha$ but to $\alpha/m$. The Bonferroni method is very appealing because it is easy to use and controls the FWER in the strong sense under any dependence structure of the underlying p-values. It is, however, overly conservative (Goeman & Solari 2014).

**Table 3.1:** Patients and tumor characteristics. The total number of patients differ among the groups due to the follow-up availability.

| | | 6 months | | | 12 months | | |
|---|---|---|---|---|---|---|---|
| | All patients | Grade 0 | Grade 1 | Grade 2 | Grade 0 | Grade 1 | Grade 2 |
| **Total patients** | 153 | 18 | 93 | 27 | 19 | 68 | 15 |
| **Age** | | | | | | | |
| Median | 61 | 61 | 61 | 61 | 60 | 61 | 62 |
| Q1–Q3 | 55–66 | 56–66 | 54–66 | 54–68 | 56–66 | 54–67 | 55–68 |
| Range | 29–82 | 49–78 | 29–82 | 43–80 | 49–79 | 34–82 | 43–76 |
| **Sex** | | | | | | | |
| Female | 37 | 5 | 21 | 6 | 6 | 16 | 5 |
| Male | 116 | 13 | 72 | 21 | 13 | 52 | 10 |
| **Tumor site** | | | | | | | |
| Hypopharynx/Larynx | 37 | 9 | 19 | 5 | 5 | 17 | 1 |
| Nasopharynx | 12 | 0 | 9 | 1 | 2 | 4 | 1 |
| Oropharynx | 99 | 8 | 63 | 19 | 12 | 46 | 11 |
| Other | 5 | 1 | 2 | 2 | 0 | 1 | 2 |
| **Radiation modality** | | | | | | | |
| IMRT | 37 | 2 | 29 | 4 | 3 | 22 | 4 |
| Tomotherapy | 116 | 16 | 64 | 23 | 16 | 46 | 11 |
| **Ipsilateral parotid mean dose** | | | | | | | |
| Median | 24.3 | 20.0 | 25.0 | 24.4 | 22.9 | 24.8 | 24.1 |
| Q1–Q3 | 20.6–27.6 | 17.4–23.4 | 21.9–28.8 | 21.5–26.4 | 17.7–24.8 | 21.4–30.0 | 22.4–26.4 |
| Range | 0.4–63.4 | 0.4–34.4 | 4.6–61.4 | 11.2–63.4 | 0.4–27.8 | 4.6–61.4 | 17.3–63.4 |
| **Contralateral parotid mean dose** | | | | | | | |
| Median | 19.9 | 16.3 | 21.3 | 19.9 | 15.5 | 20.1 | 20.3 |
| Q1–Q3 | 15.4–23.1 | 11.5–19.6 | 16.1–24.3 | 16.8–23.1 | 7.5–21.4 | 16.6–23.8 | 18.0–23.4 |
| Range | 0.3–30.9 | 0.3–22.1 | 4.1–30.9 | 5.2–26.2 | 0.3–27.5 | 4.1–28.7 | 12.7–26.1 |

To understand the *problem* with the Bonferroni correction, let us analyze the reasoning behind it. The Bonferroni method is based on the Boole's inequality given by

$$P\left(\bigcup_{i=1}^{k} E_i\right) \leq \sum_{i=1}^{k} P\left(E_i\right) \tag{3.2}$$

for any set of events $E_1, \ldots, E_k$. If $q_1, \ldots, q_{m_0}$ are the p-values of the true hypotheses, it follows from Boole's inequality that

$$\mathrm{FWER} = P\left(\bigcup_{i=1}^{m_0}\left\{q_i \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{i=1}^{m_0} P\left(q_i \leq \frac{\alpha}{m}\right) \leq m_0 \frac{\alpha}{m} \leq \alpha\,. \tag{3.3}$$

This relation shows that the Bonferroni method does not only bound the significance level $\alpha$ but also $m_0\alpha/m$. For this reason, the Bonferroni correction is especially conservative when there are many false hypotheses.

Holm (1979) presented a uniformly more powerful algorithm, later called the Holm-Bonferroni method. Similarly to the Bonferroni correction, it provides strong control of the FWER and does not require assumptions on the p-values dependence. First, the p-values are sorted in increasing order. Then, the maximum index $k$ is found such that

$$k = \max\left\{1 \leq i \leq m : p_i \leq \frac{\alpha}{m - i + 1}\right\} \tag{3.4}$$

and hypotheses $H_1, \ldots, H_k$ are rejected. If no such $k$ exists, no hypothesis is rejected.

There were also other methods developed that have even greater power than the Holm-Bonferroni method, for example, Hochberg's step-up procedure (Hochberg 1988). They require, however, assumptions on the joint distribution of the test statistics, which limit their applicability.

#### Expected incidence of xerostomia

NTCP models predict complication probability for a patient based on a features vector associated with this patient and model parameters estimated during model training. In case of mean dose models of xerostomia, the feature vector is usually a mean dose to both parotid glands or a mean dose to a contralateral parotid gland. Having NTCP probability scores for all patients in a cohort, one can calculate the expected prevalence of a side effect using the following formula

$$\overline{\mathrm{NTCP}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{NTCP}(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{3.5}$$

where $N$ is the number of patients in the cohort, $\boldsymbol{x}_i$ is a feature vector of the $i$th patient, and $\boldsymbol{\theta}$ is the model's parameterization.

In order to investigate whether the observed xerostomia rates in the HD cohort could have been predicted by an external NTCP model based on a mean dose to parotid glands, an

NTCP model trained on the PARSPORT data was used. The model published by Miah et al. (2013) was the LKB model based on the mean dose to the contralateral parotid gland. It was trained to predict LENT-SOMA subjective G2+ xerostomia at 12 months after treatment. The model training resulted in the following parameterization: $d_{50}$ = 28.7 Gy, $m$ = 0.20, $n$ = 1. The $n$ parameter value was not estimated but was fixed *a priori* at 1.

### 3.1.3 Model training

The model architecture chosen to build mean-dose models for the HD cohort was the LKB model (discussed in section 1.1.1) with parameterization proposed by Mohan et al. (1992)

$$\mathrm{NTCP} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx \,, \tag{3.6}$$

where

$$t = \frac{d_{eud} - d_{50}}{m \cdot d_{50}} \tag{3.7}$$

and

$$d_{eud} = \left( \sum_{i=i}^{N} v_i d_i^{1/n} \right)^n \,. \tag{3.8}$$

The LKB model was chosen due to its popularity in NTCP modeling and straightforward comparison of model parameters to the values reported in other studies.

In this study, the value of $n$ parameter was fixed at 1, which is commonly done in studies on xerostomia toxicity (Deasy et al. 2010). With this parameterization, the equivalent uniform dose $d_{eud}$ is equivalent to the mean dose. Due to a low number of observations in the near zero-dose region, a zero-dose correction was applied (Braam et al. 2005). Thus, artificial zero-dose dose-volume histograms were added to ensure maximal NTCP of 5% for a zero dose for every model.

As explained previously, the models were build for late xerostomia evaluated at 6 and 12 months after treatment for G1+ and G2+ severity. The models were based either on the mean dose to the contralateral parotid gland or the mean dose to both parotid glands weighted by parotid volumes. This resulted in eight models in total.

#### Maximum likelihood estimation

The parameters $\{d_{50}, m\}$ of the models were fitted using maximum likelihood estimation (MLE) defined by

$$\{\hat{d}_{50}, \hat{m}\} = \arg\max_{d_{50},m} \ln \mathcal{L}(d_{50}, m | d_{eud}, y) \,, \tag{3.9}$$

where the log-likelihood $\log \mathcal{L}$ is given by

$$\ln \mathcal{L}(d_{50}, m | d_{eud}, y) = \sum_{c=0}^{1} \sum_{i:y_i=c} \ln \mathrm{NTCP}(d_{eud,i} | d_{50}, m) \tag{3.10}$$

and

$$y_i = \begin{cases} 1 & \text{if } i \text{ is a positive case,} \\ 0 & \text{if } i \text{ is a negative case.} \end{cases} \tag{3.11}$$

Maximum likelihood estimation is a widely used method of estimating parameters of statistical models. It is based on the idea that the most reasonable values of the model parameters are those for which the probability of the observed data is largest (Hastie et al. 2009). It is often more convenient to work with the natural logarithm of the likelihood called log-likelihood, which does not change values of the estimated parameters as the log-likelihood assumes maximum values for the same values as the likelihood function.

An advantage of the MLE is that the likelihood function captures all the information about a certain parameter in the data, including its uncertainty. The equation 3.9 is realized by finding parameter values for which partial derivatives of the log-likelihood are zero

$$\nabla \ln \mathcal{L}(d_{50}, m) = \mathbf{0}\,. \tag{3.12}$$

At the log-likelihood maximum, the second derivative of the function is negative. Therefore the curvature of the log-likelihood function can be expressed as

$$\mathbf{I}(d_{50}, m) = -\nabla^2 \ln \mathcal{L}(d_{50}, m)\,, \tag{3.13}$$

where $\mathbf{I}$ is the observed Fisher information. Large values of $\mathbf{I}$ correspond to steep peak of the log-likelihood, hence low uncertainty about the estimate. The Fisher information is an estimator of the asymptotic covariance matrix of the parameters

$$\mathrm{cov}(\hat{d}_{50}, \hat{m}) = \mathbf{I}^{-1}(\hat{d}_{50}, \hat{m})\,. \tag{3.14}$$

The curvature of the log-likelihood can be often approximated by the normal distribution

$$\{\hat{d}_{50}, \hat{m}\} \approx \mathcal{N}((d_{50}, m), \mathbf{I}^{-1}(\hat{d}_{50}, \hat{m}))\,. \tag{3.15}$$

For that reason, the standard errors SE of the parameters are given by the square roots of the diagonal elements of the covariance matrix

$$\mathrm{SE}(\hat{d}_{50}) = \sqrt{\mathbf{I}^{-1}_{d_{50},d_{50}}}\,, \tag{3.16a}$$

$$\mathrm{SE}(\hat{m}) = \sqrt{\mathbf{I}^{-1}_{m,m}}\,. \tag{3.16b}$$

It follows that the confidence intervals CI for the parameters can be calculated as

$$\mathrm{CI}(\hat{d}_{50}; C) = \hat{d}_{50} \pm \Phi^{-1}\left(\frac{1-C}{2}\right)\mathrm{SE}(\hat{d}_{50})\,, \tag{3.17a}$$

$$\mathrm{CI}(\hat{m}; C) = \hat{m} \pm \Phi^{-1}\left(\frac{1-C}{2}\right)\mathrm{SE}(\hat{m})\,. \tag{3.17b}$$

**Table 3.2:** Confusion matrix. TP - true positives, FN - false negatives, FP - false positives, TN - true negatives.

| | Positive cases | Negative cases | |
|---|---|---|---|
| Positive predictions | TP | FP | Precision = $\frac{TP}{TP+FP}$ |
| Negative predictions | FN | TN | |
| | Sensitivity = $\frac{TP}{TP+FN}$ | Specificity = $\frac{TN}{FP+TN}$ | |

where $\Phi^{-1}$ marks the probit function[1] and $C$ is the confidence level. The topic of likelihood-based modeling is extensively covered in Pawitan (2001).

### 3.1.4 Model evaluation

The LKB model is a continuous classifier, which means it does not classify patients as positive or negative cases. It rather gives a probability that a certain patient will develop a side effect. One way to evaluate performance of such a model is to select a threshold, for example a mean dose of 20 Gy, and transform in this way a continuous model into a binary model. This would allow to calculate a fraction of patients correctly classified as patients who will develop the side effect (sensitivity) and a fraction of patients correctly classified as patients who will not develop the side effect (specificity) (Table 3.2). The downside of this approach is that selection of a different threshold would result in a different sensitivity and specificity. As a result, there is no particular value of sensitivity or specificity that would characterize the overall accuracy of a model, but rather an entire range of values that would vary depending on what was used as the decision threshold. The trade-off between the sensitivity and the specificity due to the decision threshold selection can be visualized with the receiver operating characteristic (ROC) curve. The ROC curve is generated by plotting sensitivity versus 1-specificity over the entire range of threshold values allowing a model-wide evaluation.

#### Area under the ROC curve

The most commonly used index related to ROC curves is the area under the curve (AUC). Intuitively, it can be interpreted as the probability that a randomly chosen patient from the positive group will be assigned a higher score than a randomly chosen patient from the negative group. It can also be interpreted as the average sensitivity over all values of specificity or, analogously, the average specificity over all values of sensitivity. A perfect model has an AUC of 1.0, whereas a random classifier gives an AUC of 0.5. The predictive performance of the models in this chapter was measured with the AUC. For brevity, $\theta$ represents the AUC in all equations in this chapter.

It has been shown that the AUC calculated with the trapezoidal rule is directly related to the Mann-Whitney $U$ statistic (Bamber 1975). Let us assume the patients who developed a complication form group A, whereas the patients who did not develop a complication form

[1] Probit function is the inverse of the cumulative distribution function of the standard normal distribution.

group B, that is

$$A = \{a_1, a_2, \ldots, a_m\}\,, \tag{3.18a}$$
$$B = \{b_1, b_2, \ldots, b_n\}\,, \tag{3.18b}$$

where the $a_i$ and $b_j$ are the model predictions for the individual patients. Then, the AUC estimate $\hat{\theta}$, can be calculated according to the trapezoidal rule as

$$\hat{\theta} = \frac{1}{mn} \sum_{i}^{m} \sum_{j}^{n} \psi(a_i, b_j)\,, \tag{3.19}$$

where $\psi$ is given by

$$\psi = \begin{cases} 1 & \text{if } a_i > b_j\,, \\ 0.5 & \text{if } a_i = b_j\,, \\ 0 & \text{if } a_i < b_j\,. \end{cases} \tag{3.20}$$

#### Confidence intervals for the AUC

Confidence intervals for the AUC can be calculated in a number of ways, such as the normal approximation, Mann-Whitney $U$ statistic, logit-transformation, and the DeLong method. A number of studies showed that usually confidence intervals closest to the true unknown confidence intervals can be estimated with the bootstrap (Obuchowski & Lieber 1998). The bootstrap is a nonparametric approach, where one generates a large number, typically thousands, of samples by sampling with replacement from the original data set. The AUCs are estimated separately for each sample, which results in a distribution of AUC estimates marked as $\hat{\Theta}^b$.

The most straightforward way to calculate confidence intervals in such setting is to use percentiles of the bootstrap distribution. For example, for the confidence level $C = 1 - 2\alpha$, the confidence intervals are given by

$$\mathrm{CI}(\hat{\theta}; C) = (\hat{\Theta}^b_{(\alpha)}, \hat{\Theta}^b_{(1-\alpha)})\,. \tag{3.21}$$

The limitation of this approach is that it does not take into account potential bias and skewness of the $\hat{\Theta}^b$ distribution. By the bias is meant the discrepancy between the median value in $\hat{\Theta}^b$ and the the $\hat{\Theta}$. The skewness, in turn, stems from the fact that the standard error of the AUC estimate with respect to the true value of the AUC is not constant. In other words, the standard normal approximation $\hat{\theta} \propto \mathcal{N}(\theta, \sigma^2_{\hat{\theta}})$ assumes that the standard error of $\hat{\theta}$ is the same for all values of $\theta$, which is often unrealistic.

For that reason, in this study the confidence intervals of the AUC estimates were calculated with the bootstrap using the bias-corrected and accelerated (BCa) method (Efron & Tibshirani 1993). The BCa interval endpoints are given by percentiles of the bootstrap

distribution

$$\mathrm{CI}(\hat{\theta}; C) = (\hat{\Theta}^b_{(\alpha_L)}, \hat{\Theta}^b_{(\alpha_H)}) \,, \tag{3.22}$$

where $\hat{\Theta}^b$ is a set of AUC bootstrap estimates, $\alpha_L$ is the lower percentile, and $\alpha_H$ is the higher percentile. The percentile values are given by

$$\alpha_L = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(\alpha)}{1 - \hat{a}(\hat{z}_0 + \Phi^{-1}(\alpha))}\right) , \tag{3.23a}$$

$$\alpha_H = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(1-\alpha)}{1 - \hat{a}(\hat{z}_0 + \Phi^{-1}(1-\alpha))}\right) , \tag{3.23b}$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution, $\Phi^{-1}$ is the probit function, $\hat{z}_0$ is the bias-correction coefficient, and $\hat{a}$ is the acceleration coefficient. The bias-correction coefficient, defined by

$$\hat{z}_0 = \Phi^{-1}\left(\frac{|\{\hat{\Theta}^b < \hat{\theta}\}|}{|\hat{\Theta}^b|}\right) , \tag{3.24}$$

measures the discrepancy between the median and the mean value of the AUC estimates. In other words, it compares the number of bootstrapped AUCs that are lower than the estimated AUC to the number of all bootstrap samples, and quantifies skewness of the bootstrap distribution via the probit function. The acceleration coefficient $\hat{a}$ is defined by

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6(\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2)^{3/2}} \,, \tag{3.25}$$

where $\hat{\theta}_{(i)}$ is the $i$th Jackknife replica, that is the AUC estimate based on the original sample with the $i$th observation removed, and $\hat{\theta}_{(\cdot)}$ is the average of Jacknife replicas given by

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \,. \tag{3.26}$$

The acceleration coefficient $\hat{a}$ describes the rate of change of the standard error of the AUC estimate with respect to the true value of the AUC. More details on the BCa method can by found in Efron (1987).

The BCa bootstrap was chosen for the confidence intervals estimation because it does not require data distribution assumptions. Moreover, it has been shown to outperform other methods, especially in small imbalanced sample settings (Obuchowski & Lieber 1998, Lasko et al. 2005).

#### Comparison of AUCs

The statistical significance of the differences between the AUCs were estimated using the DeLong method (DeLong et al. 1988, Robin et al. 2011). The DeLong method is a non-

parametric asymptotic approach which means that it makes no parametric assumptions of the data. An intuitive explanation of the algorithm can be found in Hanley & Hajian-Tilaki (1997). The method transforms each model prediction value to a placement value. The placements values are given by

$$V_{Ai} = \frac{1}{n} \sum_{j=1}^{n} \psi(a_i, b_j) \,, \tag{3.27a}$$

$$V_{Bj} = \frac{1}{m} \sum_{i=1}^{m} \psi(a_i, b_j) \,. \tag{3.27b}$$

Intuitively, the placement value for a measurement from group A is its percentile among the values of group B. The variance of the AUC estimate is then given by

$$\mathrm{var}[\hat{\theta}] = \frac{\mathrm{var}[V_A]}{m} + \frac{\mathrm{var}[V_B]}{n} \,. \tag{3.28}$$

In order to compare AUCs of two ROC curves generated by two different models using the same data set, one needs to calculate the difference between the areas as well as the variance of that difference. Because both models used the same data, the ROC curves are correlated. Therefore, the variance of the difference between two AUC estimates defined as

$$\mathrm{var}[\hat{\theta}^\alpha - \hat{\theta}^\beta] = \mathrm{var}[\hat{\theta}^\alpha] + \mathrm{var}[\hat{\theta}^\beta] - 2\,\mathrm{cov}[\hat{\theta}^\alpha, \hat{\theta}^\beta] \,, \tag{3.29}$$

must comprise a covariance term given by

$$\mathrm{cov}[\hat{\theta}^\alpha, \hat{\theta}^\beta] = \frac{\mathrm{cov}[V_A^\alpha, V_A^\beta]}{m} + \frac{\mathrm{cov}[V_B^\alpha, V_B^\beta]}{n} \,. \tag{3.30}$$

As a result, the statistical significance of a difference between two AUC estimates can be calculated from the following statistic

$$\frac{\hat{\theta}^\alpha - \hat{\theta}^\beta}{\sqrt{\mathrm{var}[\hat{\theta}^\alpha - \hat{\theta}^\beta]}} \,, \tag{3.31}$$

which follows the standard normal distribution.

In addition to the ROC curves, the precision-recall (PR) curves were calculated. The PR curves demonstrate interdependence of precision and recall[2] on a decision threshold. Similarly to the ROC curves, the PR curves allow a model-wide evaluation. Moreover, they can expose differences between models that are not apparent in the ROC plots (Davis & Goadrich 2006). Whereas the ROC curves show ability of a model to recognize positive (sensitivity) and negative (specificity) cases, the PR curves inform how sensitivity of a model affects reliability of positive predictions (precision) and vice versa.

[2]Recall is another term for sensitivity.

**Table 3.3:** Variability of xerostomia grades.

| Time point | Grade 0 | Grade 1 | Grade 2 |
|---|---|---|---|
| 6 months | 0.0% | 3.4% | 32.7% |
| 12 months | 0.0% | 2.6% | 30.0% |

## 3.2 Results

The analysis showed no variability in follow-up reports of patients who did not develop xerostomia and little variability in follow-up reports of patients categorized as xerostomia G1. For G2 patients, however, the variability was rather high reaching over 30%, both at 6 and 12 months after treatment. In other words, almost every third follow-up document reported a grade different than grade 2. The variability of xerostomia grades is summarized in Table 3.3.

There was no significant difference in xerostomia prevalence between the IMRT and the tomotherapy groups at the FWER = 0.05. The comparison involved G0 versus G1, G0 versus G2, and G0-1 versus G2 at 6 and 12 months after treatment. The observed xerostomia toxicity rates stratified by endpoint are presented in Table 3.1.

The analysis of the mean doses to parotid glands revealed that 52% of patients in the HD cohort had their contralateral parotid glands spared to a mean dose lower than 20 Gy. Moreover, 72% of patients received an average mean dose to both parotid glands lower than 25 Gy. At least one QUANTEC constraint was satisfied in 74% of patients. Figure 3.2 shows the distribution of ipsilateral and contralateral mean doses among G0, G1, and G2 patients.

A comparison of the mean dosage to the parotid glands showed that the average mean dose to the contralateral glands in the HD group was lower than in the PARSPORT cohort (18.7 Gy versus 24.9 Gy). Furthermore, the difference in the average mean dose to the ipsilateral glands between the two cohorts was even more pronounced (25.4 Gy versus 45.7 Gy). A visual comparison of the parotid gland mean dose distributions between the two cohorts is provided in Figure 3.3.

The difference in the average mean doses to parotid glands corresponded to a difference in the observed xerostomia rates. The incidence of xerostomia at 12 months in the PARSPORT cohort was 38% of LENT-SOMA subjective G2+ compared to 15% of CTCAE G2+ in the HD cohort. The low incidence in the HD cohort was also predicted by the LKB model trained on the PARSPORT data. Figure 3.4 shows the NTCP curve produced by this model and the observations from the HD cohort. The expected prevalence of LENT-SOMA subjective G2+ at 12 months in the HD cohort predicted by this model was estimated as 11%.

The parameters of the fitted models together with 95% confidence intervals are presented in Table 3.4. For the G1+ xerostomia models based on the mean dose to the contralateral parotid gland, $d_{50}$ increased slightly from 10.7 Gy to 12.0 Gy when moving from 6 to 12 months time point. A similar time point-dependent shift of $d_{50}$ was observed for the

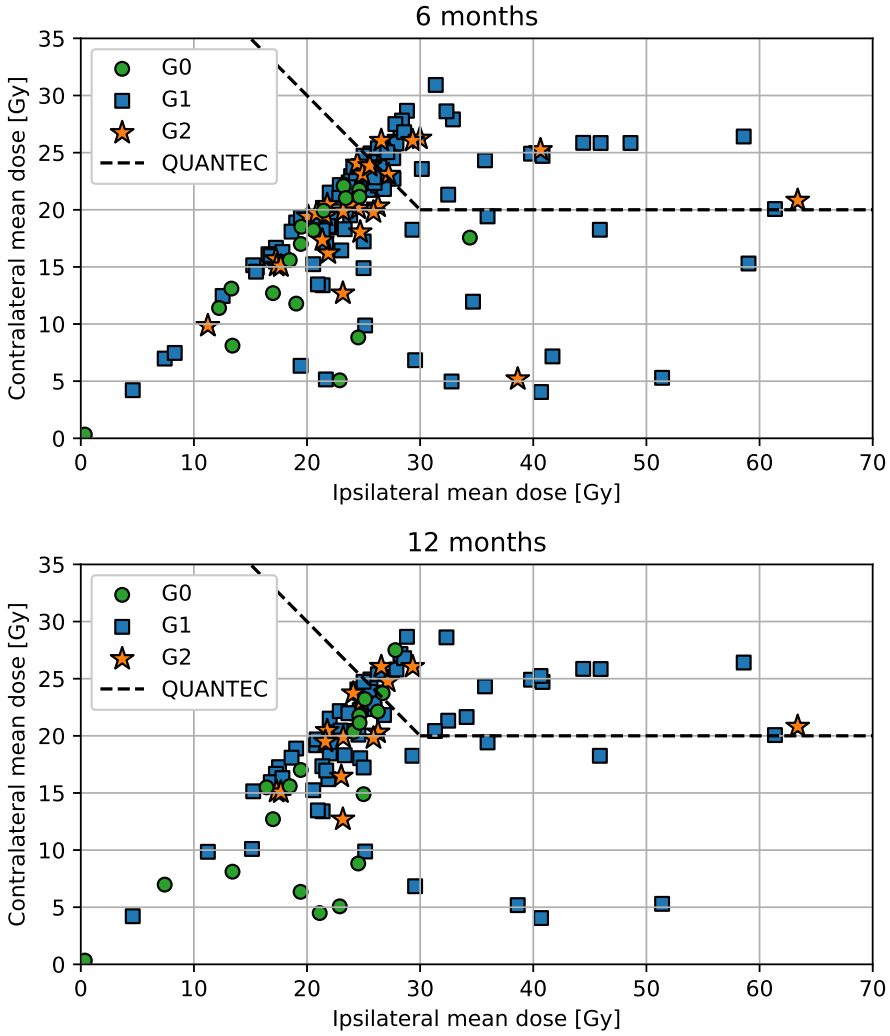



**Figure 3.2:** Ipsi- and contralateral mean dose to parotid glands and corresponding xerostomia grades at 6 and 12 months after treatment in the HD cohort.

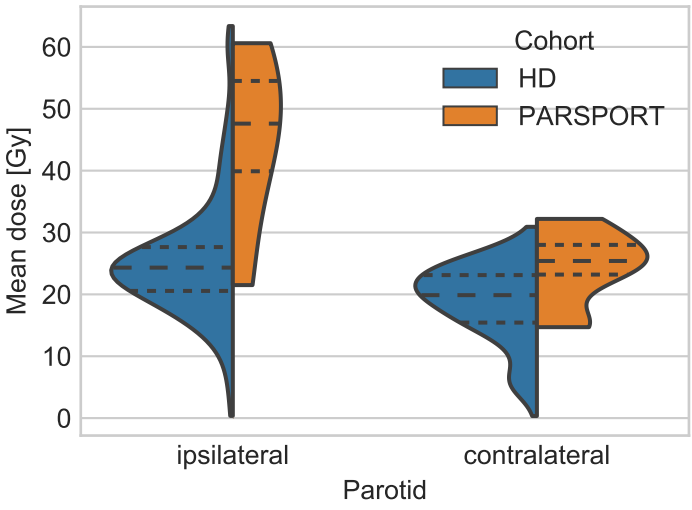



**Figure 3.3:** Comparison of parotid glands mean dose distributions between the HD and the PARSPORT data sets. The distributions are visualized as kernel density plots. Wide dashed lines mark the medians, whereas narrow dashed lines correspond to the first and the third quartiles.

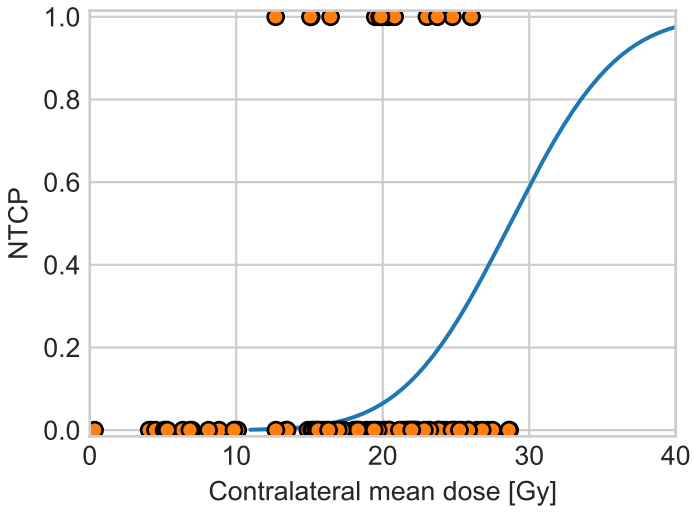



**Figure 3.4:** NTCP curve for G2+ xerostomia at 12 months based on the data from the PARSPORT clinical trial (Miah et al. 2013) with observations from the HD cohort.

**Table 3.4:** The parameters and performance of the LKB models with 95% confidence intervals.

| Endpoint | Gland | $d_{50}$ [Gy] | $m$ | AUC |
|---|---|---|---|---|
| G1+ 6 months | Contralateral | 10.7 (7.5–13.9) | 0.61 (0.46–0.75) | 0.73 (0.62–0.83) |
| G1+ 6 months | Both | 12.2 (6.4–18.0) | 0.61 (0.37–0.84) | 0.76 (0.66–0.87) |
| G1+ 12 months | Contralateral | 12.0 (7.8–16.1) | 0.61 (0.32–0.89) | 0.69 (0.54–0.83) |
| G1+ 12 months | Both | 13.6 (5.9–21.3) | 0.61 (0.32–0.89) | 0.73 (0.60–0.86) |
| G2+ 6 months | Contralateral | 42.8 (5.7–80.0) | 0.60 (0.39–0.82) | 0.49 (0.38–0.60) |
| G2+ 6 months | Both | 51.2 (8.0–94.4) | 0.60 (0.40–0.81) | 0.50 (0.39–0.62) |
| G2+ 12 months | Contralateral | 51.6 (-41.8–145.1) | 0.60 (0.20–1.00) | 0.56 (0.41–0.70) |
| G2+ 12 months | Both | 64.8 (-31.8–161.3) | 0.60 (0.29–0.91) | 0.52 (0.38–0.67) |

G1+ xerostomia models based on the mean dose to both parotid glands, that is, 12.2 Gy at 6 months and 13.6 Gy at 12 months. For all G1+ xerostomia models, $m$ parameter was 0.61. The NTCP curves for G1+ xerostomia are presented in Figure 3.5. The G2+ xerostomia models failed to fit the data which was revealed by extremely wide confidence intervals of $d_{50}$ (Table 3.4).

Predictive performance of the LKB models ranged from AUC = 0.69 to 0.76 for G1+ xerostomia. The G1+ models based on the mean dose to both parotid glands scored higher than the contralateral mean dose models for both analyzed time points. Nonetheless, the difference was not statistically significant at the significance level of 0.05. The PR curves also did not reveal any substantial advantage of one predictor over the other. Figure 3.5 shows the ROC and the PR curves for the LKB G1+ models. The predictive performance of all G2+ LKB models was close to a random classifier, that is, AUC = 0.50. For that reason, the ROC and the PR curves were not presented on the plots. The AUCs of all models together with confidence intervals are summarized in Table 3.4.

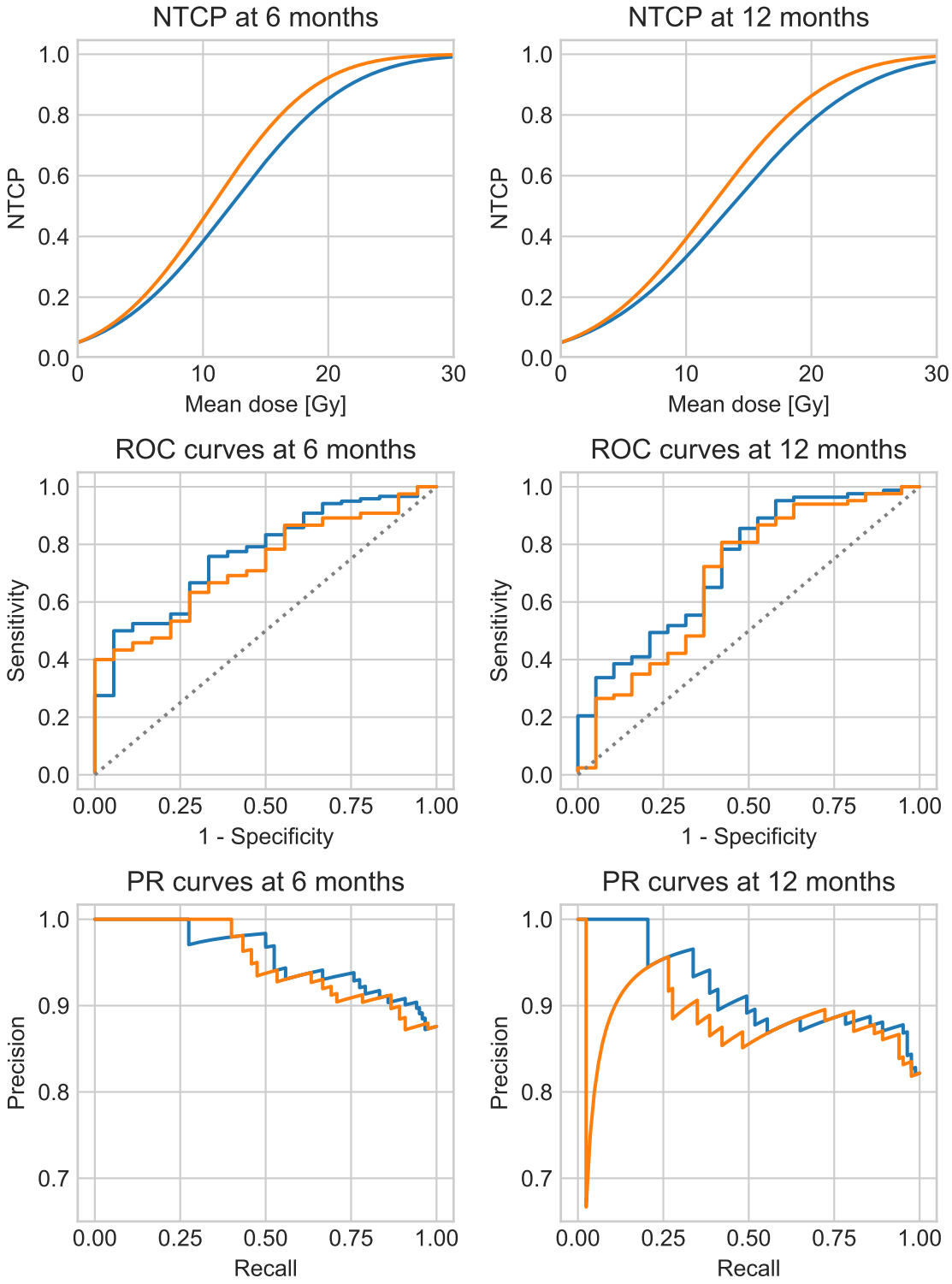



**Figure 3.5:** NTCP, ROC, and PR curves for G1+ xerostomia. The curves are presented for models based on both parotid glands (—) or the contralateral parotid gland alone (—).

## 3.3 Discussion

In recent years, considerable attention has been given to the validation of the QUANTEC guidelines to avoid radiation-induced xerostomia (Lee & Fang 2013, Beetz et al. 2014). Most of the studies supported the QUANTEC recommendations. Nonetheless, the common conclusion was that further dose reduction was advisable.

Median mean doses of 46 and 25 Gy to ipsi- and contralateral parotid glands, respectively, corresponded to 38% LENT-SOMA subjective G2+ xerostomia prevalence at 12 months in the PARSPORT clinical trial. Beetz et al. (2012) reported similar moderate-to-severe xerostomia prevalence measured with patient-based EORTC QLQ-H&N35 questionnaires. The average mean dose to parotid glands was 30 Gy. Leung & Lee (2013) showed that median mean doses of 22 and 20 Gy to ipsi- and contralateral parotid glands, respectively, resulted in RTOG/EORTC G2+ xerostomia in only two of 72 patients in their study. Moreover, Lee et al. (2015) presented a patient cohort with median ipsi- and contralateral parotid gland mean doses of 23 Gy and 22 Gy, respectively, and only 24% EORTC QLQ-H&N35 moderate-to-severe xerostomia prevalence at 6 months after treatment. The results presented in this chapter seem to be in accordance with the dose-response pattern emerging from the recently published studies. Namely, sparing both parotid glands to a mean dose as low as possible resulted in low G2+ xerostomia prevalence at 6 and 12 months after treatment.

Comparison of toxicity scores based on different toxicity grading systems is often difficult, as concordance between the assigned scores can be low. Quantitative evaluations, such as saliva flow measurements, are highly variable and often translate poorly to patient's quality of life (Deasy et al. 2010). Also, purely observer-based systems, such as the RTOG/EORTC, suffer from misinterpretation and omission errors (Trotti et al. 2007) and can underestimate patient's subjective xerostomia (Deasy et al. 2010, Meirovitz et al. 2006). The common notion is that patient-reported scores should be used for xerostomia assessment either as a solely measure (Meirovitz et al. 2006, Ho et al. 2010) or as a supplementary information at least (Deasy et al. 2010, Trotti et al. 2007). Ho et al. (2010) compared two such patient-based scoring systems, namely EORTC QLQ-HN35 with the LENT-SOMA subjective, and found good agreement between them. This finding likely extends to the CTCAE system because the LENT-SOMA items were incorporated into the CTCAE v3.0.

The LKB model proved to be unsuitable for G2+ xerostomia prediction in the HD patient cohort. The widths of the confidence intervals were in range of hundreds of grays showing that the model failed to fit the data. The lack of predictive power was supported by low AUC values. Due to direct correspondence between the AUC and the Mann-Whitney $U$-test, low AUC values were a hallmark of weak class separation with respect to the predictor. The results show that neither the mean dose to both parotid glands, nor the mean dose to the contralateral parotid gland differentiated well patients at risk of G2+ xerostomia in the HD data set. Furthermore, even though all G2+ xerostomia patients from the HD cohort were in the low-risk region of the NTCP curve (Figure 3.4), for some cases the NTCP was up to

20–30%.

G1+ xerostomia was described fairly well by the LKB model. AUC values exceeding 0.70 indicate a clear dose dependence of mild xerostomia. It is important to note that the steepest part of the NTCP curve of the LKB model corresponds to the doses close to $d_{50}$. Hence, a mean dose reduction in this region will have relatively the largest effect on a complication probability reduction. Since $d_{50}$ estimates were in the range of 11–14 Gy, the mean dose to the parotid glands should be kept as low as possible to reduce incidence of G1+ xerostomia, even in the sub-20 Gy region.

It was commonly reported in the literature that for patients who received radiation doses exceeding QUANTEC guidelines, the LKB model well discriminated patients at risk of G2+ xerostomia (Lee & Fang 2013, Beetz et al. 2012, Houweling et al. 2010). Nowadays, however, due to effective parotid sparing techniques, the prevalence of G3 is rare and patients at risk of G2+ xerostomia are often not recognized by mean dose models. The results of the analysis presented in this chapter underscore a need for the development of new NTCP models of xerostomia, which could be better suited for modeling of xerostomia in highly conformal radiotherapy treatments. Several studies emphasized the importance of the submandibular gland function as a potential factor in xerostomia prediction (Deasy et al. 2010, Lee et al. 2015). For this reason, information about surgical resection or the radiation dose to this gland may be an valuable predictor in future studies on NTCP models of xerostomia. Additionally, it could be beneficial to examine classification algorithms exploiting more complex dosimetric (dosiomics) and organ-shape (radiomics) features (Buettner et al. 2012, Dean et al. 2016). Such models may provide clinical guidelines for treatment planning in cases where a reduction of the mean dose to parotid glands is not attainable.

CHAPTER 4

# RADIOMICS, DOSIOMICS, AND DEMOGRAPHICS TO PREDICT XEROSTOMIA

*The results presented in this chapter were published in Gabryś et al. (2018). Under the journal's copyright policy, the authors retained the copyright to their work. Figures 4.1 and 4.3, and parts of section 4.1 have been quoted verbatim from Gabryś et al. (2018).*

In the course of routine medical procedures, large amounts of data are generated. Medical documentation stores information on patients' age, sex, martial status, habits, and other demographic characteristics. Also, various kinds of medical imaging, such as computed tomography, magnetic resonance imaging, and positron emission tomography are often involved. These images are usually inspected visually by an expert to aid diagnostic or therapeutic decisions. However, quantitative information that they contain can also be extracted and further analyzed. Such analysis is the realm of radiomics (Lambin et al. 2012). In radiomics, complex descriptors, such as Haralick textures, wavelets, and Laplacian transforms, are used to describe semantic image features, such as shape, vascularity, and spiculation (Gillies et al. 2015). Radiomics has been successfully used in many applications in oncology, for example, decoding tumor phenotype (Aerts et al. 2014), prediction of metastases (Vallières et al. 2015), and cancer detection (Cameron et al. 2015). Likewise, similar techniques can be applied to describe the spatial dose distribution in radiotherapy treatment plans. For instance, tubular organs can be described using dose-surface maps (Sanchez-Nieto et al. 2001) that quantify the dose deposited to the walls of the organ. For *volumetric* organs, such as prostate and parotid gland, three-dimensional moments have been successfully used for the spatial-dose description (Buettner et al. 2012).

In recent years, a number of studies have examined a variety of factors in hope of more precise xerostomia predictions. Lee et al. (2014) investigated predictive power of patient-specific predictors related to demographics, such as age, financial status, education, smoking, and alcohol intake. Hawkins et al. (2018) studied how the mean dose not only to parotid glands but also to submandibular glands and the oral cavity affects treatment out-

comes. Luijk et al. (2015) showed that it is sufficient to spare a subvolume of the parotid gland which contains stem cells to avoid parotid gland disfunction one year after treatment. Other researches focused on relating radiomic features or spatial dose shape within parotid glands to xerostomia. Van Dijk et al. (2017) examined CT image biomarkers to predict late xerostomia and sticky saliva, whereas Buettner et al. (2012) showed that the use of three-dimensional scale-invariant dose moments can predict late xerostomia better than the mean dose to parotid glands.

In this chapter, the relation of various features describing parotid shape (radiomics), dose shape (dosiomics), and demographic characteristics to radiation-induced xerostomia was investigated. The conducted analysis aimed to recognize informative predictors beyond the mean dose that would allow better prediction of early, late, and long-term xerostomia.

## 4.1 Material and Methods

### 4.1.1 Endpoints

The follow-up reports after treatment were collected at approximately three-months intervals. As already discussed in previous chapters, the number of available reports and the exact date of evaluation varied from patient to patient. For this reason, three intervals representing early, late, and long-term xerostomia were defined:

- 0–6 months (early xerostomia)
- 6–15 months (late xerostomia)
- 15–24 months (long-term xerostomia)

The intervals are visually marked in Figure 4.1. Like in chapter 3, if a given patient received more than one xerostomia evaluation in a certain time interval, the final toxicity score was calculated as the arithmetic mean rounded to the nearest integer with x.5 being rounded up. In this chapter, the focus was on xerostomia grade two or higher (G2+) according to the CTCAE v4.03. Patient and tumor characteristics stratified by endpoint are presented in Table 4.1.

### 4.1.2 Feature definitions

The candidate xerostomia predictors are listed in Table 4.2. The features were subdivided in nine feature groups: demographics, parotid shape, dose-volume histogram, subvolume mean dose, spatial dose gradient, spatial dose spread, spatial dose correlation, spatial dose skewness, and spatial dose coskewness. As explained in chapter 2, the radiomic and the dosiomic features were extracted from the CT- and the dose-cubes read from treatment planning DICOM files. The features were analyzed in terms of ipsi- and contralateral rather than left and right parotid glands. For this reason, the cubes were flipped through the sagittal plane for cases with the mean dose to the right parotid gland higher than the mean dose to

**Table 4.1:** Patients and tumor characteristics. The total number of patients differ among the groups due to the follow-up availability.

| | | 0–6 months | | | 6–15 months | | | 15–24 months | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | All patients | Grade 0 | Grade 1 | Grade 2 | Grade 0 | Grade 1 | Grade 2 | Grade 0 | Grade 1 | Grade 2 |
| **Total patients** | 153 | 17 | 87 | 30 | 19 | 99 | 13 | 15 | 53 | 9 |
| **Age** | | | | | | | | | | |
| Median | 61 | 60 | 60 | 62 | 60 | 61 | 61 | 61 | 61 | 61 |
| Q1-Q3 | 55–66 | 54–66 | 54–64 | 53–69 | 57–63 | 53–66 | 54–68 | 55–68 | 52–66 | 54–68 |
| Range | 29–82 | 44–78 | 29–82 | 43–80 | 49–75 | 29–82 | 43–74 | 47–80 | 39–78 | 41–80 |
| **Sex** | | | | | | | | | | |
| Female | 37 | 5 | 19 | 7 | 6 | 24 | 2 | 2 | 9 | 4 |
| Male | 116 | 12 | 68 | 23 | 13 | 75 | 11 | 13 | 44 | 5 |
| **Tumor site** | | | | | | | | | | |
| Hypopharynx/Larynx | 37 | 7 | 20 | 7 | 7 | 20 | 2 | 3 | 15 | 0 |
| Nasopharynx | 12 | 0 | 8 | 2 | 2 | 8 | 1 | 0 | 5 | 0 |
| Oropharynx | 99 | 9 | 57 | 20 | 10 | 69 | 9 | 11 | 32 | 9 |
| Other | 5 | 1 | 2 | 1 | 0 | 2 | 1 | 1 | 1 | 0 |
| **Radiation modality** | | | | | | | | | | |
| IMRT | 37 | 2 | 25 | 5 | 1 | 29 | 2 | 2 | 18 | 1 |
| Tomotherapy | 116 | 15 | 62 | 25 | 18 | 70 | 11 | 13 | 35 | 8 |
| **Ipsilateral parotid mean dose [Gy]** | | | | | | | | | | |
| Median | 24.3 | 22.9 | 25.0 | 23.0 | 19.5 | 24.8 | 25.9 | 22.9 | 23.8 | 24.5 |
| Q1-Q3 | 20.6–27.6 | 18.5–24.6 | 21.4–29.0 | 21.4–25.4 | 16.8–24.3 | 21.8–28.7 | 21.8–27.2 | 18.5–31.5 | 20.8–26.4 | 21.6–26.2 |
| Range | 0.4–63.4 | 0.4–36.0 | 7.4–61.4 | 4.6–59.0 | 0.4–32.9 | 4.6–61.4 | 17.3–63.4 | 0.4–51.4 | 4.6–46.0 | 17.3–63.4 |
| **Contralateral parotid mean dose [Gy]** | | | | | | | | | | |
| Median | 19.9 | 19.4 | 20.3 | 19.6 | 15.6 | 20.5 | 20.4 | 12.7 | 19.7 | 20.1 |
| Q1-Q3 | 15.4–23.1 | 13.1–21.8 | 15.2–23.8 | 16.5–22.0 | 10.3–20.7 | 16.3–23.8 | 19.8–23.1 | 5.2–17.9 | 16.3–23.7 | 16.4–22.3 |
| Range | 0.3–30.9 | 0.3–24.9 | 4.1–28.6 | 4.2–26.2 | 0.3–27.9 | 4.1–30.9 | 15.1–26.2 | 0.3–27.9 | 4.1–27.2 | 15.1–26.0 |

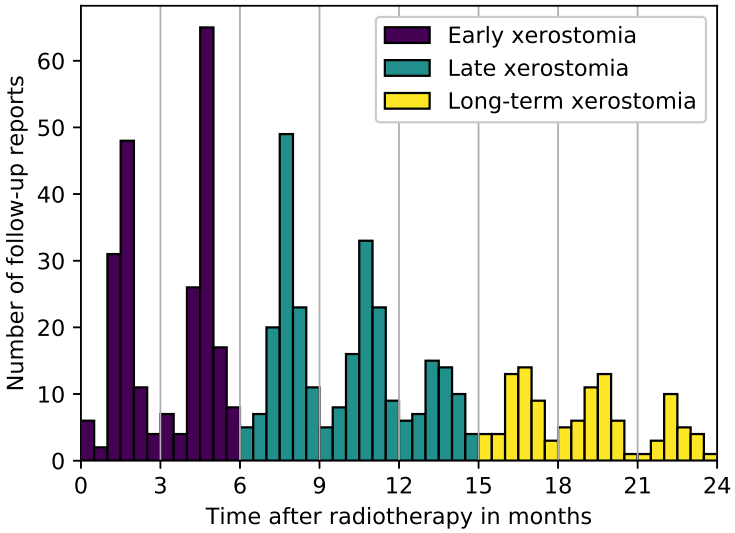



**Figure 4.1:** Frequency of the follow-up reports collection.

**Table 4.2:** Feature sets before and after removal of highly correlated pairs. Removal of correlated features is provided in section 4.1.3.

| Feature group | Initial feature set | Final feature set |
|---|---|---|
| Demographics | age, sex | age, sex |
| Parotid shape | volume, area, sphericity, eccentricity, compactness, $\lambda_1$, $\lambda_2$, $\lambda_3$ | volume, sphericity, eccentricity |
| Dose-volume histogram | mean, spread, skewness, D2, D98, D10, D20, D30, D40, D50, D60, D70, D80, D90, V10, V15, V20, V25, V30, V35, V40, V45, entropy, uniformity | mean, spread, skewness |
| Subvolume mean dose | $s_x^1$, $s_x^2$, $s_x^3$, $s_y^1$, $s_y^2$, $s_y^3$, $s_z^1$, $s_z^2$, $s_z^3$ | |
| Spatial dose gradient | $\text{gradient}_x$, $\text{gradient}_y$, $\text{gradient}_z$ | $\text{gradient}_x$, $\text{gradient}_y$, $\text{gradient}_z$ |
| Spatial dose spread | $\eta_{200}$, $\eta_{020}$, $\eta_{002}$ | $\eta_{200}$, $\eta_{020}$, $\eta_{002}$ |
| Spatial dose correlation | $\eta_{110}$, $\eta_{101}$, $\eta_{011}$ | $\eta_{110}$, $\eta_{101}$, $\eta_{011}$ |
| Spatial dose skewness | $\eta_{300}$, $\eta_{030}$, $\eta_{003}$ | $\eta_{300}$, $\eta_{030}$, $\eta_{003}$ |
| Spatial dose coskewness | $\eta_{012}$, $\eta_{021}$, $\eta_{120}$, $\eta_{102}$, $\eta_{210}$, $\eta_{201}$ | $\eta_{012}$, $\eta_{021}$, $\eta_{120}$, $\eta_{102}$, $\eta_{210}$, $\eta_{201}$ |

the left parotid gland. This allowed to avoid situations where certain spatial features, such as dose gradient, would have either positive or negative value, depending on whether the tumor was located on the patient's left or right side. All feature definitions were based on the LPS coordinate system, that is (right to left, anterior to posterior, inferior to superior). MATLAB code written for data preprocessing and feature extraction was made publicly available on GitHub[1].

### Demographic features

The demographic features group consisted of only two predictors: the patient's age and sex. This information was readily available in the medical documentation. Inclusion of other

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

factors, such as smoking, financial status and education, was unfeasible due to retrospective character of the study and, resulting from this, variability in the available information.

#### Parotid shape features

Parotid shape features described size and shape of the parotid glands.

**Volume** Volume $V$ of the parotid gland was given by

$$V = \sum_{x,y,z} I(x,y,z)\,, \tag{4.1}$$

where $I(x,y,z)$ is the three-dimensional logical array indicating whether a voxel of coordinates $x$, $y$, and $z$ is within the parotid gland.

**Surface area** Surface area $A$ of the parotid gland was calculated as

$$\begin{aligned} A = &\sum_{yz}[I(1,y,z) + \sum_{x=2}^{n_x-1} |I(x,y,z) - I(x-1,y,z)| + I(n_x,y,z)] \\ &+ \sum_{xz}[I(x,1,z) + \sum_{y=2}^{n_y-1} |I(x,y,z) - I(x,y-1,z)| + I(x,n_y,z)] \\ &+ \sum_{xy}[I(x,y,1) + \sum_{z=2}^{n_z-1} |I(x,y,z) - I(x,y,z-1)| + I(x,y,n_z)]\,, \end{aligned} \tag{4.2}$$

where $n_x$, $n_y$, and $n_z$ are the number of voxels in x, y, and z direction respectively.

**Sphericity** Parotid sphericity $\Psi$ was defined as the ratio of the surface area of a sphere of the same volume as the parotid to the actual surface area of the parotid

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}\,. \tag{4.3}$$

**Compactness** Parotid compactness $\kappa$ was given by the ratio of the parotid surface area $A$ to the parotid volume $V$

$$\kappa = \frac{A}{V}\,. \tag{4.4}$$

**Eccentricity** Eccentricity $\varepsilon$ measured elongation of the parotid gland given by

$$\varepsilon = 1 - \sqrt{\frac{\lambda_{min}}{\lambda_{max}}}\,, \tag{4.5}$$

where the eigenvalues $\lambda_i$ of the parotid shape covariance matrix correspond to the dimensions of the parotid gland along the principal axes defined by the eigenvectors. Larger

asymmetry of the gland corresponded to larger values of $\varepsilon$. The covariance matrix needed to calculate the eigenvectors is defined as

$$\mathrm{cov}[I(x,y,z)] = \begin{pmatrix} \mu_{200} & \mu_{110} & \mu_{101} \\ \mu_{110} & \mu_{020} & \mu_{011} \\ \mu_{101} & \mu_{011} & \mu_{002} \end{pmatrix}, \tag{4.6}$$

where $\mu_{pqr}$ denotes central moments of the parotid given by

$$\mu_{pqr} = \sum_{x,y,z} (x-\bar{x})^p (y-\bar{y})^q (z-\bar{z})^r I(x,y,z) \tag{4.7}$$

and

$$\bar{x} = \frac{\sum_{x,y,z} xI(x,y,z)}{\sum_{x,y,z} I(x,y,z)}, \tag{4.8a}$$

$$\bar{y} = \frac{\sum_{x,y,z} yI(x,y,z)}{\sum_{x,y,z} I(x,y,z)}, \tag{4.8b}$$

$$\bar{z} = \frac{\sum_{x,y,z} zI(x,y,z)}{\sum_{x,y,z} I(x,y,z)}. \tag{4.8c}$$

#### Dose-volume histogram features

This feature group quantified characteristics of a differential dose-volume histogram (DVH).

**Mean** The mean dose $\bar{d}$ to the parotid gland.

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} d_i, \tag{4.9}$$

where $d_i$ denotes the dose to $i$th voxel and $N$ denotes total number of voxels in the parotid gland.

**Spread** The spread $\sigma$ of the DVH was quantified with the standard deviation of the dose within the parotid gland, that is

$$\sigma = \sqrt{\frac{1}{N-1} \sum |d_i - \bar{d}|^2}. \tag{4.10}$$

Low value of $\sigma$ indicate a narrow DVH and, therefore, rather homogeneous dose distribution.

**Skewness** Skewness $\gamma_1$ of the DVH was defined as

$$\gamma_1 = \frac{\frac{1}{N}\sum(d_i - \bar{d})^3}{(\sqrt{\frac{1}{N}\sum(d_i - \bar{d})^2})^3}\,, \tag{4.11}$$

which is the third standardized moment. Negative skewness corresponds to a DVH skewed toward a low-dose region, whereas positive skewness describes a DVH skewed toward a high-dose region.

**Dx** Minimum dose to $x\%$ *hottest* volume of the parotid gland was given by

$$Dx = \inf\{d \in \mathbb{R} : x \leq \sum_{d_i > d} 1\}\,, \tag{4.12}$$

where $v_i$ is a partial volume of the $i$th voxel.

**Vx** Percentage volume of the parotid gland receiving at least $x$ Gy was defined as

$$Vx = \frac{\sum_{d_i \geq x} v_i}{\sum v_i}\,. \tag{4.13}$$

**Entropy** Entropy $H$ (Gonzalez & Woods 2006), given by

$$H = -\sum_{i=1}^{256} m(d_i) \log m(d_i)\,, \tag{4.14}$$

measures smoothness of the dose within the parotid gland. The $d_i$ is the dose delivered to the $i$th voxel and $m(d_i)$ is the corresponding histogram. Entropy is a nonnegative quantity reaching zero only for a perfectly uniform distribution.

**Uniformity** Uniformity $U$ (Gonzalez & Woods 2006) of the dose was defined as

$$U = \sum_{i=1}^{256} m^2(d_i)\,. \tag{4.15}$$

Uniformity $U$, sometimes also called energy, assumes values form zero to one. For a uniform dose $U = 1$.

#### Subvolume mean dose

Parotid gland subvolumes were defined by axial, coronal, and sagittal splits that cut parotid glands in thirds along the patient's axes. The cuts were positioned in such a way that each subvolume comprised approximately the same number of voxels. As a result, nine, not exclusive, subvolumes were defined: three in x, three in y, and three in z direction. For each subvolume the mean radiation dose was calculated, for example, the mean dose to the

anterior third of the parotid gland ($s_y^1$) or the mean dose to the superior third of the parotid gland ($s_z^3$).

#### Dose gradients

Dose gradients measured average change of the dose along one of patient axes and were defined as

$$\text{gradient}_\text{x} = \frac{\sum_{x,y,z}(D(x+1,y,z) - D(x-1,y,z))I(x,y,z)}{2\sum_{x,y,z} I(x,y,z)}, \tag{4.16a}$$

$$\text{gradient}_\text{y} = \frac{\sum_{x,y,z}(D(x,y+1,z) - D(x,y-1,z))I(x,y,z)}{2\sum_{x,y,z} I(x,y,z)}, \tag{4.16b}$$

$$\text{gradient}_\text{z} = \frac{\sum_{x,y,z}(D(x,y,z+1) - D(x,y,z-1))I(x,y,z)}{2\sum_{x,y,z} I(x,y,z)}. \tag{4.16c}$$

where $x$, $y$, and $z$ are the coordinates of the voxel, $D(x,y,z)$ is the dose delivered to the voxel, and $I(x,y,z)$ is a logical array indicating whether a voxel belongs to the parotid.

#### Three-dimensional dose moments

The scale-invariant dose moments allowed to quantify the three-dimensional shape of the dose distribution within the parotid gland. Visualization of the moments can be found in Buettner et al. Supplementary figure 1–3 (Buettner et al. 2012). The moments were defined as

$$\eta_{pqr} = \frac{\sum_{x,y,z}(x-\bar{x})^p(y-\bar{y})^q(z-\bar{z})^r D(x,y,z)I(x,y,z)}{\left(\sum_{x,y,z} D(x,y,z)I(x,y,z)\right)^{\frac{p+q+r}{3}+1}}, \tag{4.17}$$

where

$$\bar{x} = \frac{\sum_{x,y,z} xI(x,y,z)D(x,y,z)}{\sum_{x,y,z} I(x,y,z)D(x,y,z)}, \tag{4.18a}$$

$$\bar{y} = \frac{\sum_{x,y,z} yI(x,y,z)D(x,y,z)}{\sum_{x,y,z} I(x,y,z)D(x,y,z)}, \tag{4.18b}$$

$$\bar{z} = \frac{\sum_{x,y,z} zI(x,y,z)D(x,y,z)}{\sum_{x,y,z} I(x,y,z)D(x,y,z)}. \tag{4.18c}$$

In particular, moments quantifying dose variance, covariance, skewness, and coskewness were considered.

**Dose variance ($\eta_{200}$, $\eta_{020}$, $\eta_{002}$)** Dose variance corresponds to spread of the dose along a given direction.

**Dose covariance($\eta_{110}$, $\eta_{101}$, $\eta_{011}$)** Dose covariance measures how the dose covaries along two axes. For example, large positive values of $\eta_{110}$ correspond to dose deposition along $xy$

direction, whereas large negative values correspond to dose deposition along the direction perpendicular to $xy$.

**Dose skewness ($\eta_{300}$, $\eta_{030}$, $\eta_{003}$)** Dose skewness measures asymmetry of the dose distribution along a given axis.

**Dose coskewness ($\eta_{210}$, $\eta_{201}$, $\eta_{120}$, $\eta_{021}$, $\eta_{012}$, $\eta_{102}$)** Dose coskewness measures how dose variance along one direction covaries with another dimension. For example, large $\eta_{210}$ would mean that variance of the dose along $x$ axis increases when moving up the $y$ axis.

### 4.1.3 Feature correlation

To reduce feature redundancy, the Kendall rank correlation coefficient was calculated for all feature pairs. Kendall's $\tau$, defined by

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}, \tag{4.19}$$

allows to measure ordinal association between two features, in other words, agreement in ranks assigned to the observations. It can be interpreted as a difference between the probability that both features would rank a random pair of observations in the same way and the probability that they would rank these observations in a different way (Abdi 2007). Feature pairs with $|\tau| > 0.5$ in both glands were considered as highly correlated and suitable for rejection from the initial feature set. This arbitrarily chosen threshold corresponds to a 75% probability that the two features would rank a random pair of observations in the same way. Whenever a pair of features was found highly correlated, the feature that was conceptually and computationally simpler was kept in the feature set, for example, mean dose over Dx, parotid volume over parotid compactness.

### 4.1.4 Predictive power of the features

In order to estimate predictive power of individual features, the Mann-Whitney $U$ statistic and the corresponding AUC were calculated for every feature. The null hypothesis for each test was that the median value of the feature was the same in the group of patients with G0 and G1 xerostomia, and the group of patients with G2+ xerostomia. The alternative hypothesis was that the median value was different between the two groups. For all AUCs, 95% confidence intervals were estimated with bias-corrected and accelerated (BCa) bootstrap. The BCa method was discussed in detail in section 3.1.4. Type I error was a recognition of a feature as informative when in fact it was not.

As already discussed in chapter 3, performing multiple tests increases the risk of reporting falsely positive results. In the context of this analysis, evaluating multiple features increases a risk of observing overoptimistic AUC values just by chance. For this reason, a correction for multiple testing was applied. The number of type I errors was controlled with

the false discovery rate (FDR) using the Gavrilov-Benjamini-Sarkar method. The choice of this particular method is explained in the next section.

#### False discovery rate

The FDR is defined as the expected proportion of falsely rejected hypotheses $V$ in the set of all the rejected hypotheses $R$, that is

$$\mathrm{FDR} = \mathrm{E}\left[\frac{V}{R}\right] . \tag{4.20}$$

Instead of controlling the probability of at least one false positive result in the family of all tests as the FWER does (*vide* section 3.1.2), it rather controls the proportion of type I errors within the set of rejected hypotheses. The FDR provides control of the FWER in the weak sense, that is, if all null hypotheses are true, the FDR controls the FWER. Whenever there are non-true null hypotheses, the FDR provides greater power than the FWER method at the cost of a greater risk of type I errors. The FDR may be especially preferable for settings were hundreds or even thousands comparisons are performed. In such situations, the FWER may be too conservative and its power too low.

Probably the most well known algorithm of controlling the FDR is the Benjamini-Hochberg method (Benjamini & Hochberg 1995). Say $m$ hypotheses $H_1, H_2, \ldots, H_m$ are tested and of these $m_0$ are true. Also, with each hypothesis there is an associated p-value $p_1, \ldots, p_m$. First, the p-values are sorted in increasing order. Then, the maximum index $k$ is found such that

$$k = \max\left\{1 \leq i \leq m : p_i \leq \frac{i}{m}\alpha\right\} \tag{4.21}$$

and hypotheses $H_1, \ldots, H_k$ are rejected. If no such $k$ exists, no hypothesis is rejected.

The Benjamini-Hochberg method is conceptually and computationally simple and appealing. Nonetheless, it can be too conservative when the proportion $m_0/m$ is small. It is because the true proportion $m_0/m$ is unknown. The consequences of this are well discussed in Benjamini et al. (2006). To address this problem, Gavrilov et al. (2009) presented an adaptive step-down procedure that provides more powerful control of the FDR, namely

$$k = \max\left\{1 \leq i \leq m : p_i \leq \frac{i}{m - i(1-\alpha) + 1}\alpha\right\} . \tag{4.22}$$

Hypotheses $H_1, \ldots, H_k$ are rejected. If no such $k$ exists, no hypothesis is rejected. The Gavrilov-Benjamini-Sarkar method has been shown to work well under both independence and positive dependence.

### 4.1.5 Tolerance values

In order to relate raw values of the features to the expected complication probability, a univariate logistic regression model was build based on each of them. Consequently, for

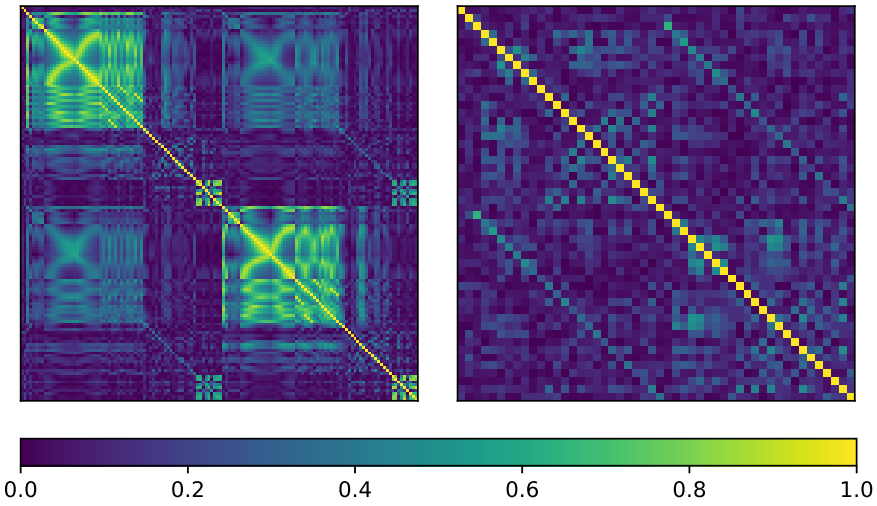



**Figure 4.2:** Correlation matrix of $|\tau|$ before (left) and after (right) feature set reduction.

each of the predictors, tolerance values corresponding to 20% (TV20), 10% (TV10), and 5% (TV5) complication probability were calculated.

## 4.2 Results

In total, 61 predictors were defined. The analysis of feature correlations and mutual redundancy started with comparison of the mean dose to parotid glands to other predictors. The mean dose was chosen because it is a gold standard in the clinical practice. The analysis revealed that all Dx, Vx, and subvolume features were highly correlated to the mean dose. Consequently, they were removed from the feature set. A number of features, such as the DVH skewness, entropy, and gradients, described inhomogeneity of the dose distribution within the parotid glands. For this reason, feature correlation to the DVH skewness was examined. The result was that entropy and uniformity were redundant with respect to the DVH skewness and could be excluded. Furthermore, there was a number of features describing shape of parotid glands. Parotid volume was considered the simplest and the most straightforward descriptor, thus other features were compared to it. As a result, surface area, compactness, and all three eigenvalues were redundant with respect to the parotid volume. Afterwards, there were 26 features left in the feature set and no feature pairs were correlated with $|\tau| > 0.5$ in both glands. The correlation matrix of features before and after removal of the correlated pairs is presented in Figure 4.2. The initial and final set of features is presented in Table 4.2.

The results of the univariate analysis are presented in Figure 4.3. There was little association between single predictors and xerostomia within the first six months after treatment. The only features that passed AUC = 0.60 threshold were the DVH skewness in both glands, contralateral $\eta_{210}$, and contralateral $\eta_{011}$. They were, however, not significant at the FDR = 0.05. Late xerostomia correlated with individual features slightly better. The most informative were the contralateral spatial dose gradients in the right-left direction and the

**Table 4.3:** AUCs and tolerance values for selected predictors of long-term xerostomia.

| Predictor | AUC | TV20 | TV10 | TV5 |
|---|---|---|---|---|
| volume$^{i}$ | 0.87 (0.75–0.95) | 9894 mm$^3$ | 15681 mm$^3$ | 21014 mm$^3$ |
| volume$^{c}$ | 0.85 (0.66–0.98) | 9169 mm$^3$ | 14533 mm$^3$ | 19475 mm$^3$ |
| gradient$^{i}_{x}$ | 0.78 (0.58–0.92) | 1.78 Gy/mm | TV10 = 1.42 Gy/mm | TV5 = 1.09 Gy/mm |
| gradient$^{c}_{x}$ | 0.84 (0.71–0.93) | 1.49 Gy/mm | TV10 = 1.29 Gy/mm | TV5 = 1.10 Gy/mm |
| gradient$^{i}_{y}$ | 0.74 (0.54–0.89) | 0.67 Gy/mm | TV10 = 0.42 Gy/mm | TV5 = 0.20 Gy/mm |
| gradient$^{c}_{y}$ | 0.72 (0.55–0.87) | 0.69 Gy/mm | TV10 = 0.36 Gy/mm | TV5 = 0.07 Gy/mm |

anterior-posterior direction. Considerable performance was achieved also by the contralateral DVH spread and ipsilateral $\eta_{210}$. Nevertheless, no late xerostomia AUC was statistically significant at the FDR = 0.05. Long-term xerostomia predictors achieved by far the highest AUCs. Parotid size described by parotid volume predicted long-term xerostomia with AUCs greater than 0.85. Furthermore, dose-gradients in the right-left and anterior-posterior predicted long-term xerostomia very well (AUCs > 0.7). Other contralateral features that that had relatively high predictive power were the DVH spread, parotid eccentricity, and coskewness in the $xy$ plane ($\eta_{210}$ and $\eta_{120}$). Also, patient's sex allowed to recognize patients at risk with AUC = 0.64. Nonetheless, at FDR = 0.05 statistically significant were only parotid gland volumes and the right-left spatial dose gradient in the contralateral gland. The corresponding AUCs and tolerance values of selected long-term xerostomia predictors are presented in Table 4.3.

## 4.3 Discussion

The investigation of correlations between the features and their mutual redundancy proved to be an important part of the analysis. Removal of highly correlated descriptors allowed to significantly reduce feature redundancy and lower the dimensionality of the data set from 61 to 26 features. Furthermore, the correlation analysis allowed to exclude complex features that have low added value over simpler and more intuitive ones. For example, the DVH skewness rendered entropy and uniformity redundant.

The analysis revealed large variability of the features predictive power depending on the analyzed time point. The most marked is probably the parotid volume. It was completely nonpredictive of early and late xerostomia. For the long-term xerostomia, however, it scored AUC greater than 0.85 for both ipsi- and contralateral parotids. Similar variability was observed for other descriptors as well, for example, spatial dose gradients. This finding underlines necessity of evaluation of radiomic and dosiomic features at various times after treatment as their predictive power may change among early, late, and long-term effects. Nonetheless, so distinct variability as observed in this study was likely emphasized by the low sample size and the retrospective character of the collected data.

Another important aspect was the number of statistically significant AUC values. Even though six features achieved AUC greater than 0.70, only three were statistically significant at the FDR = 0.05. It is important to remember that high AUC does not have to translate to

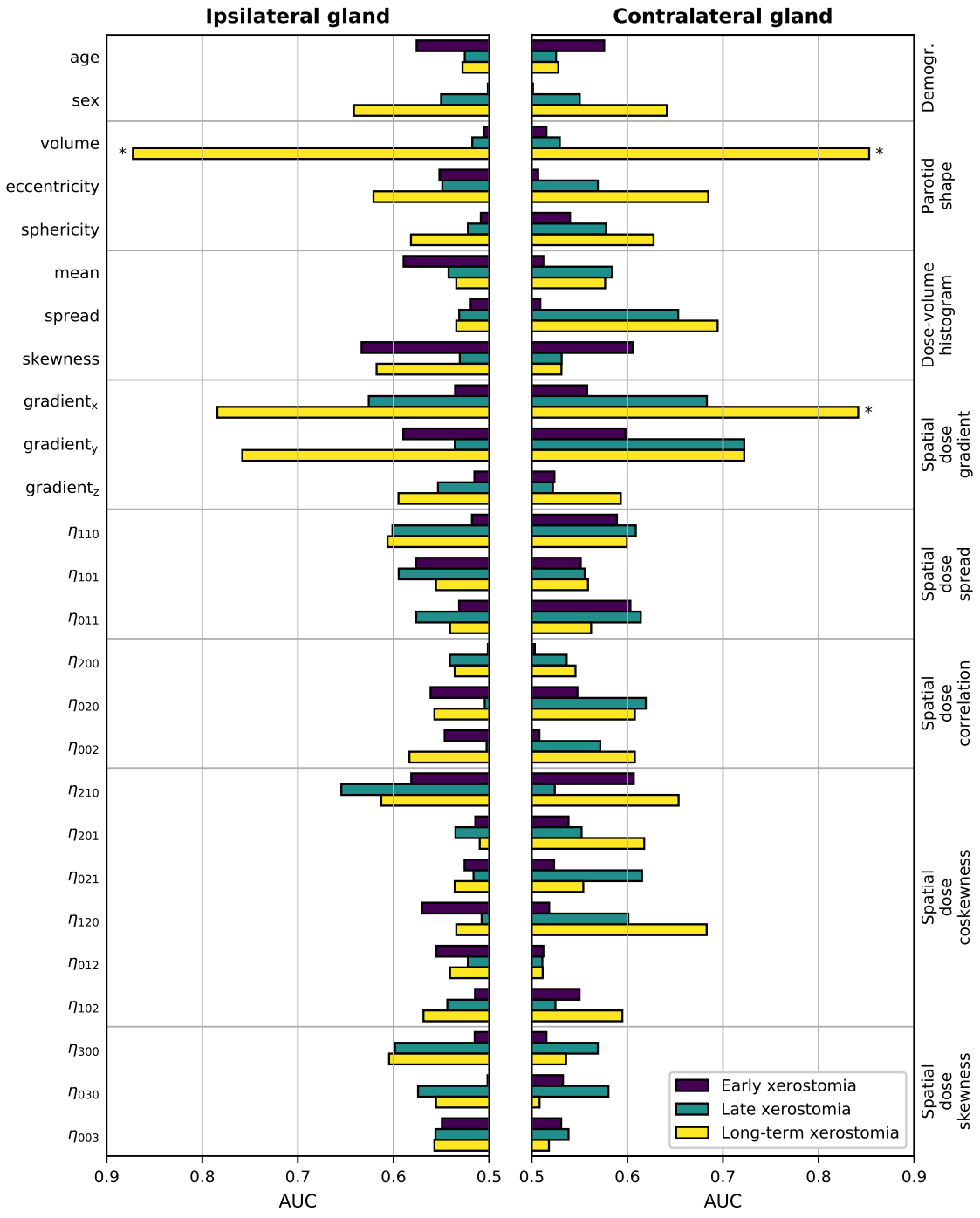



**Figure 4.3:** Predictive power of individual features in the time-specific models measured with the area under the receiver operating characteristic curve (AUC). The left-hand side vertical axis lists the features, the right-hand side vertical axis lists the feature groups. The AUCs were calculated from the corresponding Mann-Whitney $U$ statistic. Bars marked with * are significant at the false discovery rate (FDR) = 0.05.

high confidence of the observed predictive power. For instance, for large samples, AUC of 0.60 can be significant, whereas for small samples, AUC of 0.80 can be nonsignificant. In other words, AUC itself informs only about the strength of the effect, whereas p-value and confidence intervals quantify uncertainty of the AUC estimate. These results emphasize the importance of correction factors and confidence intervals estimation in multiple testing settings.

The univariate analysis showed that predictors describing shape of the dose and shape of the parotid gland can be highly predictive of xerostomia. Patients with small parotid were significantly more likely to develop long-term xerostomia than patients with large parotid glands. Average volume of parotid gland in the negative group was 14374 $mm^3$ in comparison to 9557 $mm^3$ in the positive group. Similarly, patients with steep dose gradients in the right-left direction were at higher risk than patients with low dose gradient in this direction. Median gradient in the negative group was 1.2 Gy/mm in contrast to 1.7 Gy/mm in the positive group. A possible explanation of these relations could be often reported shrinkage of parotid glands and their movement toward the medial direction during the course of radiotherapy.

Nonetheless, good predictive power of the spatial dose gradients and poor performance of the mean dose can be put into perspective of other studies validating mean-dose models. In cohorts where patients received a high radiation dose to parotid glands, the mean dose allowed achieving AUC greater than 0.80 (Houweling et al. 2010, Beetz et al. 2012). It seems that for relatively high doses, the mean dose alone is a good xerostomia predictor, irrespective of the dose gradient, whereas in the highly conformal regime of modern radiotherapy, the dose gradients are more informative and the mean dose is less predictive.

CHAPTER 5

# MACHINE LEARNING MODELS OF XEROSTOMIA

*The results presented in this chapter were published in Gabryś et al. (2018). Under the journal's copyright policy, the authors retained the copyright to their work. All the figures and parts of sections 5.1, 5.2, and 5.3 have been quoted verbatim from Gabryś et al. (2018).*

A multitude of miscellaneous predictors can be designed for NTCP modeling. Demographic, radiomic, and dosiomic features can have high predictive power and recognize well patients at risk of side effects (*vide* chapter 4). Nonetheless, models based on a single predictor, that is, univariate models, are significantly limited. This is because various features can explain different parts of variation between positive and negative groups, thus being mutually complementary. The Lyman model, for instance, takes into account not only the mean dose but also the partial volume of the organ receiving that dose. It may seem only logical to include as many predictors as possible in the model in order to capture the internal structure of the data and better predict a complication risk of future patients.

Unfortunately, numerous issues are related to transition from low- to high-dimensional feature spaces. These phenomena are often referred to as *curse of dimensionality*, a term coined by Bellman (1961). One of these issues is sparsity of data in high dimensions. For instance, say, a sample of $N$ observations is considered dense for a univariate model. In order to obtain the same sampling density in case of a d-dimensional model, one would require $N^d$ observations. As a result, the required sample size grows exponentially and quickly becomes unfeasible.

A related problem is overfitting. In many model architectures, increasing the dimensionality increases the number of free parameters that can be tuned to fit the data. Consequently, a model may become too flexible and fit the training data so well that apart from the present pattern, or signal, it learns noise and eventually fails to generalize to new data sets. In order to prevent this from happening, one could increase the size of the training set to keep the risk of overfitting on the same level (Hastie et al. 2009), which, as stated before, quickly becomes unrealistic for real-life samples.

Another disadvantage of high-dimensional models is poor interpretability. Developing

an intuition for a model based on a few features is usually not difficult but interpreting a model based on dozens of covariates may be challenging if not impossible. For some applications this may not be a problem, however, in NTCP models designed to aid clinical decisions interpretability is an important factor.

Machine learning provides means to solve these issues. Various methods exist to recognize subsets of relevant informative features and discard the redundant ones. This allows to reduce model dimensionality and facilitate *understanding* the model. Furthermore, model flexibility can be tuned which allows to prevent overfitting. Additionally, sampling and data cleaning algorithms are able to remove noisy observations and improve density of the data set by creating artificial observations based on the available training samples.

Nonetheless, despite the growing interest in data-driven methods (El Naqa et al. 2018), there have been no published studies so far systematically evaluating how different machine learning techniques can be used to address challenges specific to NTCP modeling. NTCP data sets are usually relatively small, rarely exceeding 300 patients. Moreover, they are frequently imbalanced, which often results in cohorts with a small positive group and a large negative group. Furthermore, there are multiple sources of uncertainty. In radiotherapy the planned dose is not the delivered dose. Among others a disconnect between the planned and delivered dose may be implied by erroneous patient positioning, organ motion, weight loss, and tumor shrinkage. Also, the follow-up reports are subject to bias introduced by (different) human observers if physically non-objective measures (as used in this study with CTCAE) are used.

Works comprehensively comparing various machine learning methods in terms of their suitability for application-specific tasks have been already presented in the fields of bioinformatics (Statnikov et al. 2005, Olson et al. 2017) and radiomics (Parmar et al. 2015). Such analysis is missing for NTCP modeling, although it seems especially relevant. For this reason, in this chapter, the suitability for multivariate NTCP modeling of xerostomia of seven machine learning classifiers, six feature selection methods, and nine sampling (data cleaning/class balancing) algorithms was evaluated. The models were based on demographic, radiomic, and dosiomic features introduced in chapter 4. The multivariate analysis allowed to examine interactions between the features and their relative relevance and redundancy. The obtained results were compared to models based on the mean dose to parotid glands and the morphological model based on the shape of the dose (Buettner et al. 2012).

Additionally, a novel longitudinal approach to NTCP modeling has been proposed. Usually, the side effects are modeled at specific time after treatment, for example, at 12 months after radiotherapy. This is a convenient setting for well-organized prospective studies with rigorous protocols of patient evaluation. In retrospective studies, however, the structure of follow-up reports collection is much more irregular and heterogeneous (*vide* section 2.2). For this reason, the longitudinal approach presented in this chapter treated the time after treatment as a model covariate. This allowed to use all available data from model training without the need for binning the observations around a certain time point and predict the NTCP at any time after treatment.

## 5.1 Material and Methods

### 5.1.1 Endpoints

In this chapter, two types of endpoints were defined, the time-specific endpoints and the longitudinal endpoint. The time-specific endpoints were the same as in chapter 4, that is: 0–6 months, 6–15 months, and 15–24 months, to investigate early, late, and long-term xerostomia, respectively (Figure 4.1). In case there were multiple follow-up reports available for individual patients, the final toxicity score was calculated as the arithmetic mean rounded to the nearest integer with x.5 being rounded up. In the longitudinal approach, no time-intervals were defined and no toxicity grades were averaged. Instead, each patient evaluation served as a separate observation and the time after treatment was included as a covariate in the model. The intention was to provide the model enough flexibility to describe early, late, and long-term effects. Patient and tumor characteristics stratified by endpoint are presented in Table 4.1.

### 5.1.2 Previously proposed NTCP models of xerostomia

Four different mean-dose models were tested to evaluate predictive power of the mean dose in the HD cohort: three univariate logistic regression models based on the ipsilateral mean dose, the contralateral mean dose, and the mean dose to both parotid glands, as well as one bivariate logistic regression model based on the mean dose to contralateral and to ipsilateral parotid glands.

As an alternative to the mean-dose models, Buettner et al. (2012) proposed a morphological multivariate logistic regression model. The model was based on scale-invariant dose moments, namely $\eta^{\mathrm{i}}_{111}$, $\eta^{\mathrm{c}}_{002}$, $\eta^{\mathrm{c}}_{300}$, and $\eta^{\mathrm{i}}_{110}$ $\eta^{\mathrm{c}}_{110}$. The model was retrained and tested on the HD data set.

### 5.1.3 Features

The multivariate model building process investigated in this chapter directly builds upon the results of the univariate feature analysis described in chapter 4. In particular, the feature set after removal of correlated features was used (*vide* Table 4.2 and Figure 4.3).

### 5.1.4 Model building

The multivariate analysis was a multi-step process comprising feature-group selection, feature scaling, sampling (data cleaning/class balancing), feature selection, and classification. The workflow is presented in Figure 5.1.

#### Workflow

The first step of the workflow was an initial, unsupervised dimensionality reduction of the feature space by random selection of feature groups (Table 4.2) used for model training.

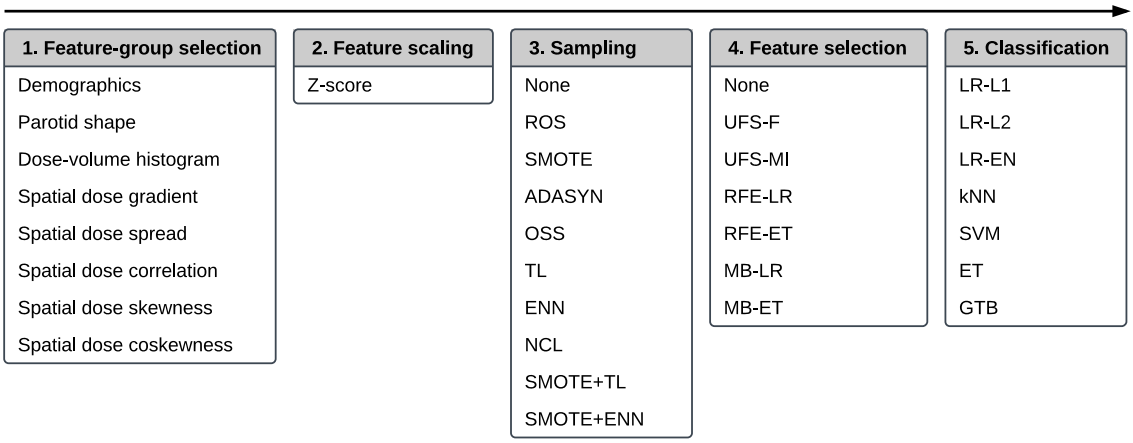



**Figure 5.1:** The workflow of a multivariate five-step model building comprising, in this order, feature-group selection, feature scaling, sampling, feature selection, and classification.

The selection was realized by performing a Bernoulli trial for every feature group with a 50% chance of success. If a given group was selected, all features in this group were accepted for further analysis. If no group was selected after performing all Bernoulli trials, the procedure was repeated for all feature groups. Reduced dimensionality of feature space often translates into better predictive performance and facilitates interpretation of a model.

In the second step, all features were scaled via Z-score normalization. Feature normalization usually improves stability and speed of optimization algorithms because scaled features are in comparable order of magnitude.

The third step was responsible for class balancing and data cleaning. Class imbalance is the difference in size between the modeled classes, in case of this study, G0-1 versus G2+ xerostomia. It was reported that small minority class together with class imbalance can hinder performance of predictive models (Japkowicz & Stephen 2002, He & Garcia 2009). This is applicable to the HD cohort as the prevalence of G2+ xerostomia ranged from 10% to 22%, depending on the endpoint. Two approaches are commonly used to reduce class imbalance: oversampling and undersampling. In oversampling, one lowers the imbalance between classes by random replication or synthetic creation of minority class observations. Conversely, in undersampling, the majority class size is reduced by discarding some of its observations. Furthermore, there are data cleaning methods which exclude some observations irrespective of their class membership. These methods remove the observations that can be considered noise or the observations close to the decision boundary. In order to investigate the influence of various sampling and data cleaning methods on the NTCP models performance, ten data sampling scenarios were examined:

- No sampling or data cleaning (None)
- Random oversampling (ROS)
- Synthetic minority oversampling (SMOTE)
- Adaptive synthetic sampling (ADASYN)

- One-sided selection (OSS)
- Tomek links (TL)
- The Wilson's edited nearest neighbor rule (ENN)
- The neighborhood cleaning rule (NCL)
- SMOTE followed by ENN (SMOTE+ENN)
- SMOTE followed by TL (SMOTE+TL)

The detailed description of the sampling algorithms is given in Appendix B.1.

The fourth step of the analysis was feature selection. One of rationales for feature selection is a reduction of model complexity, which decreases the risk of overfitting and often improves model performance (Guyon & Elisseeff 2003). Also, the reduced dimensionality of the model facilitates understanding of the relations between the predictors and the modeled outcome (here: xerostomia). In this study, seven feature selection scenarios were tested:

- No feature selection (None)
- Univariate feature selection by F-score (UFS-F)
- Univariate feature selection by mutual information (UFS-MI)
- Recursive feature elimination by logistic regression (RFE-LR)
- Recursive feature elimination by extra-trees (RFE-ET)
- Model-based feature selection by logistic regression (MB-LR)
- Model-based feature selection by extra-trees (MB-ET)

The details on the feature selection algorithms are provided in Appendix B.2.

The last step of the workflow was classification. Selection of a particular classification algorithm is an important part of model building which determines shape and flexibility of the model's decision boundary. On the one hand, a too flexible model can result in overfitting and low generalizability. On the other hand, a too simple model can fail to capture the complexity of the true decision boundary and result in underfitting. Moreover, the interpretability of a model depends strongly on the type of the chosen algorithm. In this study seven machine learning classification algorithms were compared:

- 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)

- Support vector machines (SVM)
- Extra-trees (ET)
- Gradient tree boosting (GTB)

More details on the classification algorithms are given in Appendix B.3.

Multivariate NTCP models of xerostomia were built for every combination of the classification, feature selection, and data sampling methods. This resulted in 490 models per endpoint and, consequently, 1960 models in total. A given classifier or a feature selection algorithm was a part of 210 time-specific and 70 longitudinal models. Every data sampling method was involved in 147 time-specific and 49 longitudinal models.

#### Model tuning

The process of model building can be divided into a number of steps. First, the input space and the modeled variable need to be specified. Second, the model architecture is defined, for example, the data preprocessing, the sampling algorithm, the feature selection method, and the classifier. Third, the model hyperparameter values need to be set. Hyperparameters control various characteristics of the models, for example, the regularization strength in logistic regression, the number of trees in extra-trees, and the number of neighbors in k-nearest neighbors. The process of finding optimal values of hyperparameters is called model tuning. Last, the model is trained, that is, optimal values of model parameters are estimated by minimization of some cost function.

A number of hyperparameter tuning methods exists; most common are grid search and random search. In this study the random search optimization was used because it was shown to find better hyperparameter values than grid search (Bergstra & Bengio 2012). The type and the range of the hyperparameters that were used for model tuning were based on previously reported values that worked well in various machine learning tasks (Tables B.1, B.2, and B.3).

The process of hyperparameter optimization by random search is the following: 1) select a random sample of hyperparameter values, 2) train the model using these hyperparameter values, 3) evaluate the model performance. This three-step procedure is repeated until a desired number of iterations is reached. The hyperparameter values that result in the best performance score are selected.

In order to evaluate model performance for a given set of hyperparameters (the third step of the random search algorithm), the available observations need to be divided into a training set and a test set. This allows to test the model on data that it did not *see* during training. In large data sets this approach usually works well. In small cohorts, however, the particular way the data set was divided can affect the obtained results. A typical solution to this problem is to repeat the process of splitting the data, model training, and model evaluation in order to obtain a set of performance scores that can be averaged to provide more robust performance estimate. This approach is called cross-validation. There are

many variations of cross-validation which differ in terms of the bias and variance of the provided performance estimates.

In this study two types of cross-validation were used, depending if the model was time-specific or had a longitudinal architecture. In the time-specific models, the stratified Monte Carlo cross-validation (MCCV) (Molinaro et al. 2005) with 300 splits and 10% of observations held out for testing at each split was used. For the longitudinal models, modified leave-pair-out cross-validation (LPOCV) (Krzanowski & Hand 1997, Airola et al. 2011) was implemented. The modification consisted in that all the training observations sharing patient ID with the test fold observations were removed at each split because such observations differed only in the time of the follow-up evaluation; not removing them from the training fold would lead to overoptimistic performance scores. Furthermore, instead of all possible positive-negative pairs, as in typical LPOCV, only a random subset of 300 positive-negative pairs was used at each split, which allowed to substantially reduce the computation time.

Finally, the random search hyperparameter optimization in this study was realized in the following way:

1. Select 300 random samples from the hyperparameter space.

2. Evaluate model performance for each hyperparameter sample using cross-validation.

3. Retrain the model using all available data with the hyperparameter configuration that maximized the cross-validated AUC.

Confidence intervals for the model tuning AUC estimates were calculated with BCa bootstrap.

### 5.1.5 Model evaluation

#### Comparison of machine learning algorithms

The machine learning algorithms described before were compared in terms of their influence on the average predictive performance of the final model. The classifiers, the feature selection methods, and the data sampling techniques were evaluated separately. Furthermore, the analysis was performed independently for the time-specific and the longitudinal models.

The statistical significance of the differences between the algorithms was evaluated with the Friedman tests (Friedman 1937, 1940) followed by the Nemenyi *post hoc* analysis (Nemenyi 1963). The Friedman test is a non-parametric test suitable for comparison of the difference among multiple related samples. For instance, seven different classifiers evaluated on 50 different models. The Friedman test computed the average performance ranks of the algorithms and tested whether they had the same influence on the AUC score. If the null hypothesis was rejected, the Nemenyi *post hoc* analysis followed. To do the Nemenyi

test, one has to first calculate the critical difference (CD), given by

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}\,, \tag{5.1}$$

where $N$ is the number of models, $k$ is the number of algorithms, and $q_\alpha$ is the critical value based on the Studentized range statistic divided by $\sqrt{2}$. When the average performance ranks of two algorithms differed by at least the CD, they were significantly different.

The analysis was repeated six times to test the classifiers, the feature selection algorithms, and the data sampling methods separately in the time-specific and the longitudinal models. The family-wise error rate (FWER) of the Friedman tests, that is the probability of at least one incorrect rejection of a true null hypothesis in any of the comparisons, was controlled with the Holm-Bonferroni method. The significance level for the FWER was set to 0.05.

#### Generalization performance

Hyperparameter optimization (model tuning) comes at a cost. On the one hand, it allows to tune the model so it well fits the training data. On the other hand, a favorable selection of hyperparameters can result in overoptimistic performance of the tuned model. To estimate performance of a tuned model on new, unseen data, that is, the *generalization* performance, the data used for hyperparameter optimization must be separate from the data used for model testing. A common way to do it, is to divide the data set into three parts: the training subset, the validation subset, and the testing subset. The model is trained on the training subset, the hyperparameter values are chosen based on the model's performance on the validation subset, and the generalization performance is estimated on the testing subset. In small data sets, however, the particular way the data set is divided into the three subsets can substantially influence the hyperparameter values and the generalization performance. For this reason, due to the modest size of the HD data set, instead of dividing the data to training, validation, and test folds, the models were tested with nested-cross validation (Cawley & Talbot 2010).

Nested cross-validation is essentially cross validation within cross validation. Part of the data is set aside for testing and the rest is used for model tuning. Next, the tuned model is tested on the part of data previously set aside for testing. Then, the procedure is repeated, that is another randomly selected part of the data is set aside for testing and the rest is used for model tuning. This is repeated until the desired number of iterations is achieved.

Unfortunately, it was not feasible to calculate the expected generalization performance of all 1960 models due to high computation cost. For this reason, nested cross-validation estimates were calculated only for best performing models in model tuning stratified by endpoint and classifier. This resulted in seven models per endpoint and 28 models in total. The inner loops of the nested cross-validation, which were responsible for model tuning, were the same as described in section dedicated to model tuning. That is, 300 hyperparameter samples were evaluated with MCCV with 300 splits and 10% of observations held

out for testing at each split for the time-specific models and modified LPOCV with 300 positive-negative pairs for the longitudinal models. The outer loops were realized by the MCCV with 100 splits and a 10% test fold for the time-specific models and the modified LPOCV with 100 positive-negative pairs for the longitudinal models. Confidence intervals for the generalization AUCs were calculated with BCa bootstrap.

### 5.1.6 Software

For the described analysis, an automated pipeline was built using existing libraries available for the Python programming language and custom code was written for specific functionalities required for this study. Open-source Python libraries:

- **scikit-learn** (Pedregosa et al. 2011) - machine learning library providing classes for model training, tuning, and testing. Also, the package provided classes for UFS and RFE feature selection methods and all classifiers except GTB.
- **XGBoost** (Chen & Guestrin 2016) - GTB classifier
- **imbalanced-learn** (Lemaitre et al. 2017) - sampling and data cleaning algorithms
- **Orange** (Demšar et al. 2013) - Friedman and Nemenyi tests as well as critical difference plots
- **StatsModels** (Seabold & Perktold 2010) - FWER and FDR
- **NumPy & SciPy** (Van der Walt et al. 2011), **Pandas** (McKinney 2010), **Matplotlib** (Hunter 2007) - miscellaneous libraries for data handling and visualization.

The author's scripts and extensions:

- **Model architecture** - a collection of classes and functions allowing to combine classes from various libraries into a single model and enable model training, tuning, and testing.
- **Feature-group selection** - a class providing feature-group selection described in section 5.1.4
- **Model-based feature selection** - a class providing model-based feature selection described in appendix B.2
- **Modified LPOCV** - a class implementing modified leave-pair-out cross-validation (LPOCV) described in section 5.1.4

## 5.2 Results

### 5.2.1 Mean-dose and morphological models

The predictive performance scores of the mean-dose models and the morphological model are presented in Table 5.1. The mean-dose models failed to predict xerostomia (AUC $<$

0.60) at all time-intervals as well as in the longitudinal approach. The morphological model achieved fair performance (AUC = 0.64) only in predicting long-term xerostomia.

### 5.2.2 Comparison of classification, feature selection, and sampling algorithms

There was a clear difference in the average performance between early (AUC $\approx$ 0.60), late (AUC $\approx$ 0.70), and long-term (AUC $\approx$ 0.90) xerostomia models (Figure 5.2). After applying the Holm-Bonferroni correction, all the Friedman tests were significant at the FWER = 0.05. Therefore, classification, feature selection, and sampling algorithms were compared with Nemenyi *post hoc* analysis for both the time-specific and the longitudinal models.

In the time-specific models, the support vector machine was by far the best scoring classifier, outperforming the other classifiers in over 70% of cases (Figure 5.3), whereas gradient tree boosting was on average the worst performing classifier (Figure 5.4). Conversely, gradient tree boosting together with support vector machines and extra-trees predicted xerostomia significantly better than all the other classifiers in the longitudinal approach.

The logistic regression-based algorithms performed significantly better than the feature selection methods based on extra-trees, in both the time-specific and the longitudinal models. Interestingly, while univariate feature selection by mutual information was the worst performing feature selection method in the time-specific models, it was one of the best in the longitudinal approach. Not performing feature selection was not disadvantageous in terms of predictive performance.

In both the time-specific and the longitudinal approach, no sampling algorithm gave a significant advantage over no sampling at all. In the time-specific models, Tomek links and the neighborhood cleaning rule performed significantly better than any oversampling algorithm. In the longitudinal models, Tomek links performed significantly better than random oversampling or ADASYN.

### 5.2.3 Generalization performance

The best performing models stratified by endpoint and classifier are listed in Table 5.2. These models were retested by nested cross-validation to estimate their generalization performance. Early xerostomia (0–6 months after treatment) was predicted fairly well only by the k-nearest neighbors classifier (AUC = 0.65). The models of late xerostomia (6–15 months after treatment) generalized slightly better with logistic regression, k-nearest neighbors, and gradient tree boosting scoring the AUC over 0.60. For long-term xerostomia (15–24 months after treatment), the models generalized best with the AUC ranging from 0.74 (k-nearest neighbors) to 0.88 (extra-trees). The longitudinal models failed to generalize except the gradient tree boosting classifier, which achieved the AUC of 0.63. Generalization AUCs were on average 0.10 lower than tuning AUCs for all the analyzed endpoints.

**Table 5.1:** Predictive performance of the mean-dose models and the morphological model proposed by Buettner et al. (2012). i - ipsilateral gland, c - contralateral gland, b - both glands.

| Endpoint | Model | AUC |
|---|---|---|
| Early | $\text{mean}^{i}$ | 0.58 (0.56–0.60) |
| | $\text{mean}^{c}$ | 0.42 (0.41–0.44) |
| | $\text{mean}^{b}$ | 0.50 (0.48–0.53) |
| | $\text{mean}^{i}$, $\text{mean}^{c}$ | 0.49 (0.48–0.51) |
| | morphological | 0.42 (0.40–0.44) |
| Late | $\text{mean}^{i}$ | 0.48 (0.44–0.51) |
| | $\text{mean}^{c}$ | 0.58 (0.55–0.61) |
| | $\text{mean}^{b}$ | 0.55 (0.52–0.58) |
| | $\text{mean}^{i}$, $\text{mean}^{c}$ | 0.54 (0.51–0.57) |
| | morphological | 0.59 (0.56–0.62) |
| Long-term | $\text{mean}^{i}$ | 0.40 (0.37–0.44) |
| | $\text{mean}^{c}$ | 0.58 (0.55–0.61) |
| | $\text{mean}^{b}$ | 0.56 (0.52–0.60) |
| | $\text{mean}^{i}$, $\text{mean}^{c}$ | 0.47 (0.44–0.50) |
| | morphological | 0.64 (0.60–0.67) |
| Longitudinal | $\text{mean}^{i}$ | 0.51 (0.45–0.56) |
| | $\text{mean}^{c}$ | 0.57 (0.51–0.62) |
| | $\text{mean}^{b}$ | 0.50 (0.44–0.55) |
| | $\text{mean}^{i}$, $\text{mean}^{c}$ | 0.52 (0.46–0.58) |
| | morphological | 0.55 (0.49–0.60) |

### 5.2.4 Model interpretation

Only the models predicting long-term xerostomia achieved high generalization scores, that is the AUC greater than 0.70. For that reason, model interpretation was performed only for this endpoint. The multivariate models of long-term xerostomia relied mostly on the parotid gland volume, the spread of the contralateral dose-volume histogram, and the parotid gland eccentricity (Figure 5.5). The contralateral dose gradient in the right-left direction, despite good univariate predictive power, was included in only one model.

## 5.3 Discussion

The mean-dose models failed to recognize patients at risk of early, late, and long-term xerostomia. The morphological model predicted patients at risk of long-term xerostomia fairly well confirming the findings of Buettner et al. (2012) that morphological descriptors of the dose distribution can improve prediction of radiation-induced xerostomia.

There was no multivariate model that would achieve generalization AUC above 0.65 for early or late effects, even though a few univariate models of late xerostomia exceeded that value. Similarly, the multivariate models of long-term xerostomia, despite their good generalization scores ($\text{AUC}_{\text{max}} = 0.88$), performed on a par with the univariate models based on the parotid volume or the contralateral dose gradient in the patient's right-left direction. Comparable performance of the univariate and the multivariate models could be caused by the small sample size, especially the small minority class. In such setting, the distribution

**Table 5.2:** Expected generalization performance of selected models evaluated by nested cross-validation. Best performing models at a given endpoint are marked with *.

| Endpoint | Classifier | Feature selection | Sampling | AUC tuning | AUC testing |
|---|---|---|---|---|---|
| Early | LR-L1 | RFE-ET | NCL | 0.62 (0.60–0.64) | 0.56 (0.53–0.60) |
| | LR-L2 | RFE-LR | NCL | 0.62 (0.60–0.64) | 0.46 (0.42–0.49) |
| | LR-EN | MB-ET | NCL | 0.62 (0.60–0.64) | 0.54 (0.50–0.57) |
| | kNN | UFS-F | SMOTE+ENN | 0.68 (0.66–0.70) | 0.65 (0.62–0.68)* |
| | SVM | UFS-F | NONE | 0.70 (0.68–0.72) | 0.57 (0.53–0.61) |
| | ET | MB-LR | NCL | 0.63 (0.61–0.65) | 0.44 (0.41–0.47) |
| | GTB | UFS-F | NONE | 0.66 (0.64–0.68) | 0.55 (0.51–0.59) |
| Late | LR-L1 | RFE-LR | NCL | 0.78 (0.75–0.80) | 0.63 (0.56–0.69) |
| | LR-L2 | RFE-LR | NCL | 0.76 (0.73–0.78) | 0.60 (0.53–0.66) |
| | LR-EN | MB-LR | SMOTE+TL | 0.73 (0.70–0.76) | 0.56 (0.51–0.62) |
| | kNN | MB-LR | NCL | 0.78 (0.76–0.80) | 0.62 (0.57–0.67) |
| | SVM | UFS-F | TL | 0.80 (0.77–0.82) | 0.52 (0.46–0.58) |
| | ET | RFE-ET | NCL | 0.78 (0.75–0.80) | 0.55 (0.50–0.61) |
| | GTB | MB-LR | OSS | 0.77 (0.75–0.79) | 0.65 (0.59–0.70)* |
| Long-term | LR-L1 | MB-LR | ROS | 0.95 (0.94–0.96) | 0.86 (0.80–0.90) |
| | LR-L2 | MB-LR | NONE | 0.96 (0.95–0.97) | 0.86 (0.81–0.90) |
| | LR-EN | MB-LR | SMOTE+ENN | 0.92 (0.90–0.93) | 0.83 (0.76–0.88) |
| | kNN | UFS-MI | TL | 0.88 (0.86–0.90) | 0.74 (0.68–0.80) |
| | SVM | RFE-LR | ENN | 0.94 (0.92–0.96) | 0.79 (0.73–0.85) |
| | ET | MB-LR | ENN | 0.93 (0.92–0.94) | 0.88 (0.84–0.91)* |
| | GTB | UFS-F | ROS | 0.89 (0.86–0.91) | 0.77 (0.71–0.83) |
| Longitudinal | LR-L1 | UFS-MI | NONE | 0.63 (0.57–0.68) | 0.52 (0.41–0.61) |
| | LR-L2 | RFE-LR | NCL | 0.60 (0.55–0.66) | 0.39 (0.29–0.48) |
| | LR-EN | UFS-MI | TL | 0.62 (0.57–0.68) | 0.52 (0.42–0.60) |
| | kNN | UFS-MI | NCL | 0.65 (0.61–0.69) | 0.58 (0.49–0.66) |
| | SVM | UFS-MI | OSS | 0.66 (0.60–0.71) | 0.57 (0.46–0.66) |
| | ET | UFS-MI | TL | 0.66 (0.61–0.71) | 0.51 (0.40–0.60) |
| | GTB | RFE-LR | ROS | 0.68 (0.62–0.72) | 0.63 (0.52–0.71)* |

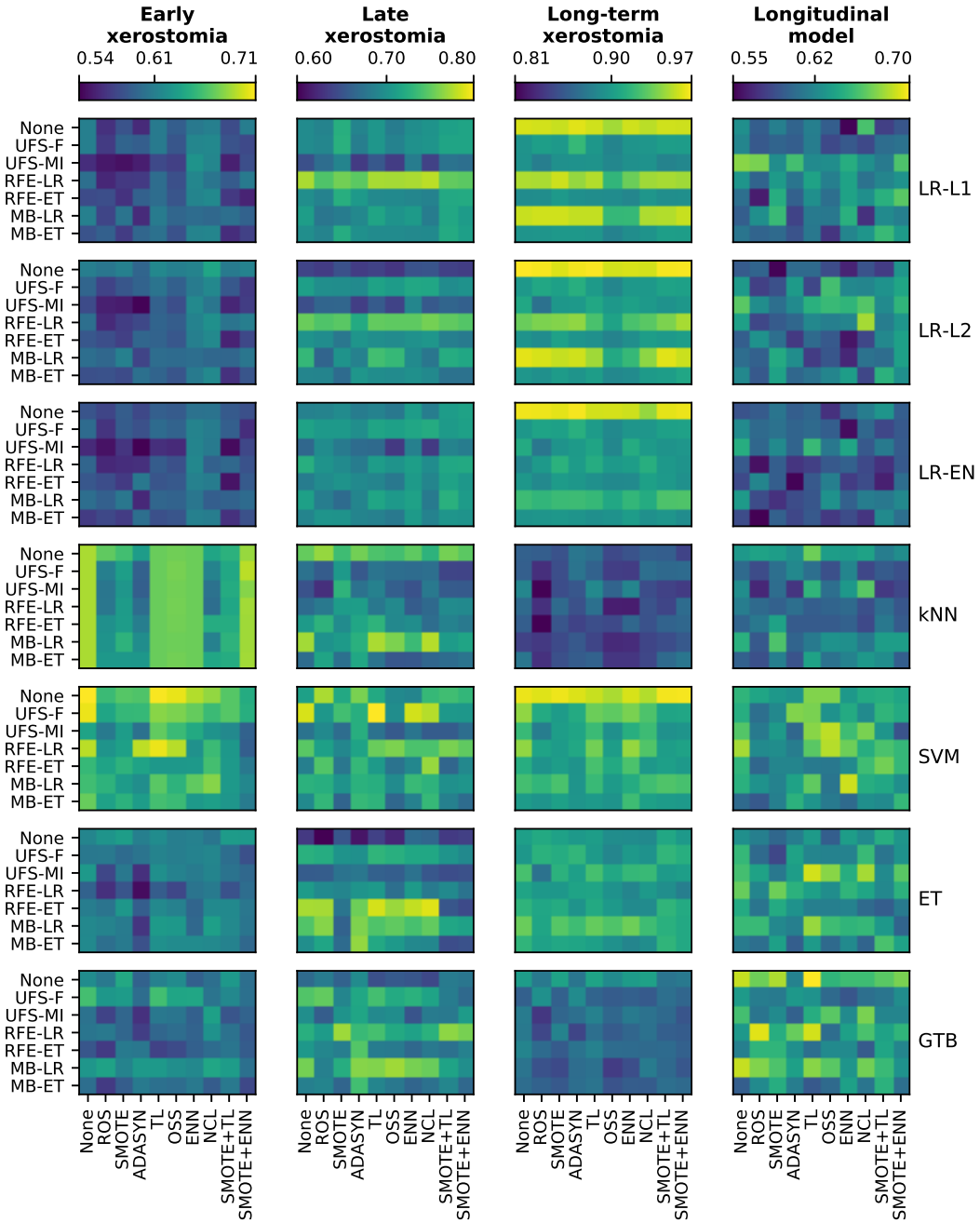



**Figure 5.2:** A comparison of classification, feature selection, and sampling algorithms in terms of their predictive performance in model tuning. All heat maps in a given column belong to a single endpoint, whereas all heat maps in a given row correspond to a single classifier. In each heat map, rows represent feature selection algorithms and columns correspond to sampling methods. The color maps are normalized per endpoint. The color bar ticks correspond to the worst, average, and the best model performance.

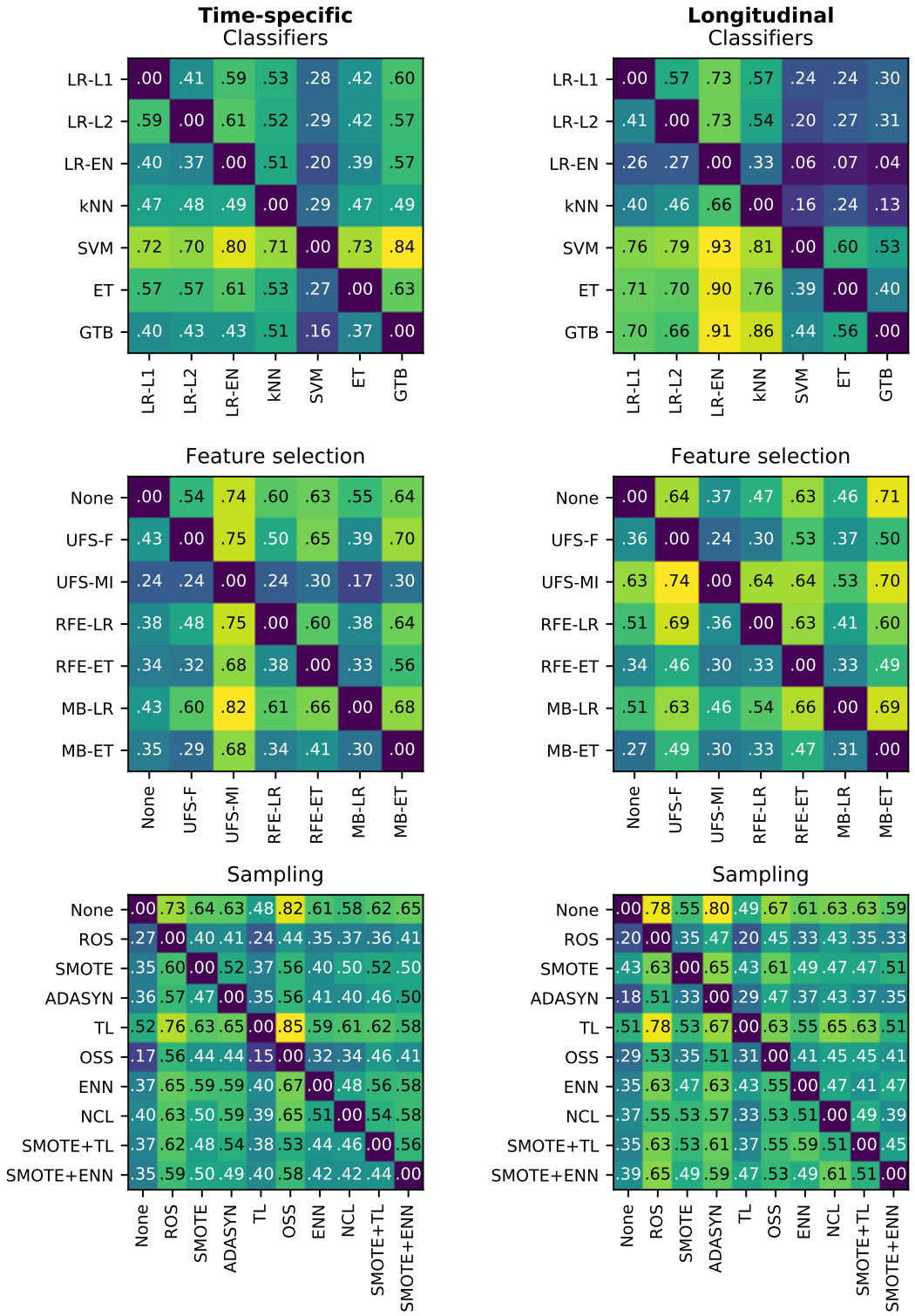



**Figure 5.3:** Heat maps showing a proportion of times a given algorithm on the vertical axis outperformed another algorithm on the horizontal axis in terms of the best AUC in model tuning.

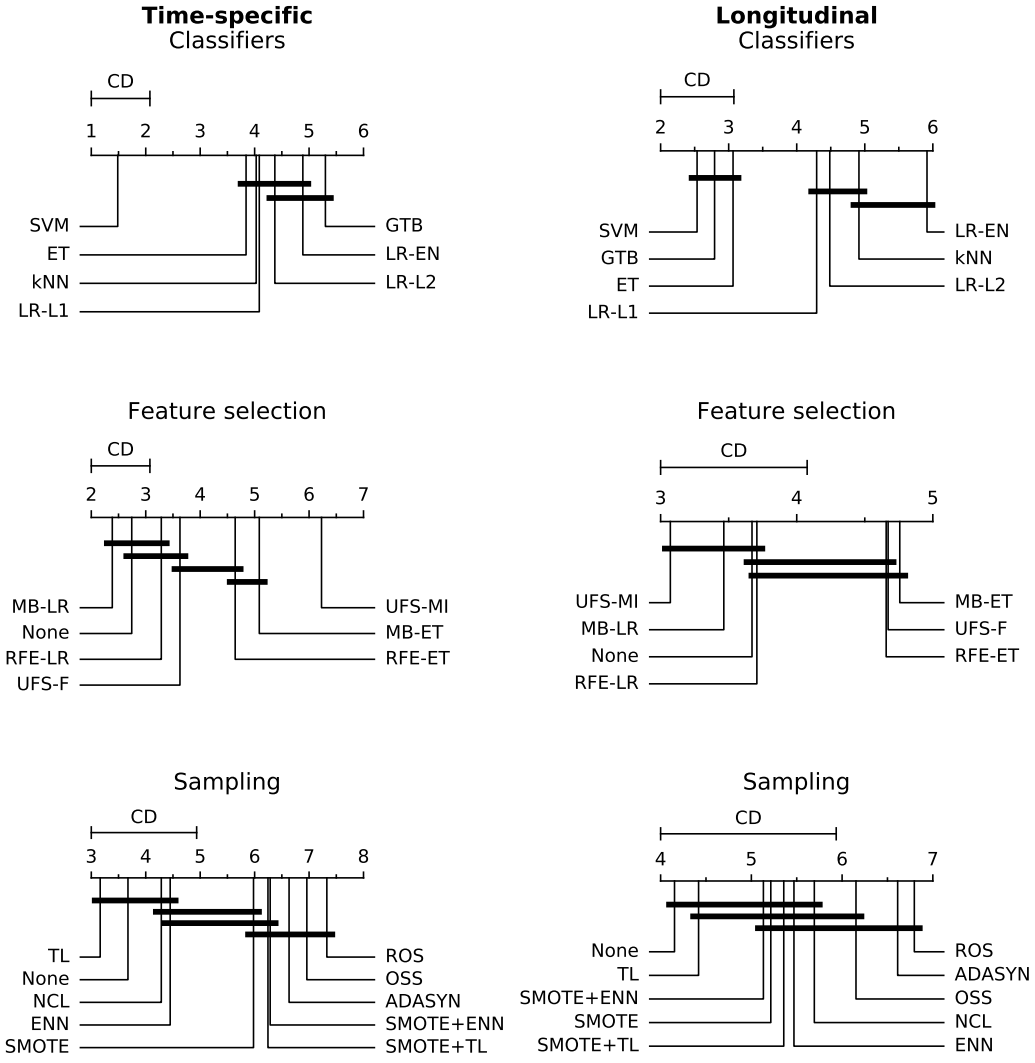



**Figure 5.4:** A comparison of classification, feature selection, and sampling methods against one another with the Nemenyi test. Lower ranks correspond to better performance of the algorithm, that is rank 1 is the best. Algorithms which ranks differ by less than the critical difference (CD) are not significantly different at 0.05 significance level and are connected by the black bars.

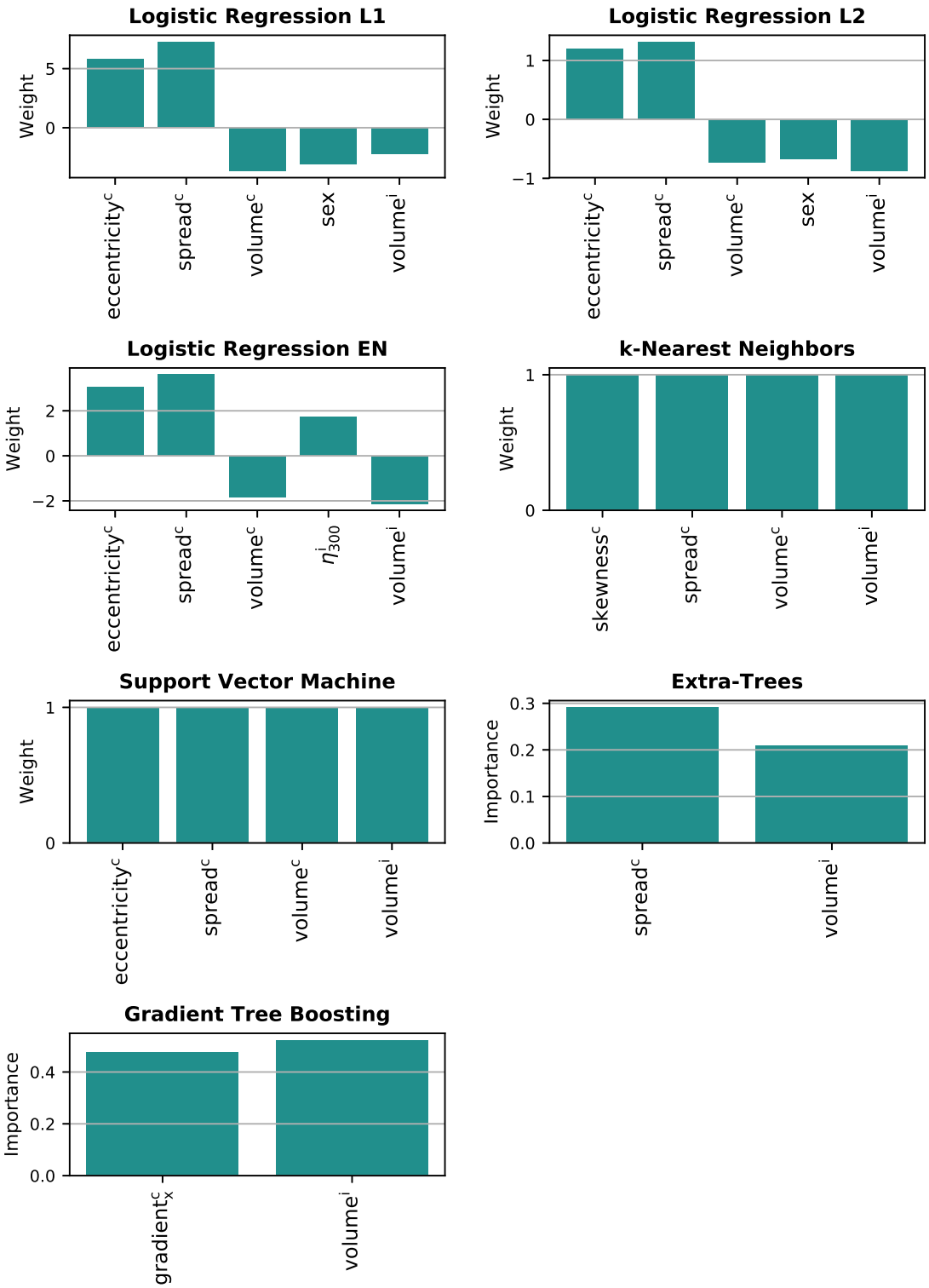



**Figure 5.5:** Features underlying the multivariate models of long-term xerostomia. i - ipsilateral gland, c - contralateral gland.

of model covariates can nonnegligibly differ between training and testing folds, hindering model training and reducing performance of the model.

The analysis of the multivariate models highlighted the importance of personalized treatment planning in radiotherapy. The models were strongly based on patient-specific and dose-independent features, such as parotid volume, parotid eccentricity, and the patient's sex. Interestingly, the dose gradient, despite relatively high predictive power, was included in only one model. Instead, the most common dosiomic feature was the spread of the contralateral dose-volume histogram quantifying the standard deviation of the dose within a parotid gland. Nevertheless, due to the geometry of the problem, the DVH spread and spatial dose gradients measured a similar characteristic of the dose distribution. That is, a large spread of the DVH was present when part of the parotid gland received high dose, whereas another part was spared.

In the time-specific models, the support vector machine was most commonly the best classifier. The other classifiers performed similarly to one another. The unexceptional performance of the ensemble methods (extra-trees and gradient tree boosting) could stem from the fact that complex models need more training samples to correctly learn the decision boundary. Among the longitudinal models, there was a more commonly observed classifier *ranking*, that is, GTB $>$ ET $>$ SVM $>$ LR $>$ kNN (Olson et al. 2017). Feature selection did not give a clear advantage over no feature selection in terms of the predictive performance. Nonetheless, feature selection allowed for a reduction of model complexity and made model interpretation easier. The best results were achieved with the logistic regression-based algorithms and feature selection by mutual information (only in the longitudinal models). There was no evidence that sampling methods improve accuracy of predictions. Moreover, it was observed that certain kinds of sampling, especially random oversampling, can significantly decrease predictive performance of the models.

Nested cross-validation proved to be an important step in the analysis. On average, the generalization AUCs were significantly lower than the AUCs achieved in model tuning. These findings corroborate the notion that single cross-validation can lead to overoptimistic performance estimates when hyperparameter tuning is involved in model building.

The results presented in this chapter show that the choice of a classifier and a feature selection algorithm can significantly influence predictive performance of the NTCP model. Moreover, in relatively small clinical data sets, simple logistic regression can perform as well as top-ranking machine learning algorithms, such as extra-trees or support vector machines. There was no significant advantage in using data cleaning or reducing the class imbalance. This study indicates the need for significantly larger patient cohorts to benefit from advanced classification methods, such as gradient tree boosting. It was shown that single cross-validation can lead to overoptimistic performance estimates when hyperparameter optimization is involved; either nested cross-validation or an independent test set should be used to estimate the generalization performance of a model.

It was demonstrated that in a highly conformal regime of modern radiotherapy, use of organ- and dose-shape features can be advantageous for modeling of treatment outcomes.

Moreover, due to strong dependence on patient-specific factors, such as the parotid shape and the patient's sex, the results highlight the need for development of personalized data-driven risk profiles in future NTCP models of xerostomia.

CHAPTER 6

# EXTERNAL VALIDATION

Predictive models, even carefully internally validated, tend to perform worse on external data than on the training data set (Moons et al. 2012). This behavior pertains especially to relatively small data sets (Bleeker et al. 2003), such as the HD cohort used for this study. Positive results of external validation broadens the impact of the research, indicating that the model is applicable beyond the institution in which it was developed. Successful validation on several external data sets can be considered the ultimate test for generalizability of any NTCP model. Nevertheless, if independent validation is not successful, it does not mean that the model is not useful. It means that its applicability may be restricted to patients treated at the institution in which the model was internally validated.

This chapter investigates how the models built on patient data from Heidelberg University Hospital, as detailed in chapter 2, generalize to a cohort from an external institution. For this purpose, patient data from the PARSPORT clinical trial was used for external validation of the models presented in chapters 4 and 5. The validation data was extracted during the author's research stay at the Institute of Cancer Research in London, United Kingdom.

## 6.1 Material and Methods

### 6.1.1 Endpoints

The endpoints selected for the external validation were late and long-term grade 2 or higher (G2+) xerostomia. This decision was motivated by the good performance of the univariate and the multivariate models at both of these time points (*vide* chapter 4 and 5).

As laid out in chapter 2, the follow-up reports in the HD cohort were collected at 3-month intervals, starting at the beginning of the treatment (Figure 4.1). In the PARSPORT cohort, in contrast, the follow-up reports were collected at 3, 6, 12, 18, and 24 months after the end of the therapy. For a meaningful comparison of the follow-up data between the cohorts, late xerostomia in the PARSPORT cohort was defined as the average complication score based on the reports from 6 and 12 months after treatment pooled together. For long-term xerostomia, it was realized by pooling the reports from 18 and 24 months after treatment. The endpoint definitions for the HD cohort were the same as in the previous chapters, that is 6–15 and 15–24 months after treatment for late and long-term xerostomia,

respectively.

Apart from the times of the follow-up report collection, there were also differences in the toxicity grading scales used in both cohorts. While the HD cohort applied the CTCAE scale, the PARSPORT cohort used the LENT-SOMA scale. However, the differences between the two scales can be considered negligible as the LENT-SOMA scale was mostly incorporated to the CTCAE grading system (*vide* chapter 2).

### 6.1.2 Covariate distribution

Patients in both cohorts were treated at different times, different institutions, and according to different protocols. It was expected that the distribution of predictors (covariates) may differ between patients treated in Heidelberg and patients from the PARSPORT trial. Investigation of such differences is important because most of machine learning algorithms is assume that the distribution of covariates within both the training and testing data sets is similar (Sugiyama & Kawanabe 2012).

To compare the distributions between the data sets, violin plots were generated for all the analyzed real-valued features. Violin plots are similar to box plots, however, they are more informative by showing the whole data distribution.

### 6.1.3 Model testing

The predictive power of individual features was compared between the HD and the PARSPORT cohorts. For this purpose, the AUC was calculated for each feature at late and long-term xerostomia endpoints.

In chapter 5, the best performing models stratified by endpoint and classifier were internally tested with nested cross-validation. Predictive performance of these models on the PARSPORT data was estimated and the results were compared to the results of the internal validation. The performance was measured with the AUC. Confidence intervals were estimated with BCa bootstrap.

Additionally, to investigate general association between model tuning and model testing scores, all late and long-term xerostomia models were evaluated on the PARSPORT data. The results were visualized with scatter plots.

Furthermore, the models at both time points were divided into well generalizing models and badly generalizing models. The threshold was the median generalization AUC among all models at a given time point. The *good* and the *bad* groups were compared in terms of the features on which the models were based.

## 6.2 Results

### 6.2.1 Covariate distribution

The comparison of covariate distribution is provided in Figures 6.1 and 6.2. The distribution of patients' age was comparable between the cohorts. Parotid gland volumes were

substantially larger in the PARSPORT cohort (median volume of 24 $cm^3$) than in the HD cohort (median volume of 18 $cm^3$). Also, parotid glands of the Heidelberg patients tended to be more round (greater sphericity). The difference in the mean dose was large, especially for the ipsilateral parotid. The absolute values of all three gradients were slightly smaller in the PARSPORT cohort than in the HD data set. The distributions of the majority of scale-invariant moments are markedly different between the two cohorts; typically, the spread of the dose-moments in the HD cohort was substantially broader.

### 6.2.2 Univariate models

The univariate analysis revealed large discrepancies in the predictive performance of the individual predictors between the cohorts. The results of the comparison are presented in Figure 6.3 for late xerostomia and in Figure 6.4 for long-term xerostomia.

For late xerostomia, good predictive power of the mean dose was observed (AUC = 0.74) in the PARSPORT cohort in contrast to the AUC of 0.58 in the HD cohort. Also, contralateral $\eta_{102}$ scored the AUC of 0.78, whereas the predictive power of this feature in the HD cohort was negligible. The only feature with substantial predictive power in both cohorts was the gradient in the anterior-posterior direction of the contralateral parotid gland ($\text{gradient}_y$) with AUCs close to 0.70.

For long-term xerostomia, the agreement in the predictive power of the contralateral $\text{gradient}_y$ was observed as well. Additionally, the contralateral gradient in the lateral (x) direction gave good results in both cohorts. On the other hand, high predictive power of the ipsilateral $\text{gradient}_x$ and $\text{gradient}_y$ in the HD cohort was not observed in the PARSPORT data. Furthermore, very high AUCs of parotid gland volumes in the HD cohort were not recapitulated in the PARSPORT cohort.

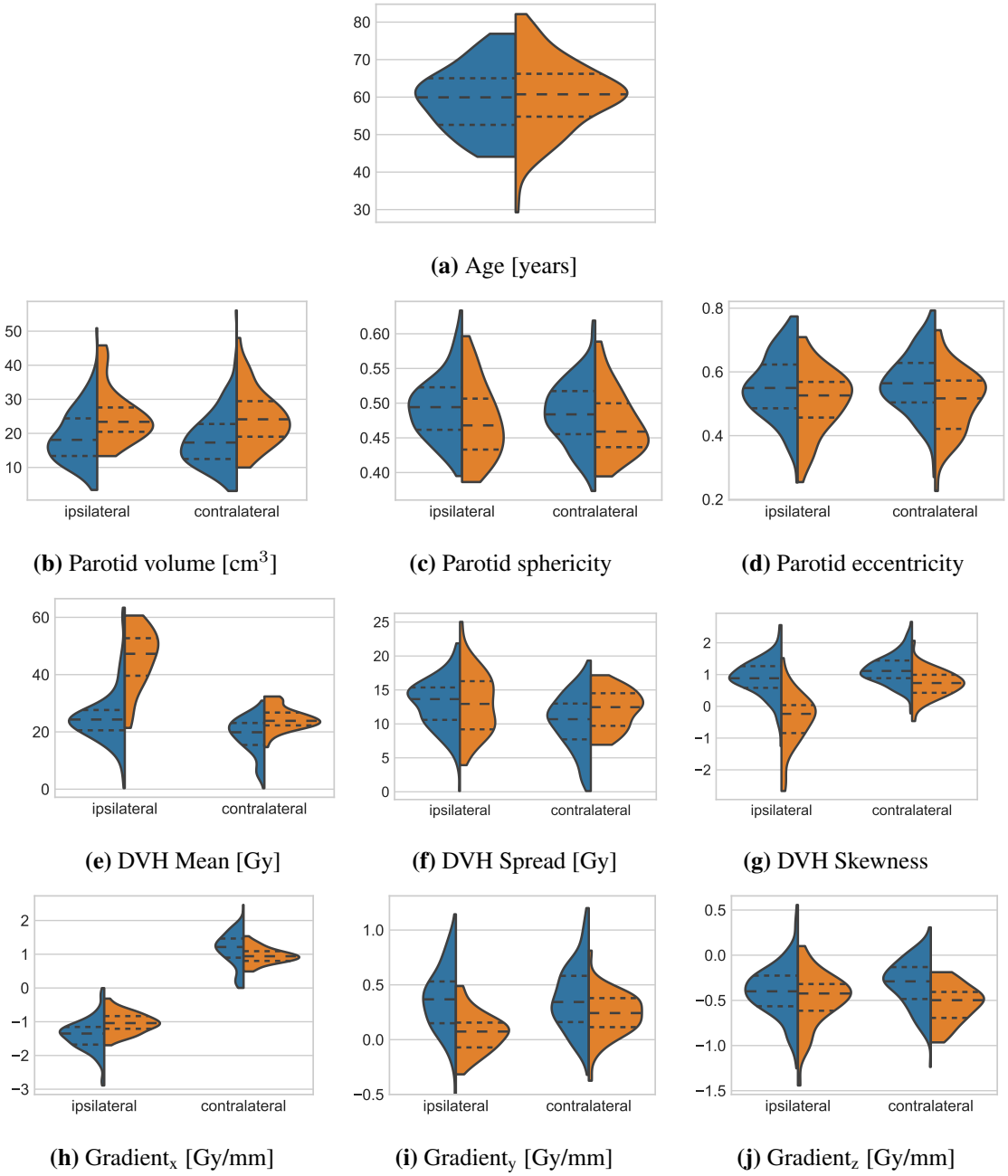
### 6.2.3 Multivariate models

The results of external validation of the models which were internally validated with nested cross-validation in chapter 5 are summarized in Table 6.1. The obtained AUCs were low for most of the models at both considered time points. The only models with reasonable performance were extra-trees for late xerostomia (AUC = 0.64) and gradient tree boosting for long-term xerostomia (AUC = 0.69). At the same time, however, the 95% confidence intervals of the obtained estimates were rather large.

The relationship between the tuning scores and the testing scores is presented in Figure 6.5. There was no clear correlation between the AUC scores achieved in model tuning and the AUC scores in model testing.

#### Feature analysis for late xerostomia models

All of the features that did well in the univariate analysis of late xerostomia were often included in the multivariate models (Figure 6.6a). Unexpectedly, the ipsilateral parotid




**(a)** Age [years]

**(b)** Parotid volume [$cm^3$]

**(c)** Parotid sphericity

**(d)** Parotid eccentricity

**(e)** DVH Mean [Gy]

**(f)** DVH Spread [Gy]

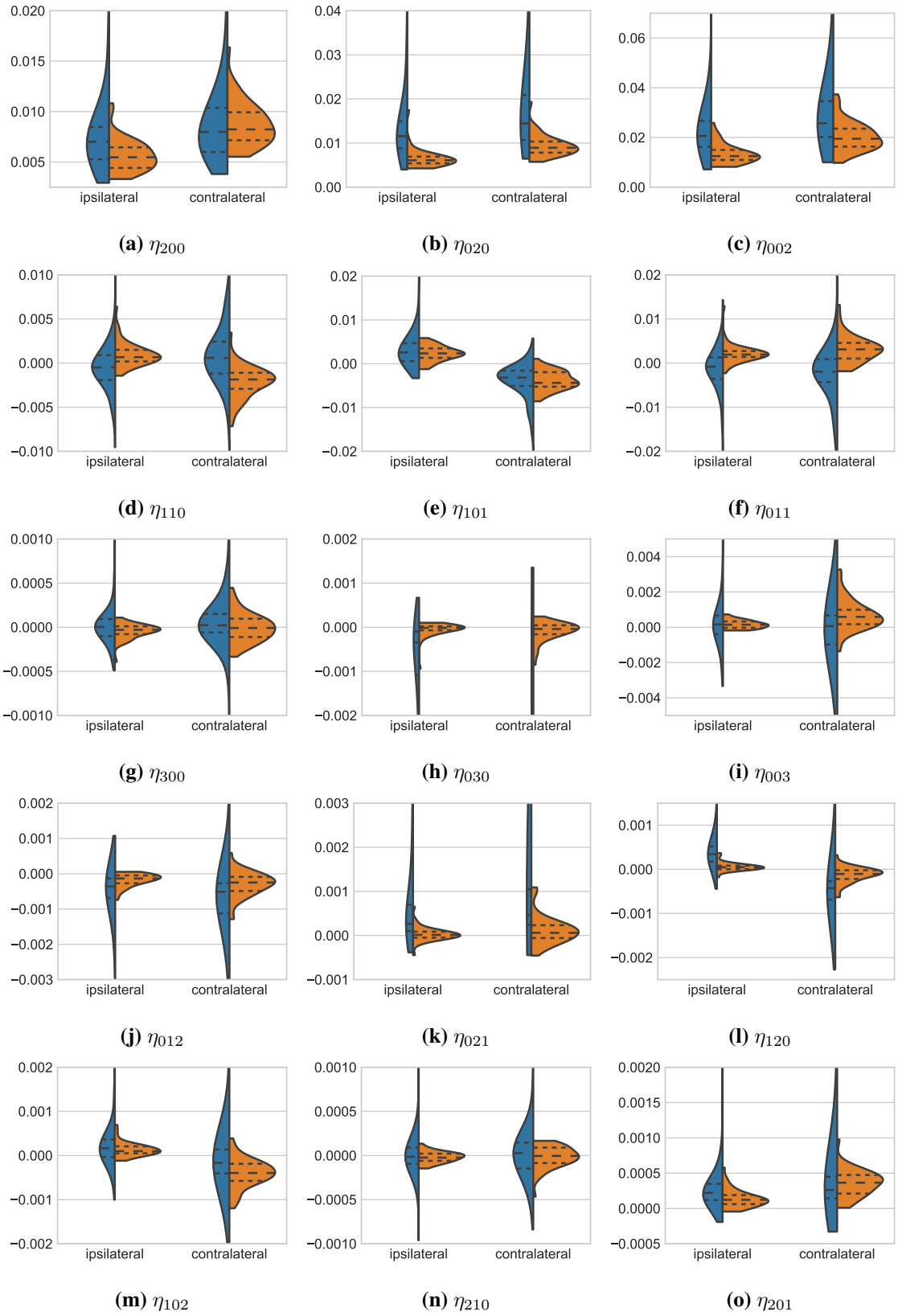
**(g)** DVH Skewness

**(h)** Gradient$_x$ [Gy/mm]

**(i)** Gradient$_y$ [Gy/mm]

**(j)** Gradient$_z$ [Gy/mm]

**Figure 6.1:** Age, parotid shape, dose-volume histogram, and spatial dose gradient features. Comparison between the HD (—) and the PARSPORT (—) cohorts. The distributions are visualized as kernel density plots. Wide dashed lines mark the medians, whereas narrow dashed lines correspond to the first and the third quartiles.




**(a)** $\eta_{200}$ **(b)** $\eta_{020}$ **(c)** $\eta_{002}$

**(d)** $\eta_{110}$ **(e)** $\eta_{101}$ **(f)** $\eta_{011}$

**(g)** $\eta_{300}$ **(h)** $\eta_{030}$ **(i)** $\eta_{003}$

**(j)** $\eta_{012}$ **(k)** $\eta_{021}$ **(l)** $\eta_{120}$

**(m)** $\eta_{102}$ **(n)** $\eta_{210}$ **(o)** $\eta_{201}$

**Figure 6.2:** Three-dimensional dose moments. Comparison between the HD (—) and the PARSPORT (—) cohorts. The distributions are visualized as kernel density plots. Wide dashed lines mark the medians, whereas narrow dashed lines correspond to the first and the third quartiles.

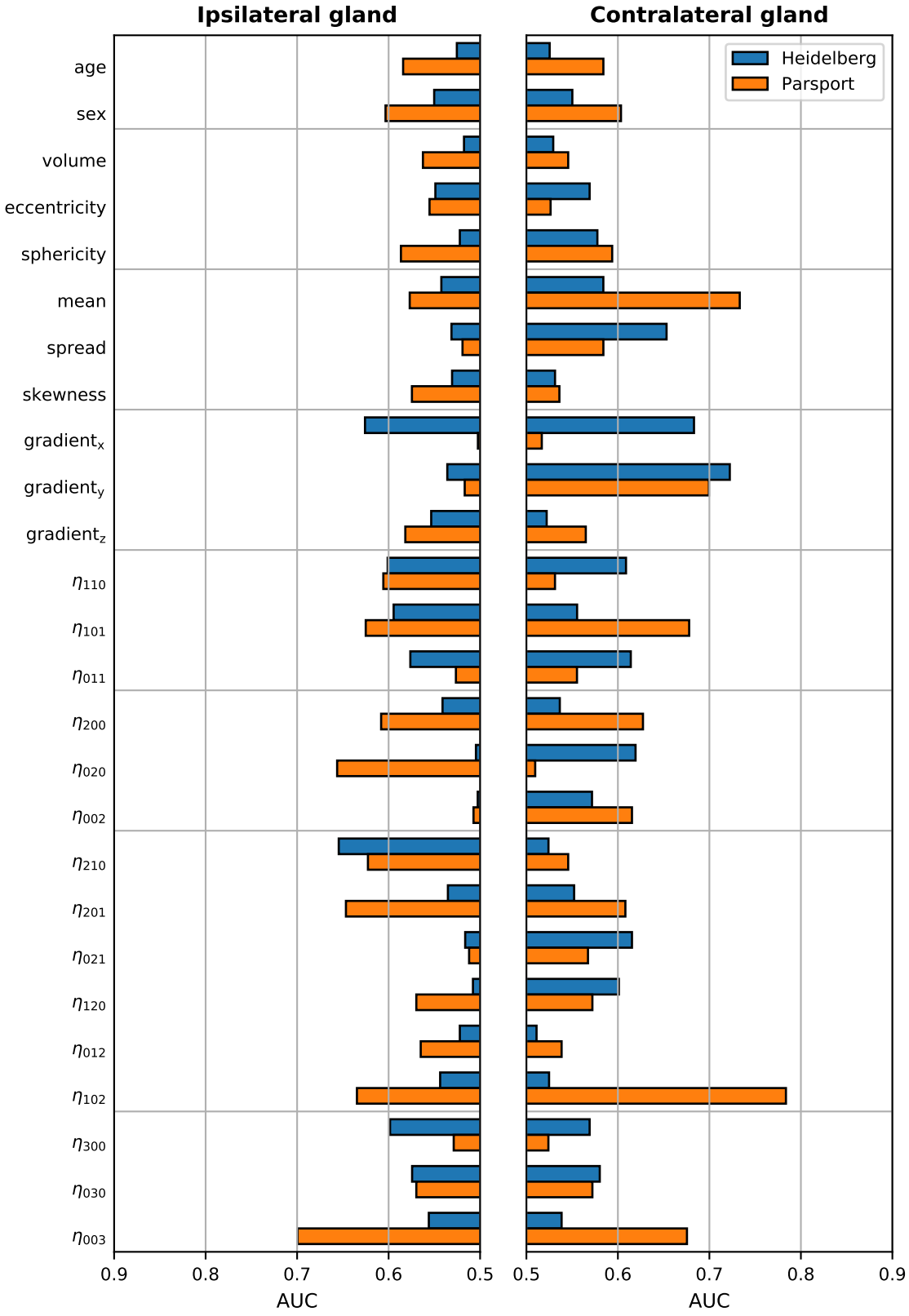



**Figure 6.3:** Predictive power of individual features for late xerostomia in the HD and the PARSPORT cohorts. Predictive power was measured with the AUC.

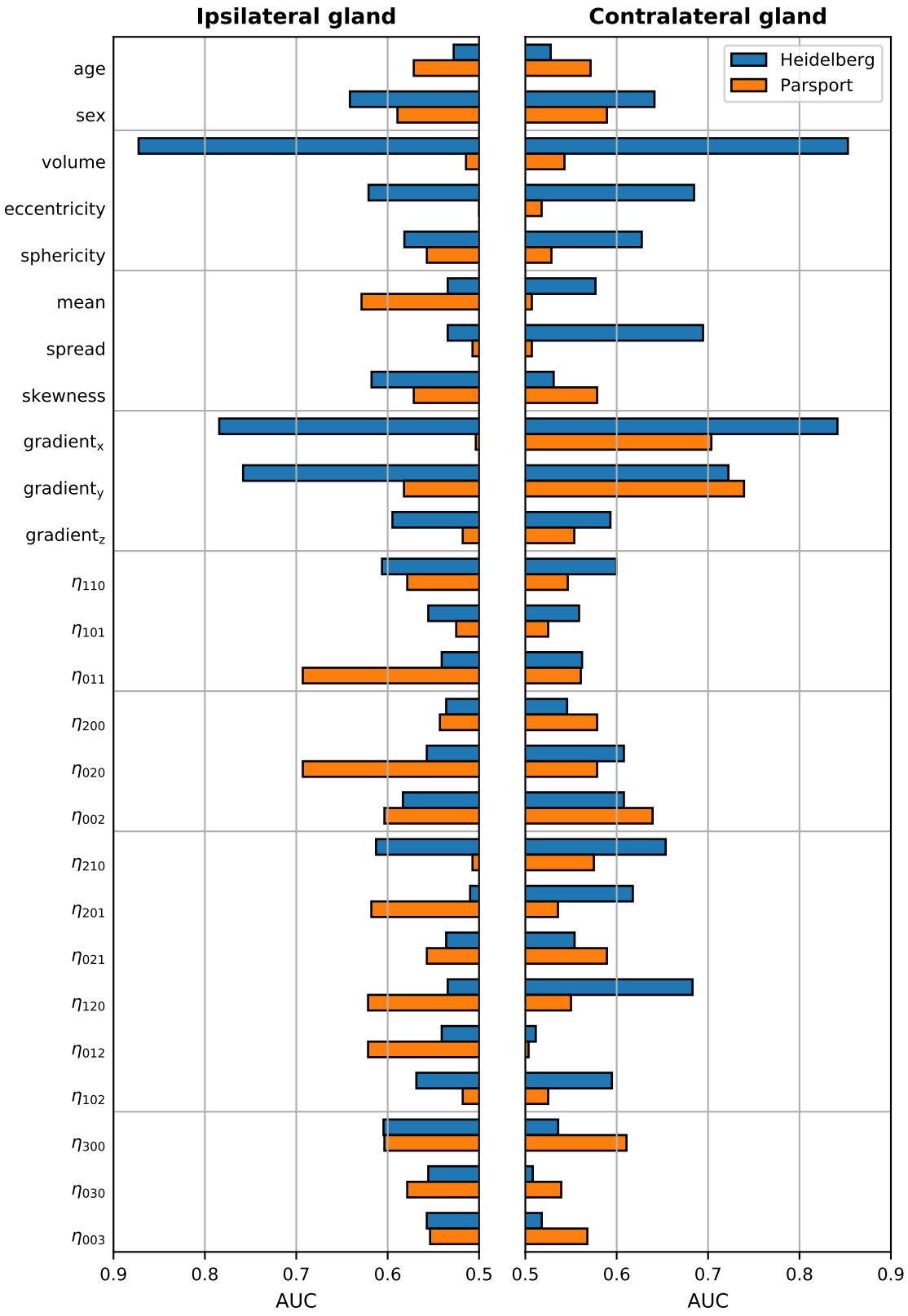
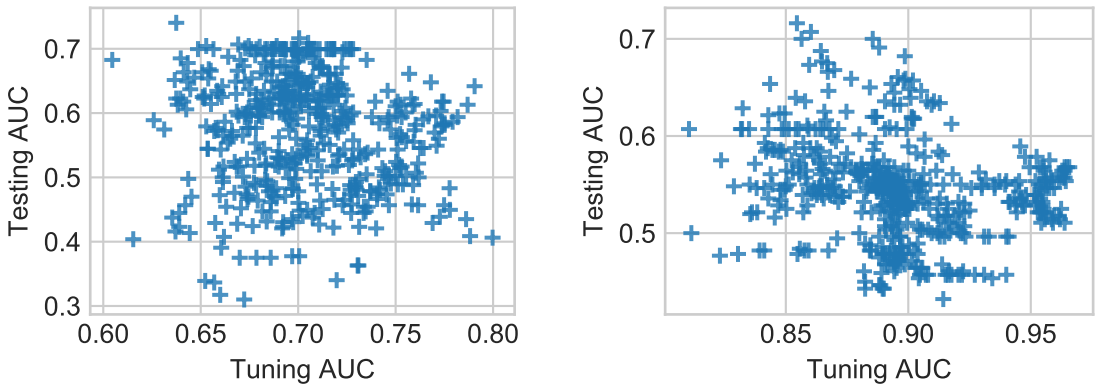



**Figure 6.4:** Predictive power of individual features for long-term xerostomia in the HD and the PARSPORT cohorts. Predictive power was measured with the AUC.

**Table 6.1:** Generalization performance of selected models evaluated by nested cross-validation and external validation.

| Endpoint | Classifier | Feature selection | Sampling | AUC tuning | AUC internal testing | AUC external testing |
|---|---|---|---|---|---|---|
| Late | LR-L1 | RFE-LR | NCL | 0.78 (0.75–0.80) | 0.63 (0.56–0.69) | 0.59 (0.40–0.75) |
| | LR-L2 | RFE-LR | NCL | 0.76 (0.73–0.78) | 0.60 (0.53–0.66) | 0.56 (0.36–0.72) |
| | LR-EN | MB-LR | SMOTE+TL | 0.73 (0.70–0.76) | 0.56 (0.51–0.62) | 0.60 (0.41–0.77) |
| | kNN | MB-LR | NCL | 0.78 (0.76–0.80) | 0.62 (0.57–0.67) | 0.45 (0.30–0.59) |
| | SVM | UFS-F | TL | 0.80 (0.77–0.82) | 0.52 (0.46–0.58) | 0.41 (0.25–0.59) |
| | ET | RFE-ET | NCL | 0.78 (0.75–0.80) | 0.55 (0.50–0.61) | 0.64 (0.46–0.79) |
| | GTB | MB-LR | OSS | 0.77 (0.75–0.79) | 0.65 (0.59–0.70) | 0.56 (0.38–0.72) |
| Long-term | LR-L1 | MB-LR | ROS | 0.95 (0.94–0.96) | 0.86 (0.80–0.90) | 0.53 (0.30–0.74) |
| | LR-L2 | MB-LR | NONE | 0.96 (0.95–0.97) | 0.86 (0.81–0.90) | 0.55 (0.32–0.77) |
| | LR-EN | MB-LR | SMOTE+ENN | 0.92 (0.90–0.93) | 0.83 (0.76–0.88) | 0.51 (0.29–0.71) |
| | kNN | UFS-MI | TL | 0.88 (0.86–0.90) | 0.74 (0.68–0.80) | 0.56 (0.44–0.69) |
| | SVM | RFE-LR | ENN | 0.94 (0.92–0.96) | 0.79 (0.73–0.85) | 0.50 (0.29–0.70) |
| | ET | MB-LR | ENN | 0.93 (0.92–0.94) | 0.88 (0.84–0.91) | 0.56 (0.36–0.75) |
| | GTB | UFS-F | ROS | 0.89 (0.86–0.91) | 0.77 (0.71–0.83) | 0.69 (0.53–0.84) |




**(a)** Late xerostomia **(b)** Long-term xerostomia

**Figure 6.5:** Distribution of AUCs achieved in model testing on the PARSPORT data against AUCs from model tuning on the HD data.

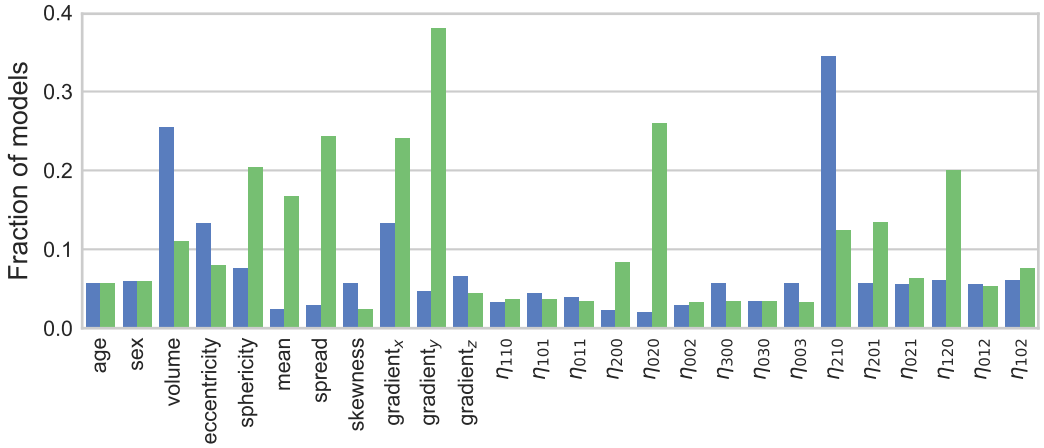
volume, the contralateral parotid sphericity, and the contralateral mean dose were included in the models of late xerostomia more often than the univariate analysis would suggest.

Over 80% of models that included gradient$_x^c$ or gradient$_x^i$ achieved better than average score in model testing (Figure 6.6b). Conversely, over 90% of models that relied on ipsilateral parotid volume, contralateral parotid sphericity or some other parotid-shape metric scored worse than average (Figure 6.6c).

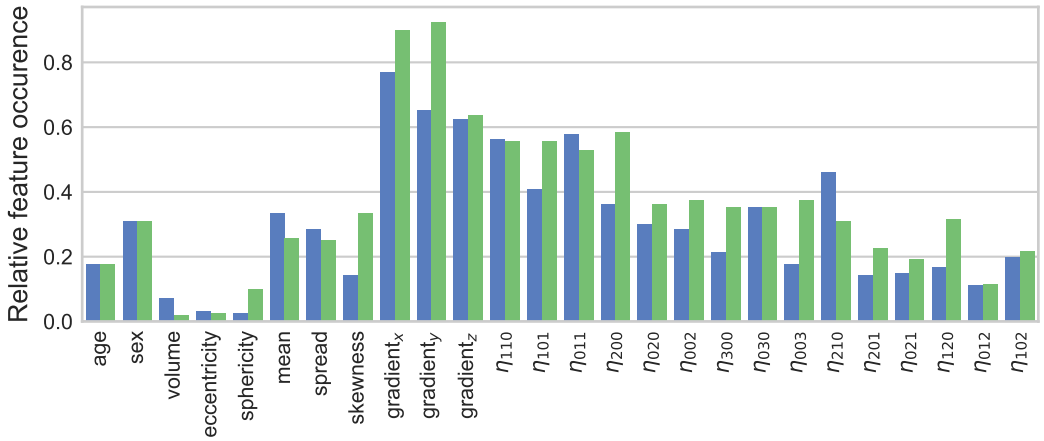
#### Feature analysis for long-term xerostomia models

In case of long-term xerostomia, the situation was slightly more complicated. In the univariate analysis the most informative features were parotid gland volumes (Figure 6.7a). Unsurprisingly, the vast majority of the models relied on the ipsilateral parotid volume (over 98%) or the contralateral parotid volume (over 80%). Other good performing features in the univariate analysis were the ipsi- and contralateral dose gradients in the right-left and the anterior-posterior directions. Remarkably, all these gradients, with the exception of the contralateral gradient$_x$, were rarely included in the multivariate models (less than 5% of the models). Moreover, the contralateral DVH spread was included in over 50% of the models, even though it had lower univariate predictive power.

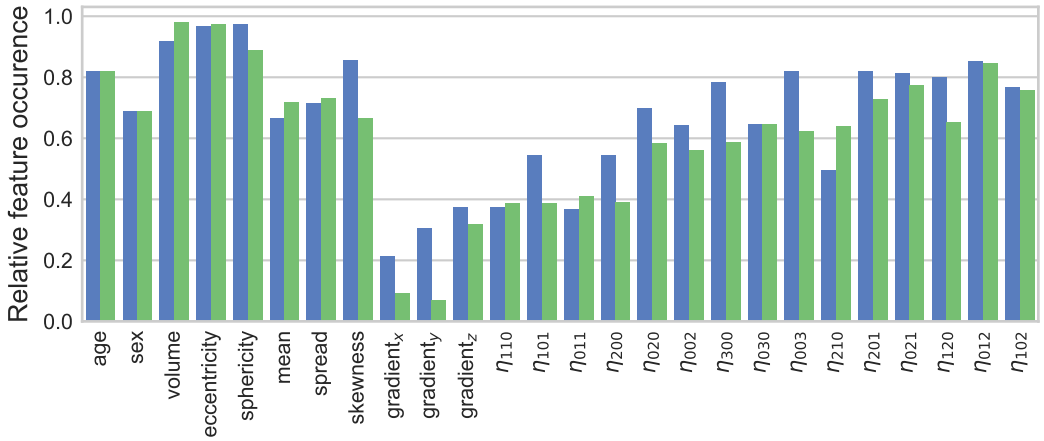
Due to the extremely high prevalence, ipsilateral parotid volume was approximately evenly distributed between the *good* and the *bad* models. Many features were present only in the well performing group (Figure 6.7b). Most of them, however, were included in only a couple of models. The only feature that was fairly often selected (approximately 30% of all models) and was usually included in the group that achieved better than average AUC was the contralateral gradient$_x$. Over 70% of models that included the contralateral gradient$_x$ generalized better than the average. As for the features present mainly in the worse generalizing group, the contralateral DVH spread was the most prevalent one (Figure 6.7c). Over 60% of models that included this covariate did worse than the average.




**(a)** Fraction of models in which a given feature was included.

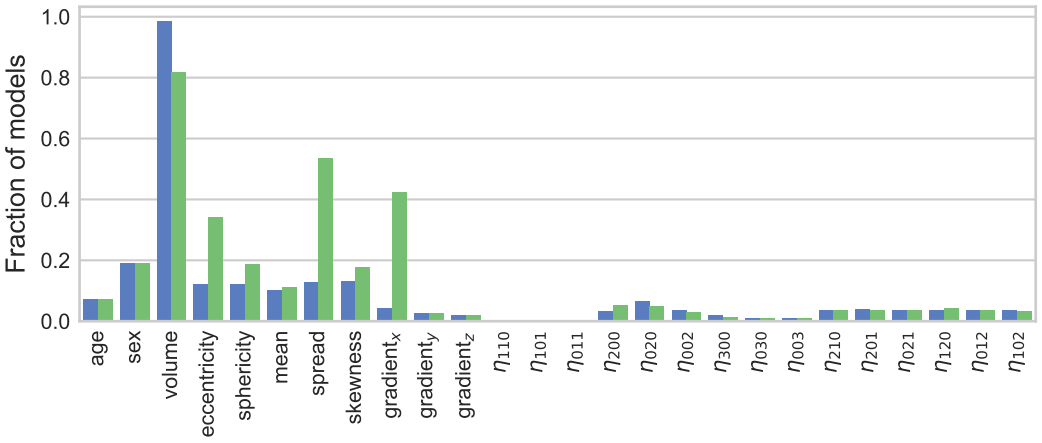



**(b)** Relative occurrence of features within models that performed better than average according to the testing AUC. For example, 1 corresponds to a feature that was only included in the better-than-average models, 0.5 corresponds to a feature that was equally often included in better- and worse-than-average performing models.

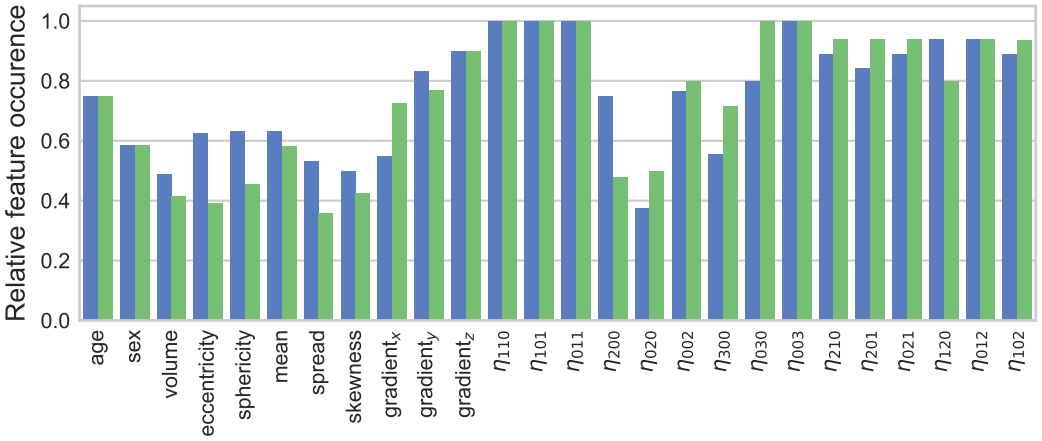



**(c)** Relative occurrence of features within models that performed worse than average according to the testing AUC. For example, 1 corresponds to a feature that was only included in the worse-than-average models, 0.5 corresponds to a feature that was equally often included in better- and worse-than-average performing models.

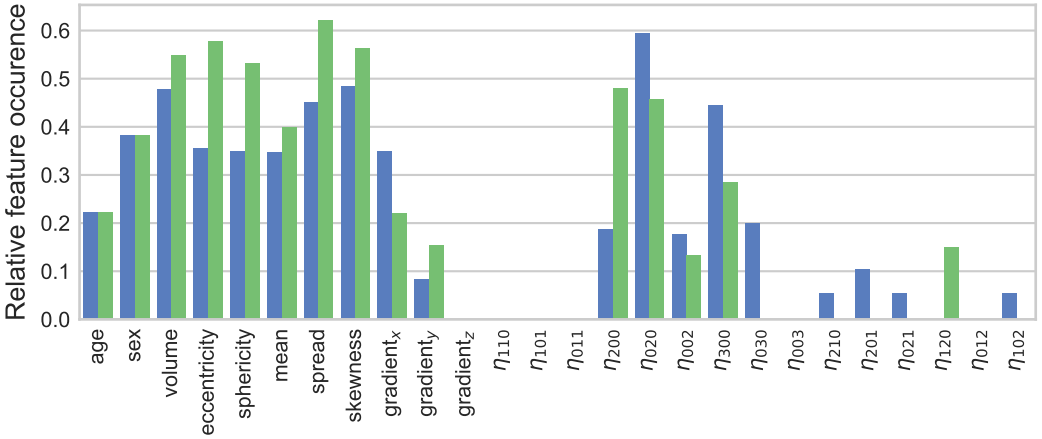
**Figure 6.6:** Features underlying multivariate models of late xerostomia. Ipsilateral features are marked with —, whereas contralateral features are marked with —.




**(a)** Fraction of models in which a given feature was included considering all models.




**(b)** Relative occurrence of features within models that performed better than average according to the testing AUC. For example, 1 corresponds to a feature that was only included in the better-than-average models, 0.5 corresponds to a feature that was equally often included in better- and worse-than-average performing models.




**(c)** Relative occurrence of features within models that performed worse than average according to the testing AUC. For example, 1 corresponds to a feature that was only included in the worse-than-average models, 0.5 corresponds to a feature that was equally often included in better- and worse-than-average performing models.

**Figure 6.7:** Features underlying multivariate models of long-term xerostomia. Ipsilateral features are marked with —, whereas contralateral features are marked with —.

## 6.3 Discussion

In this chapter the external validation of univariate- and multivariate NTCP models of late and long-term xerostomia was presented. In general, the analyzed models did not generalized well to the external cohort. Especially performance of the long-term xerostomia models was disappointing due to their high predictive performance in the HD cohort.

The head on comparison of covariate distributions between the training and validation data sets revealed that even though patients in both cohorts were treated with the goal of sparing parotid glands, the distribution of various features differed significantly between the groups. Moreover, the differences were apparent not only in the dosiomic features, which are treatment-specific, but also in radiomic features. Such differences may suggest different organ delineation protocols between the data sets. The difference in dosage to parotid glands was substantial, especially for the ipsilateral parotid which was much better spared in the HD cohort. The differences between the cohorts in the three-dimensional scale-invariant dose moments concerned not so much a shift of the distributions but rather the spread of the distributions. As nicely illustrated in the violin plots in Figure 6.2, the dose moments were significantly more scattered in the HD data set than in the PARSPORT cohort. This may suggest that for highly conformal treatments, spatial dose moments are too unstable.

The analysis revealed large discrepancies in the predictive performance of the individual predictors between the cohorts. For instance, high predictive power of the parotid volumes was not observed in the PARSPORT cohort, which may suggest that it was only a statistical fluke and a false positive. After all, the good performance of the volumes was observed only for the long-term xerostomia endpoint in the HD cohort and for no other endpoints. On the other hand, the volume definitions in the PARSPORT cohort seem to be slightly larger. Furthermore, even though parotid volume was not predictive of late xerostomia alone, it was the fourth most often included feature in the multivariate models of late xerostomia. This suggests that parotid volume may provide NTCP-related information complementary to other features.

The only feature group that exhibited good predictive performance in both cohorts were spatial dose gradients. A possible explanation of this finding could be that parotid glands typically shrink and move toward the medial direction during the course of radiotherapy. As a result, the gradient can be a proxy for the change of any dose-related metric subject to motion. As such, this might be an indicator of neglected motion and deformation effects during the modeling process.

Given the covariate shift between the HD and the PARSPORT cohorts it would be interesting to validate the results of chapters 4 and 5 against a data set which has a more similar distribution of the covariates. Furthermore, covariate-shift adaptation methods could be investigated (Sugiyama & Kawanabe 2012).

The observed marked differences in the covariate distributions between the cohorts together with the variability in predictive power of the features suggest it could be beneficial

in the future to report not only the predictors themselves but also some metrics of their distribution, such as median and interquartile range, when publishing novel multivariate NTCP models. In the previous chapter, one of the conclusions was that larger data sets could improve performance of state-of-the-art machine learning methods. The results presented in this chapter may, furthermore, suggest that not only larger but also more heterogeneous data sets could be beneficial. That is, data sets coming from different institutions, various modalities (tomotherapy, Volumetric Arc Therapy (VMAT), etc.) which could allow to build less cohort-specific but more general models.

CHAPTER 7

# CONCLUSIONS

This thesis investigated suitability of state-of-the-art machine learning and high-dimensional feature spaces for building NTCP models of xerostomia.

The analysis presented in chapter 3 showed that modern radiotherapy treatments in the HD cohort allowed to spare parotid glands below QUANTEC recommendations for 74% of patients. QUANTEC guidelines state that to avoid severe xerostomia, one should spare at least one parotid gland to a mean dose less than 20 Gy or both parotids to a mean dose less than 25 Gy. In the HD cohort, the former recommendation was satisfied in 54% of patients, whereas the latter one in 72% of patients. Consequently, this work dealt with xerostomia prediction models in highly conformal treatments.

The corresponding rates of xerostomia were low. Grade 2 (moderate) xerostomia pertained to approximately 15–20% of patients and no patient suffered from grade 3 (severe) xerostomia. These results show that a reduction of the mean dose to parotid glands beyond QUANTEC recommendations is beneficial and allows to reduce overall risk of G2+ xerostomia. Interestingly, conventional xerostomia prediction models, which are based on the mean dose, were not predictive of patients at risk of G2+ xerostomia in this low-dose regime. The presented findings indicate that other factors, apart from the mean dose to parotid glands, may influence the development of G2+ xerostomia and that these factors dominate in highly-conformal modern radiotherapy treatments.

The examination of 61 demographic, radiomic, and dosiomic features revealed high predictive power of parotid volumes and the spatial dose gradients within the parotid glands in the right-left and the anterior-posterior directions (*vide* chapters 4 and 5). Moreover, the good predictive performance of gradients was also recognized in the PARSPORT cohort (Figure 6.3 and 6.4). This finding is especially important because it indicates that the observed pattern is not limited to the HD data set alone. The consistent observation of gradients, being a predictive signal in two independent and strongly differing patient cohorts, may be interpreted as a strong indicator that the assumption of a completely parallel parotid gland structure for NTCP modeling (as done in mean-dose models) falls short of the actual organ complexity. Nonhomogeneous radiosensitivity of parotid glands was also observed by Luijk et al. (2015), who identified the stem cell-containing region responsible for the functional regeneration. However, predictors based on the spatial dose gradient within the

parotid are easy to calculate and do not require creating additional delineations during the treatment planning phase. Additionally, the high predictive power of dose gradients may have roots in a neglected organ motion and deformation effects. Parotid glands typically shrink and move toward the medial direction during the course of radiotherapy (Osorio et al. 2008). Effectively, this means that parotids move toward the high-dose region. In any case, in light of poor performance of the mean dose, the findings presented in this thesis undermine the applicability of the LKB model for xerostomia predictions in highly-conformal low-dose domains.

The comparison of various machine learning methods for the modeling of xerostomia was presented in chapter 5. The results showed that the choice of a classifier and a feature selection algorithm can significantly influence the predictive performance of an NTCP model. Interestingly, simple logistic regression often performed on a par with top-ranking machine learning algorithms, such as extra-trees and support vector machines. There is a lot of evidence from multiple studies and machine learning competitions that certain algorithms tend to work better than others (Olson et al. 2017). However, more complex models usually require more data to learn effectively. The advantage of larger training cohorts became apparent when the performance of machine learning classifiers between the time-specific and the longitudinal model was compared (Figure 5.4). Support vector machines and tree-based models clearly took advantage of a larger number of observations and performed significantly better than logistic regression and k-nearest neighbors. The results of this study indicate that significantly larger patient cohorts are needed to fully benefit from sophisticated classification methods, such as gradient tree boosting, in NTCP modeling.

The external validation of the models showed that most of the models built on the HD data set did not generalize to the PARSPORT cohort (*vide* chapter 6). The models that performed well in both data sets relied on the spatial dose gradients, especially within the contralateral gland (Figures 6.6 and 6.7). The underlying assumption of most supervised models in machine learning is that the distribution of covariates is similar between the training and the testing cohort. It was observed, however, that distributions of many predictors were extremely different between the HD and the PARSPORT cohorts (Figures 6.1 and 6.2). The patients in the PARSPORT trial were treated from 2003 to 2007 and the NTCP models based on the PARSPORT data were published in 2012 and 2013 (Buettner et al. 2012, Miah et al. 2013). The patients in the HD cohort were treated from 2010 to 2015. This means that at the time the patients from the HD cohort were being treated, the PARSPORT NTCP models that were about to be published could have already been unsuitable for some of the contemporary patient cohorts. For these reasons, it seems to be reasonable to include information about the covariate distribution when publishing novel NTCP models. Even the median value with the interquartile range would be informative, allowing to predetermine whether a given model might be applicable to a given cohort.

In order to perform the analysis presented in this thesis, an original multi-stage pipeline was developed. The pipeline implemented feature extraction, unsupervised feature selection, data cleaning, class balancing, supervised feature selection, and classification. It was

the first, to the author's knowledge, such a complex approach to NTCP modeling. Moreover, it is straightforward to use this pipeline for follow-up studies or the modeling of other indications. Selected parts of the pipeline were made publicly available on GitHub to facilitate similar studies for other researchers.

It would be interesting to analyze the performance of advanced machine learning algorithms on a lager patient cohort. A relatively small sample size of the HD data set often translated to wide confidence intervals of the estimated parameters and the performance metrics. Also, the applicability of sophisticated machine learning methods which require large number of samples could be better evaluated. With a sufficiently large cohort one could even think of designing an NTCP model based on a deep neural network architecture. Such models proved to be extremely effective in numerous computer vision tasks and require neither feature engineering nor feature selection. Deep neural networks are, however, difficult to interpret, which may impede adoption of such models in the clinic. Another interesting research direction could be investigation of image-guided NTCP models. On the one hand, high conformality of the dose distribution allows to efficiently spare organs at risk. On the other hand, any effects related to uncertainty in dose delivery during treatment are intensified. Image guidance could allow to keep track of these uncertainties and, perhaps, explain large part of unexplained variability observed in NTCP models based solely on the treatment planning data. The lack of image guidance data in the HD data set prevented correcting for interfractional anatomical changes that often take place during head-and-neck radiotherapy. Such analysis would be especially interesting, considering the observed high importance of the spatial dose gradients in this study. Last but not least, some studies indicated that the condition of submandibular glands after radiotherapy might affect the severity of patient-reported xerostomia (Dijkema et al. 2012). Due to the observed efficient sparing of parotid glands, the relative impact of submandibular glands injury on xerostomia is increasing. For this reason, inclusion of submandibular gland-based predictors may be beneficial for future NTCP models of xerostomia.

# SUMMARY

In routine clinical practice, the risk of xerostomia is typically managed by limiting the mean radiation dose to parotid glands. This approach used to give satisfying results. In recent years, however, several studies have reported mean-dose models to fail in the recognition of xerostomia risk. This can be explained by a strong improvement of overall dose conformality in radiotherapy due to recent technological advances, and thereby a substantial reduction of the mean dose to parotid glands. This thesis investigated novel approaches to building reliable normal tissue complication probability (NTCP) models of xerostomia in this context.

For the purpose of the study, a cohort of 153 head-and-neck cancer patients treated with radiotherapy at Heidelberg University Hospital was retrospectively collected. The predictive performance of the mean-dose to parotid glands was evaluated with the Lyman-Kutcher-Burman (LKB) model. In order to examine the individual predictive power of predictors describing parotid shape (radiomics), dose shape (dosiomics), and demographic characteristics, a total of 61 different features was defined and extracted from the DICOM files. These included the patient's age and sex, parotid shape features, features related to the dose-volume histogram, the mean dose to subvolumes of parotid glands, spatial dose gradients, and three-dimensional dose moments. In the multivariate analysis, a variety of machine learning algorithms was evaluated: 1) classification methods, that discriminated patients between a high and a low risk of complication, 2) feature selection techniques, that aimed to select a number of highly informative covariates from a large set of predictors, 3) sampling methods, that reduced the class imbalance, 4) data cleaning methods, that reduced noise in the data set. The predictive performance of the models was validated internally, using nested cross-validation, and externally, using an independent patient cohort from the PARSPORT clinical trial.

The LKB model showed fairly good performance on mild-to-severe (G1+) xerostomia predictions. The corresponding dose-response curve revealed that even small doses to parotid glands increase the risk of xerostomia and should be kept as low as possible. For the patients who did develop moderate-to-severe (G2+) xerostomia, the mean dose was not an informative predictor, even though the efficient sparing of parotid glands allowed to achieve low G2+ xerostomia rates. The features describing the shape of a parotid gland and

the shape of a dose proved to be highly predictive of xerostomia. In particular, the parotid volume and the spatial dose gradients in the transverse plane explained xerostomia well. The results of the machine learning algorithms comparison showed that a particular choice of a classifier and a feature selection method can significantly influence predictive performance of the NTCP model. In general, support vector machines and extra-trees achieved top performance, especially for the endpoints with a large number of observations. For the endpoints with a smaller number of observations, simple logistic regression often performed on a par with the top-ranking machine learning algorithms. The external validation showed that the analyzed multivariate models did not generalize well to the PARSPORT cohort. The only features that were predictive of xerostomia both in the Heidelberg (HD) and the PARSPORT cohort were the spatial dose gradients in the right-left and the anterior-posterior directions. Substantial differences in the distribution of covariates between the two cohorts were observed, which may be one of the reasons for the weak generalizability of the HD models.

The results presented in this thesis undermine the applicability of NTCP models of xerostomia based only on the mean dose to parotid glands in highly conformal radiotherapy treatments. The spatial dose gradients in the left-right and the anterior-posterior directions proved to be predictive of xerostomia both in the HD and the PARSPORT cohort. This finding is especially important as it is not limited to a single cohort but describes a general pattern present in two independent data sets. The performance of the sophisticated machine learning methods may indicate a need for larger patient cohorts in studies on NTCP models in order to fully benefit from their advantages. Last but not least, the observed covariate-shift between the HD and the PARSPORT cohort motivates, in the author's opinion, a need for reporting information about the covariate distribution when publishing novel NTCP models.

# ZUSAMMENFASSUNG

In der klinischen Routine wird dem Risiko einer Xerostomie typischerweise dadurch begegnet, dass die mittlere Strahlendosis auf die Parotis begrenzt wird. Dieser Ansatz lieferte gute Ergebnisse. In den letzten Jahren haben jedoch mehrere Studien berichtet, dass Modelle mit mittlerer Dosis als prädiktiven Parameter bei der Erkennung eines Xerostomie-Risikos versagen. Dies kann durch eine starke Verbesserung der Dosiskonformität in der Strahlentherapie aufgrund neuerer technologischer Fortschritte und einer dadurch wesentlichen Verringerung der mittleren Dosis für die Parotiden erklärt werden. Die vorliegende Arbeit untersucht neue Ansätze zum Aufbau zuverlässigerer Vorhersagemodelle für die Nebenwirkungswahrscheinlichkeit (engl.: normal tissue complication probability, NTCP) von Xerostomie.

Im Rahmen dieser Arbeit wurde eine Kohorte von 153 Kopf-Hals-Karzinompatienten, die mit Strahlentherapie am Universitätsklinikum Heidelberg behandelt wurden, wurde gesammelt. Die Vorhersagekraft der mittleren Dosis für Parotiden wurde mit dem Lyman-Kutcher-Burman (LKB) Modell bewertet. Um die individuelle Vorhersagekraft von Prädiktoren zu untersuchen, die jeweils Parotidenform (radiomics), Dosisform (dosiomics) und demographische Merkmale beschreiben, wurden insgesamt 61 verschiedene Indikatoren definiert und aus den DICOM-Dateien extrahiert. Dazu gehörten Alter und Geschlecht des Patienten, Charakteristiken der Parotiden, extrahierte Indikatoren aus Dosis-Volumen-Histogrammen, die mittlere Dosis für Subvolumina der Parotiden, räumliche Dosisgradienten sowie dreidimensionale Dosismomente. In der multivariaten Analyse wurden mehrere maschinelle Lernalgorithmen evaluiert: 1) Klassifikationsmethoden, die Patienten nach hohem und niedrigem Komplikationsrisiko einstufen, 2) Merkmalauswahltechniken, die eine Anzahl relevanter Kovariaten aus einer großen Menge von Prädiktoren auswählen, 3) Stichprobenverfahren, die das Klassenungleichgewicht reduzieren, 4) Datenbereinigungsmethoden, die das Rauschen im Datensatz reduzieren. Die prädiktive Leistung der Modelle wurde intern mithilfe einer verschachtelten Kreuzvalidierung (engl.: nested cross-validation) und extern mithilfe einer unabhängigen Patientenkohorte aus der PARSPORT-Studie evaluiert.

Das LKB-Modell zeigte eine annehmbare Leistung bei der Vorhersage von leicht bis schwerer Xerostomie. Die mit dem LKB-modell angepasste NTCP-Kurve zeigte, dass bereits eine geringe Dosierung der Parotiden das Risiko einer Xerostomie erhöhen kann

und die Dosis daher hier so gering wie möglich gehalten werden sollte. Für Patienten, die eine moderate bis schwere Xerostomie entwickelten, war die mittlere Dosis hingegen kein aussagekräftiger Prädiktor, obwohl durch die effiziente Schonung der Parotiden niedrige Xerostomie-Raten erreicht wurden. Die Merkmale, die die Form der Parotiden oder der Dosis beschreiben, erwiesen sich hier als prädiktiv für die Xerostomie. Insbesondere Parameter wie die Parotisvolumina und die räumlichen Dosisgradienten in der Transversalebene konnten Xerostomie gut vorhersagen. Die Ergebnisse des Vergleichs mit einem maschinellen Lernalgorithmus zeigten, dass die Wahl eines Klassifikators und eines Merkmalsauswahlverfahrens die Vorhersagefähigkeit des NTCP-Modells signifikant beeinflussen kann. Im Allgemeinen erzielten Support Vector Machine und Extra-Trees höchste Vorraussagekraft, insbesondere für die Endpunkte mit einer großen Anzahl von Beobachtungen. Für die Endpunkte mit geringerer Anzahl von Beobachtungen erzielt einfache logistische Regression oft gute Ergebnisse. Die externe Validierung zeigte, dass die analysierten multivariaten Modelle eher nicht auf die PARSPORT-Kohorte verallgemeinerten. Die einzigen Merkmale, die sowohl in der Heidelberger (HD) als auch in der PARSPORT-Kohorte auf Xerostomie hindeuteten, waren die räumlichen Dosisgradienten in der Rechts-Links-Richtung und der Anterior-Posterior-Richtung. In der Verteilung der Kovariaten zwischen den beiden Kohorten wurden erhebliche Unterschiede festgestellt, was einer der Gründe für die schwache Generalisierbarkeit der HD-Modelle sein könnte.

Die in dieser Arbeit vorgestellten Ergebnisse zweifeln an der Anwendbarkeit von NTCP-Modellen für Xerostomie, die nur auf der mittleren Dosis der Parotis bei hochkonformen Strahlentherapien beruhen. Die räumlichen Dosisgradienten in der links-rechts- und der anterior-posterior-Richtung hingegen erwiesen sich sowohl in der HD- als auch in der PARSPORT-Kohorte als prädiktiv für Xerostomie. Diese Erkenntnis ist besonders wichtig, da sie nicht auf eine einzige Kohorte beschränkt ist, sondern ein allgemeines Muster in zwei unabhängigen Datensätzen beschreibt. Die Leistung der ausgefeilten Methoden des maschinellen Lernens kann auf die Notwendigkeit größerer Patientenkohorten in Studien über NTCP-Modelle hinweisen, um ihre Vorteile voll auszunutzen. Die beobachtete Kovariatenverschiebung (engl.: covariate shift) zwischen der HD- und der PARSPORT-Kohorte motiviert die Notwendigkeit, Informationen über die Kovariatenverteilung in neuartigen NTCP-Modellen zu melden.

APPENDIX A

# DICOM IMPORT AND DATA PREPROCESSING

**Algorithm A.1:** Generate CT cube and directional vectors from DICOM CT images. Source code: load_ct_cube.m

```
   // Generate a CT cube and zVec from DICOM CT images
 1 i = 0;
 2 foreach dcm do
 3 |   read dcm;
 4 |   cube(:,:,i) = dcm.PixelData;
 5 |   zVec(i) = dcm.ImagePositionPatient(3);
 6 |   i = i+1;
 7 end
   // Generate yVec
 8 if dcm.ImageOrientationPatient(1:3) == (1,0,0) then
 9 |   yVec = dcm.ImagePositionPatient(2) + [0..dcm.Rows-1] * dcm.PixelSpacing(2);
10 else if dcm.ImageOrientationPatient(1:3) == (-1,0,0) then
11 |   yVec = dcm.ImagePositionPatient(2) - [0..dcm.Rows-1] * dcm.PixelSpacing(2);
12 |   yVec = sort(yVec);
13 |   cube = flip(cube,1);
14 else
15 |   error('Not supported patient's orientation');
16 end
   // Generate xVec
17 if dcm.ImageOrientationPatient(4:6) == (0,1,0) then
18 |   xVec = dcm.ImagePositionPatient(1) + [0..dcm.Columns-1] * dcm.PixelSpacing(1);
19 else if dcm.ImageOrientationPatient(4:6) == (0,-1,0) then
20 |   xVec = dcm.ImagePositionPatient(1) - [0..dcm.Columns-1] * dcm.PixelSpacing(1);
21 |   xVec = sort(xVec);
22 |   cube = flip(cube,2);
23 else
24 |   error('Not supported patient's orientation');
25 end
   // Ensure that the zVec is monotonic
26 [zVec, idx] = sort(zVec);
27 cube = cube(:,:,idx);
```

**Algorithm A.2:** Generate a dose cube and directional vectors from DICOM RT Dose. Source code: load_dose_cube.m

```
   // Generate dose cube and zVec from DICOM RT dose
 1 read dcm;
 2 cube = dcm.PixelData;
 3 cube = cube * dcm.DoseGridScaling;
 4 if dcm.GrifFrameOffsetVector(1) != 0 then
 5 |   offset = dcm.GridFrameOffsetVector - dcm.GridFrameOffsetVector(1);
 6 else
 7 |   offset = dcm.GridFrameOffsetVector;
 8 end
 9 zVec = dcm.ImagePositionPatient(3) + offset;
10 if zVec(2) - zVec(1) < 0 then
11 |   zVec(:,1) = flip(zVec);
12 |   cube = flip(cube, 3);
13 end
   // Generate yVec
14 if dcm.ImageOrientationPatient(1:3) == (1,0,0) then
15 |   yVec = dcm.ImagePositionPatient(2) + [0..dcm.Rows-1] * dcm.PixelSpacing(2);
16 else if dcm.ImageOrientationPatient(1:3) == (-1,0,0) then
17 |   yVec = dcm.ImagePositionPatient(2) - [0..dcm.Rows-1] * dcm.PixelSpacing(2);
18 |   yVec = sort(yVec);
19 |   cube = flip(cube,1);
20 else
21 |   error('Not supported patient's orientation');
22 end
   // Generate xVec
23 if dcm.ImageOrientationPatient(4:6) == (0,1,0) then
24 |   xVec = dcm.ImagePositionPatient(1) + [0..dcm.Columns-1] * dcm.PixelSpacing(1);
25 else if dcm.ImageOrientationPatient(4:6) == (0,-1,0) then
26 |   xVec = dcm.ImagePositionPatient(1) - [0..dcm.Columns-1] * dcm.PixelSpacing(1);
27 |   xVec = sort(xVec);
28 |   cube = flip(cube,2);
29 else
30 |   error('Not supported patient's orientation');
31 end
```

**Algorithm A.3:** Calculate masks for structures from DICOM RT Structure Set. Source code: calc_struct_masks.m

```
1  xVec, yVec, zVec ← DICOM RT Dose and/or DICOM CT;
2  dcm ← read DICOM RT Structure Set;
3  contoured_structs ← dcm.ROIContourSequence;
4  defined_structs ← dcm.StructureSetROISequence;
5  foreach c_struct in contoured_structs do
       // Get structure name
6      foreach d_struct in defined_structs do
7          if c_struct.ReferencedROINumber == d_struct.ROINumber then
8              struct_name ← d_struct.ROIName;
9              break;
10         end
11     end
       // Generate logical mask for each structure slice
12     slices ← struct.ContourSequence;
13     zCoords ← empty list;
14     slice_masks ← empty list;
15     foreach slice in slices do
16         if slice.ContourGeometricType != 'POINT' then
               // Structure vertices in the slice
17             zCoord ← slice.ContourData(3);
18             add zCoord to zCoords;
19             xCoord, yCoord ← slice.ContourData(1:2);
               // Generate meshgrid
20             X, Y ← xVec, yVec;
               // Calculate the logical mask
21             slice_mask ← inpolygon(X, Y, xCoord, yCoord);
22             add slice_mask to slice_masks;
23         end
24     end
       // Move to the next structure if this one is of type POINT
25     if slice.ContourGeometricType == 'POINT' then
26         continue;
27     end
       // Convert 2D slice masks to a 3D structure mask
28     struct_mask ← zCoords, slice_masks;
       // Interpolate structure mask to DICOM RT Dose resolution
29     struct_mask ← struct_mask, xVec, yVec, zVec;
30 end
```

APPENDIX B

# DATA SAMPLING, FEATURE SELECTION, AND CLASSIFICATION ALGORITHMS

*Parts of this appendix has been quoted verbatim from Gabryś et al. (2018). Under the journal's copyright policy, the authors retained the copyright to their work.*

## B.1 Data cleaning and class balancing

### Random oversampling

Class imbalance is reduced by randomly duplicating observations from the minority class.

### Synthetic minority oversampling

Synthetic minority oversampling (SMOTE) was proposed by Chawla et al. (2002). The algorithm generates new synthetic minority observations by considering $k$ nearest neighbors of a randomly selected minority observation. Next, the difference between the observation feature vector and one of the nearest neighbors feature vector is taken. This difference is then multiplied by a random weight between zero and one and added to the observation feature vector to generate new synthetic observation. Approximately equal number of synthetic observations is created for each minority class observation.

### Adaptive synthetic sampling

Adaptive synthetic sampling (ADASYN) (He et al. 2008), similarly to SMOTE, generates synthetic minority class observations by interpolating feature vectors between a minority class observation and a randomly selected nearest neighbor. The key difference to SMOTE is that ADASYN aims to create more synthetic data for minority class observations that are hard to learn. For that reason, a learning difficulty weight is calculated for each minority class observation, based on the number of majority class observations in its neighborhood.

Based on these weights, more synthetic observations are created for *difficult* minority class observations.

### Tomek links

A pair of observations $(E_i, E_j)$ stemming from different classes and with distance $d(E_i, E_j)$ form a Tomek link if there is no observation $E_l$, such that $d(E_i, E_l) < d(E_i, E_j)$ or $d(E_j, E_l) < d(E_i, E_j)$ (Tomek 1976). As an undersampling method, all the observations in the majority class forming Tomek links are removed; when used as a data cleaning method, both the observation from the majority and the observation from the minority class are eliminated.

### Condensed Nearest Neighbor Rule

The condensed nearest neighbor rule (CNN) proposed by Hart (1968) undersamples the data set to find a consistent subset $\hat{E}$ of all observations $E$. First, all minority class observations and one randomly selected majority class observation are moved to $\hat{E}$. Next, the rest of the majority class observations are classified using 1-nearest neighbor rule and during this process every misclassified observation is moved to subset $\hat{E}$. The procedure continues until all misclassified observations are in the subset $\hat{E}$ (Kubat & Matwin 1997). Intuitively, CNN reduces the number of redundant observations in majority class that are far from the decision border and therefore less informative in learning.

### One-sided selection

One-sided selection (OSS) (Kubat & Matwin 1997) is an undersampling method realized by Tomek links algorithm followed by CNN. Tomek links undersample the majority class and remove noisy and borderline class observations. CNN, on the other hand, removes observations from the majority class that are distant from the decision border and likely are not informative.

### Wilson's edited nearest neighbor rule

The Wilson's edited nearest neighbor rule (ENN) (Wilson 1972) removes all observations which class label differ from the class of its $k$ nearest neighbors.

### Neighborhood cleaning rule

The neighborhood cleaning rule (NCL) (Laurikkala 2001) is a modification of the ENN algorithm. As in the ENN, the class of each observation is compared with the classes of its $k$ nearest neighbors. If the analyzed observation belongs to the majority class, the procedure is the same as in the ENN. However, if the observation belongs to the minority class and its $k$ nearest neighbors to the majority class, the minority class observation is kept in the data set and the $k$ nearest neighbors are removed.

#### SMOTE + TL

First, the original data set is oversampled with SMOTE, and then Tomek links are identified and removed. The method aims to produce a balanced data set with well-defined class clusters (Batista et al. 2004).

#### SMOTE + ENN

This method is similar to SMOTE+TL but with stronger data cleaning component realized by the ENN (Batista et al. 2004).

## B.2 Feature selection

### Univariate feature selection

Univariate feature selection methods evaluate each feature separately relying solely on the relation between one feature characteristic and the modeled variable. After all the features were graded, the features with the highest rankings are selected. A disadvantage of univariate feature selection is that the algorithm fails to select features which have relatively low individual scores but a high score when combined together. Also, due to the fact that univariate feature selection methods evaluate features individually, they are unable to handle feature redundancy (Gu et al. 2012, Tang et al. 2014).

#### Fisher score

Intuitively, Fisher score is a ratio of the between-class scatter to the within-class scatter. As a result, high Fisher scores correspond to features with well defined class clusters (low within-class scatter) that are distant from each other (large between-class scatter) (Duda et al. 2012). Fisher score is commonly used in supervised classification tasks due to its low computational cost and general good performance (Gu et al. 2012).

Fisher score of feature $X$ was calculated using the following formula (Lowry 2014):

$$F(X) = \frac{\frac{1}{C-1}\sum_{c=1}^{C} N_c(\bar{x}_c - \bar{x})^2}{\frac{1}{N-C}\sum_{c=1}^{C}\sum_{i:y_i=c}(x_i - \bar{x}_c)^2}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{x}_c = \frac{1}{N_c}\sum_{i:y_i=c} x_i$$

where $C$ is the number of classes, $N$ total number of observations, $N_c$ number of observations in class $c$, $\bar{x}$ mean value of feature $X$, and $\bar{x}_c$ mean value of feature $X$ in class $c$.

#### Mutual information

This univariate feature selection method measures mutual information between each feature and the modeled variable. Intuitively, mutual information measures how much knowing the feature $X$ value reduces uncertainty about the class label $Y$, and vice versa (Murphy 2012). This can be expressed by the formula:

$$MI(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

where $H(X)$ is the entropy of $X$ and $H(X|Y)$ is the entropy of $X$ after observing class $Y$.

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i)$$

$$H(X|Y) = -\sum_{i=1}^{N} p(y_i) \sum_{k=1}^{N} p(x_k|y_i) \log p(x_k|y_i)$$

Features with high mutual information are considered informative.

#### Recursive feature elimination

In the first step of recursive feature elimination (RFE), an induction algorithm is trained using the full set of features. Next, the features are ranked according to a given criterion, such as feature weight in logistic regression or feature importance in ensemble models. Then, the feature or the features with the smallest ranks are removed from the feature set. This procedure is repeated iteratively until the desired number of features is achieved (Kohavi & John 1997, Guyon et al. 2002).

In contrast to univariate feature selection, recursive feature elimination methods can capture feature interactions. For this reason, they can select not only good univariate predictors but also features which have low predictive power alone but high predictive power when pooled together.

The ability to handle feature redundancy depends on the induction algorithm used with RFE. For instance, L1 penalized logistic regression tends to select one of highly correlated features, hence reducing feature redundancy (Hastie et al. 2009). On the contrary, L2 penalized logistic regression tends to give similar weights to correlated features distributing the total feature importance among them. For the recursive feature elimination in this study two induction algorithms were used: logistic regression and extra-trees.

#### Model based feature selection

Model based feature selection can be considered a special case of recursive feature elimination with only one iteration step. The induction algorithm is trained using the full set of features and the desired number of lowest scoring features is removed. Similarly to RFE, logistic regression and extra-trees were employed as the induction algorithms.

## B.3 Classification

### Logistic Regression

Logistic regression is a simple linear model allowing to estimate probability of a binary response based on a number of risk factors. In order to avoid overfitting, logistic regression is usually regularized via either L1, L2, or elastic net penalty. L1 penalty outperforms L2 penalty in terms of handling irrelevant and redundant features (Ng 2004). Its ability to bring feature weights to zero results in sparse models and improves model interpretability (Bishop 2006). On the other hand, L1 tends to randomly select one of highly correlated features which can result in model variability (Zou & Hastie 2005). The elastic method brings in a way the two worlds together and applies a penalty that is a convex combination of L1 and L2 regularization (Zou & Hastie 2005).

The advantages of logistic regression are its simplicity, interpretability, and easy tuning (only one hyperparameter with L1 or L2 regularization or two hyperparameters with elastic net regularization). The biggest disadvantage is a linear hypersurface decision boundary that may not be flexible enough to describe the real decision boundary.

### k-Nearest Neighbors

The k-nearest neighbor (kNN) classifier looks at the $k$ points in the training set that are nearest to the test input. The object is classified based on a majority of vote of its neighbors (Murphy 2012). kNN has a much more flexible decision boundary compared to logistic regression. It will likely outperform logistic regression when the true decision boundary is highly irregular. Nevertheless, the curse of dimensionality has a considerable impact on the performance of the k-nearest neighbors classifier making feature selection crucial when working with high-dimensional data sets.

### Support Vector Machines

Similarly to the k-nearest neighbors algorithm, the support vector machine (SVM) does not learn a fixed set of parameters corresponding to the features of the input. It rather remembers the training examples and classifies new observations based on some similarity function. The two main concepts behind support vector machines are the kernel trick and the large margin principle. The kernel trick guarantees high flexibility of the decision boundary by allowing to operate in features spaces of very high, even infinite, dimensionality. The large margin principle, ensures model sparsity by discarding all observations not laying on maximum margin hypersurfaces. Support vector machines proved to be very successful in various classification tasks, including NTCP modeling. Unfortunately, interpretation of support vector machines with nonlinear kernels is a challenge (Burges 1998).

### Extra-Trees

The extra-trees classifier is an ensemble of decision trees. Each tree is built either on the full learning sample or on a bootstrap replica. At each node, a random subset of features is selected and for each feature a random cut-point is drawn. The best feature-cutpoint pair is selected to split the node. The tree is grown until the minimum sample size for splitting a node is reached. The ensemble predictions are the results of the majority vote of predictions of individual trees (Geurts et al. 2006). A great advantage of the extra-trees algorithm is that it works *out-of-the-box* with no or minimal hyperparameter tuning.

### Gradient Tree Boosting

Similarly to extra-trees, gradient tree boosting uses an ensemble of decision trees. Gradient tree boosting iteratively fits small decision trees to the data set in an adaptive fashion. After each iteration, training samples are reweighted to focus on the instances misclassified by the previous trees. When all trees are grown, the prediction is obtained by the weighted majority vote of the trees (Hastie et al. 2009, Freund & Schapire 1997).

Gradient tree boosting proved to be a very successful algorithm often outperforming neural networks, support vector machines, and other ensemble models. However, tuning the hyperparameters may be challenging.

**Table B.1:** Hyperparameters used to tune the sampling algorithms. Hyperparameters not listed in this table assumed the default values of imbalanced-learn package (Lemaitre et al. 2017).

| Algorithm | Hyperparameters | Values |
|---|---|---|
| ROS | - | - |
| SMOTE | **k_neighbors**: Number of nearest neighbors used to construct synthetic samples. | $\{3, 4, 5\}$ |
| | **m_neighbors**: Number of nearest neighbors used to determine if a minority sample is in danger. | $\{7, 8, 9\}$ |
| | **kind**: Type of SMOTE algorithm. | {'regular', 'borderline1', 'borderline2'} |
| ADASYN | **n_neighbors**: Number of nearest neighbors to use to construct synthetic samples. | $\{3, 5, 8\}$ |
| OSS | - | - |
| TL | - | - |
| ENN | **n_neighbors**: Number of nearest neighbors. | $\{2, 3, 5\}$ |
| | **kind_sel**: Type of ENN algorithm. | {'all', 'mode'} |
| NCL | **n_neighbors**: Number of nearest neighbors. | $\{2, 3, 5\}$ |
| SMOTE+TL | - | - |
| SMOTE+ENN | - | - |

**Table B.2:** Hyperparameters used to tune the feature selection algorithms. Hyperparameters not listed in this table assumed the default values of scikit-learn package (Pedregosa et al. 2011).

| **Algorithm** | **Hyperparameters** | **Values** |
|---|---|---|
| UFS-F | **k**: Number of features to select. | $\{2, 3, 4, 5, 6\}$ |
| UFS-MI | **k**: Number of features to select. | $\{2, 3, 4, 5, 6\}$ |
| RFE-LR | **k**: Number of features to select.<br>**step**: Number of features to remove at each iteration.<br>**class_weight**: Whether class weights are equal or inversely proportional to class frequencies.<br>**C**: Inverse of regularization strength.<br>**penalty**: Type of regularization. | $\{2, 3, 4, 5, 6\}$<br>1<br>{None, 'balanced'}<br>$\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$<br>'l2' |
| RFE-ET | **k**: Number of features to select.<br>**step**: Fraction of features to remove at each iteration.<br>**class_weight**: Whether class weights are equal or inversely proportional to class frequencies.<br>**n_estimators**: Number of decision trees. | $\{2, 3, 4, 5, 6\}$<br>0.5<br>{None, 'balanced', 'balanced_subsample'}<br>$[90, 140]$ |
| MB-LR | **k**: Number of features to select.<br>**class_weight**: Whether class weights are equal or inversely proportional to class frequencies.<br>**C**: Inverse of regularization strength.<br>**penalty**: Type of regularization. | $\{2, 3, 4, 5, 6\}$<br>{None, 'balanced'}<br>$\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$<br>{'l1', 'l2'} |
| MB-ET | **k**: Number of features to select.<br>**class_weight**: Whether class weights are equal or inversely proportional to class frequencies.<br>**n_estimators**: Number of decision trees. | $\{2, 3, 4, 5, 6\}$<br>{None, 'balanced', 'balanced_subsample'}<br>$[90, 140]$ |

**Table B.3:** Hyperparameters used to tune the classification algorithms. Hyperparameters not listed in this table assumed the default values of scikit-learn (Pedregosa et al. 2011) and xgboost (Chen & Guestrin 2016) packages.

| Algorithm | Hyperparameters | Values |
|---|---|---|
| LR-L1/L2 | **class_weight**: Whether class weights are equal or inversely proportional to class frequencies. | {None, 'balanced'} |
| | **C**: Inverse of regularization strength. | $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$ |
| LR-EN | **class_weight**: Whether class weights are equal or inversely proportional to class frequencies. | {None, 'balanced'} |
| | **alpha**: Regularization strength. | $\{2^{-10}, 2^{-9.985}, 2^{-9.97}, \ldots, 2^{5}\}$ |
| | **l1_ratio**: Ratio between L1 and L2 penalty. | $[0, 1]$ |
| kNN | **n_neighbors**: Number of nearest neighbors. | $\{1, 2, 3, \ldots, 9\}$ |
| | **p**: Power parameter of the Minkowski distance. | $\{1, 2, \infty\}$ |
| SVM | **class_weight**: Whether class weights are equal or inversely proportional to class frequencies. | {None, 'balanced'} |
| | **C**: Inverse of regularization strength. | $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$ |
| | **gamma**: Parameter of the RBF kernel. | $\{2^{-15}, 2^{-14.982}, 2^{-14.964}, \ldots, 2^{3}\}$ |
| ET | **n_estimators**: Number of decision trees. | $[90, 230]$ |
| | **class_weight**: Whether class weights are equal of inversely proportional to class frequencies. | {None, 'balanced'} |
| | **criterion**: The function to measure the quality of a split. | {'gini', 'entropy'} |
| | **max_features**: Number of features to consider when calculating the best split. | $\{0.05, 0.10, 0.15, \ldots, 1\}$ |
| | **min_samples_split**: The minimum number of samples required to split a node. | $\{2, 3, 4, \ldots, 20\}$ |
| | **min_samples_leaf**: The minimum number of samples required to be at a leaf node. | $\{1, 2, 3, \ldots, 20\}$ |
| GTB | **n_estimators**: Number of decision trees. | $[200, 2000]$ |
| | **learning_rate**: Boosting learning rate. | $\{2^{-7}, 2^{-6.994}, 2^{-6.988}, \ldots, 2^{-1}\}$ |
| | **max_depth**: Maximum tree depth. | $\{1, 2, 3, \ldots, 6\}$ |
| | **gamma**: Minimum loss reduction required to make a further partition on a leaf node of the tree. | $\{0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$ |
| | **min_child_weight**: Minimum sum of instance weight(hessian) needed in a child. | $\{1, 3, 5, 7\}$ |
| | **subsample**: Ratio of the training samples used to grow trees. | $\{0.6, 0.65, 0.70, \ldots, 1\}$ |
| | **reg_lambda**: L1 regularization term on weights. | $[0, 1]$ |
| | **reg_alpha**: L2 regularization term on weights. | $[0, 1]$ |

# BIBLIOGRAPHY


Abdi, H. (2007), The Kendall rank correlation coefficient, *in* Salkind, Neil J., ed., 'Encyclopedia of Measurement and Statistics', SAGE, Thousand Oaks, CA, USA, pp. 509–510.

Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J. & Lambin, P. (2014), 'Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.', *Nat Commun* **5**, 4006.

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B. & Salakoski, T. (2011), 'An experimental comparison of cross-validation techniques for estimating the area under the ROC curve', *Comput Stat Data Anal* **55**(4), 1828–1844.

Bamber, D. (1975), 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', *J Math Psychol* **12**(4), 387–415.

Barnard, G. A. (1945), 'A new test for 2x2 tables', *Nature* **156**, 177.

Barnard, G. A. (1947), 'Significance tests for 2x2 tables', *Biometrika* **34**, 123–138.

Batista, G. E. A. P. A., Prati, R. C. & Monard, M. C. (2004), 'A study of the behavior of several methods for balancing machine learning training data', *SIGKDD Explor* **6**(1), 20–29.

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. M. C., Van Der Laan, B. F. a. M., Vissink, A., Steenbakkers, R. J. H. M. & 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', *Radiother Oncol* **105**(1), 86–93.

Beetz, I., Steenbakkers, R. J. H. M., Chouvalova, O., Leemans, C. R., Doornaert, P., van der Laan, B. F. a. M., Christianen, M. E. M. C., Vissink, A., Bijl, H. P., van Luijk, P. & Langendijk, J. A. (2014), 'The QUANTEC criteria for parotid gland dose and their efficacy to prevent moderate to severe patient-rated xerostomia.', *Acta Oncol* **53**(5), 597–604.

Bellman, R. E. (1961), *Adaptive Control Processes: A Guided Tour*, Princeton University Press.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *J R Stat Soc Series B Stat Methodol* **57**(1), 289–300.

Benjamini, Y., Krieger, A. M. & Yekutieli, D. (2006), 'Adaptive linear step-up procedures that control the false discovery rate', *Biometrika* **93**(3), 491–507.

Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Ten Haken, R. K. & Yorke, E. D. (2010), 'Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): An Introduction to the Scientific Issues', *Int J Radiat Oncol Biol Phys* **76**(3), 3–9.

Bentzen, S. M., Dörr, W., Anscher, M. S., Denham, J. W., Hauer-Jensen, M., Marks, L. B. & Williams, J. (2003), 'Normal tissue effects: reporting and analysis', *Semin Radiat Oncol* **13**(3), 189–202.

Bentzen, S. M. & Tucker, S. L. (1997), 'Quantifying the position and steepness of radiation dose-response curves', *Int J Radiat Biol* **71**(5), 531–542.

Bergstra, J. & Bengio, Y. (2012), 'Random search for hyper-parameter optimization', *J Mach Learn Res* **13**, 281–305.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, Springer, New York, NY, USA.

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R., Derksen-Lubsen, G., Grobbee, D. E. & Moons, K. G. (2003), 'External validation is necessary in prediction research: a clinical example', *J Clin Epidemiol* **56**(9), 826–832.

Braam, P. M., Roesink, J. M., Moerland, M. a., Raaijmakers, C. P., Schipper, M. & Terhaard, C. H. (2005), 'Long-term parotid gland function after radiotherapy', *Int J Radiat Oncol Biol Phys* **62**(3), 659–664.

Braam, P. M., Terhaard, C. H., Roesink, J. M. & Raaijmakers, C. P. (2006), 'Intensity-modulated radiotherapy significantly reduces xerostomia compared with conventional radiotherapy', *Int J Radiat Oncol Biol Phys* **66**(4), 975–980.

Brahme, A. (1984), 'Dosimetric precision requirements in radiation therapy', *Acta Oncol* **23**(5), 379–391.

Buettner, F., Miah, A. B., Gulliford, S. L., Hall, E., Harrington, K. J., Webb, S., Partridge, M. & 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', *Radiother Oncol* **103**(1), 82–87.

Burges, C. J. C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Min Knowl Discov* **2**(2), 121–167.

Cameron, A., Khalvati, F., Haider, M. & Wong, A. (2015), 'MAPS: a quantitative radiomics approach for prostate cancer detection', *IEEE Trans Biomed Eng* **63**(6), 1145–1156.

Cawley, G. C. & Talbot, N. L. C. (2010), 'On over-fitting in model selection and subsequent selection bias in performance evaluation', *J Mach Learn Res* **11**, 2079–2107.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'SMOTE: synthetic minority over-sampling technique', *J Artif Intell Res* **16**, 321–357.

Chen, S., Zhou, S., Yin, F.-F., Marks, L. B. & Das, S. K. (2007), 'Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis', *Med Phys* **34**(10), 3808–3814.

Chen, T. & Guestrin, C. (2016), XGBoost: a scalable tree boosting system, *in* 'KDD', ACM, New York, NY, USA, pp. 785–794.

Coates, J., Souhami, L. & El Naqa, I. (2016), 'Big Data Analytics for Prostate Radiotherapy', *Front Oncol* **6**(June), 149.

Cox, J. D. & Stetz, J. (1995), 'Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC)', *Int J Radiat Oncol Biol Phys* **31**(5), 1341–1346.

Davis, J. & Goadrich, M. (2006), 'The relationship between precision-recall and ROC curves', *Proc Int Conf Mach Learn* pp. 233–240.

Dean, J. A., Wong, K. H., Welsh, L. C., Jones, A. B., Schick, U., Newbold, K. L., Bhide, S. A., Harrington, K. J., Nutting, C. M. & Gulliford, S. L. (2016), 'Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy', *Radiother Oncol* **120**(1), 21–27.

Deasy, J. O., Moiseenko, V., Marks, L., Chao, K. S. C., Nam, J. & Eisbruch, A. (2010), 'Radiotherapy dose-volume effects on salivary gland function', *Int J Radiat Oncol Biol Phys* **76**(3 SUPPL.), 58–63.

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. (1988), 'Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach', *Biometrics* **44**(3), 837–845.

Demšar, J., Curk, T., Erjavec, A., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. & Zupan, B. (2013), 'Orange: data mining toolbox in Python', *J Mach Learn Res* **14**, 2349–2353.

Dijkema, T., Raaijmakers, C. P. J., Braam, P. M., Roesink, J. M., Monninkhof, E. M. & Terhaard, C. H. J. (2012), 'Xerostomia: a day and night difference', *Radiother Oncol* **104**(2), 219–223.

Dijkema, T., Terhaard, C. H. J., Roesink, J. M., Braam, P. M., van Gils, C. H., Moerland, M. a. & Raaijmakers, C. P. J. (2008), 'Large cohort dose-volume response analysis of parotid gland function after radiotherapy: intensity-modulated versus conventional radiotherapy', *Int J Radiat Oncol Biol Phys* **72**(4), 1101–1109.

Duda, R. O., Hart, P. E. & Stork, D. G. (2012), *Pattern Classification*, 2 edn, John Wiley and Sons, New York, NY, USA.

Efron, B. (1987), 'Better bootstrap confidence intervals', *J Am Stat Assoc* **82**(397), 171–185.

Efron, B. & Tibshirani, R. J. (1993), *An introduction to the bootstrap*, CRC Press.

El Naqa, I., Brock, K., Yu, Y., Langen, K. & Klein, E. E. (2018), 'On the fuzziness of machine learning, neural networks, and artificial intelligence in radiation oncology', *Int J Radiat Oncol Biol Phys* **100**(1), 1–4.

Fisher, R. A. (1925), *Statistical methods for research workers*, Genesis Publishing Pvt Ltd.

Freund, Y. & Schapire, R. E. (1997), 'A decision-theoretic generalization of on-Line learning and an application to boosting', *J Comput Syst Sci* **55**(1), 119–139.

Friedman, M. (1937), 'The use of ranks to avoid the assumption of normality implicit in the analysis of variance', *J Am Stat Assoc* **32**(200), 675–701.

Friedman, M. (1940), 'A comparison of alternative tests of significance for the problem of m rankings', *Ann Math Stat* **11**(1), 86–92.

Gabryś, H. S., Buettner, F., Sterzing, F., Hauswald, H. & Bangert, M. (2017), 'Parotid gland mean dose as a xerostomia predictor in low-dose domains', *Acta Oncol* **56**(9), 1197–1203.

Gabryś, H. S., Buettner, F., Sterzing, F., Hauswald, H. & Bangert, M. (2018), 'Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia', *Front Oncol* **8**, 35.

Gavrilov, Y., Benjamini, Y. & Sarkar, S. K. (2009), 'An adaptive step-down procedure with proven FDR control under independence', *Ann Stat* **37**(2), 619–629.

Geurts, P., Ernst, D. & Wehenkel, L. (2006), 'Extremely randomized trees', *Mach Learn* **63**(1), 3–42.

Gillies, R. J., Kinahan, P. E. & Hricak, H. (2015), 'Radiomics: images are more than pictures, they are data', *Radiology* **278**(2), 563–577.

Goeman, J. J. & Solari, A. (2014), 'Multiple hypothesis testing in genomics', *Stat Med* **33**(11), 1946–1978.

Goitein, M. (1979), 'The utility of computed tomography in radiation therapy: an estimate of outcome', *Int J Radiat Oncol Biol Phys* **5**(10), 1799–1807.

Gonzalez, R. C. & Woods, R. E. (2006), *Digital Image Processing*, 3rd edn, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Gray, H. (1918), *Anatomy of the Human Body*, Lea & Febiger, Philadelphia, PA, USA.

Gu, Q., Li, Z. & Han, J. (2012), 'Generalized fisher score for feature selection', *arXiv preprint arXiv:1202.3725* .

Gulliford, S. (2015), Modelling of Normal Tissue Complication Probabilities (NTCP): review of application of machine learning in predicting NTCP, *in* 'Machine Learning in Radiation Oncology', Springer, pp. 277–310.

Guyon, I. & Elisseeff, A. (2003), 'An introduction to variable and feature selection', *J Mach Learn Res* **3**, 1157–1182.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Mach Learn* **46**(1-3), 389–422.

Hanley, J. A. & Hajian-Tilaki, K. O. (1997), 'Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update', *Acad Radiol* **4**(1), 49–58.

Hart, P. E. (1968), 'The condensed nearest neighbour rule', *IEEE Trans Inf Theory* **14**(5), 515–516.

Hastie, T., Tibshirani, R. J. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2 edn, Springer, New York, NY, USA.

Hawkins, P. G., Lee, J. Y., Mao, Y., Li, P., Green, M., Worden, F. P., Swiecicki, P. L., Mierzwa, M. L., Spector, M. E., Schipper, M. J. & Eisbruch, A. (2018), 'Sparing all salivary glands with IMRT for head and neck cancer: longitudinal study of patient-reported xerostomia and head-and-neck quality of life', *Radiother Oncol* **126**(1), 68–74.

He, H., Bai, Y., Garcia, E. a. & Li, S. (2008), ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *in* 'Proc Int Jt Conf Neural Netw', IEEE, pp. 1322–1328.

He, H. & Garcia, E. a. (2009), 'Learning from imbalanced data', *IEEE Trans Knowl Data Eng* **21**(9), 1263–1284.

Ho, F. K., Farnell, D. J. J., Routledge, J. A., Burns, M. P., Sykes, A. J., Slevin, N. J. & Davidson, S. E. (2010), 'Comparison of patient-reported late treatment toxicity (LENT-SOMA) with quality of life (EORTC QLQ-C30 and QLQ-H&N35) assessment after head and neck radiotherapy', *Radiother Oncol* **97**(2), 270–275.

Hochberg, Y. (1988), 'A sharper Bonferroni test for multiple tests of significance', *Biometrika* **75**(4), 800–802.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scand Stat Theory Appl* **6**, 65–70.

Holthusen, H. (1936), 'Erfahrungen über die Verträglichkeitsgrenze für Röntgenstrahlen und deren Nutzanwendung zur Verhütung von Schäden', *Strahlentherapie* **57**, 254–269.

Houweling, A. C., Philippens, M. E. P., Dijkema, T., Roesink, J. M., Terhaard, C. H. J., Schilstra, C., Ten Haken, R. K., Eisbruch, A. & Raaijmakers, C. P. J. (2010), 'A comparison of dose-response models for the parotid gland in a large group of head-and-neck cancer patients', *Int J Radiat Oncol Biol Phys* **76**(4), 1259–1265.

Hunter, J. D. (2007), 'Matplotlib: a 2D graphics environment', *Comput Sci Eng* **9**(3), 99–104.

Japkowicz, N. & Stephen, S. (2002), 'The class imbalance problem: a systematic study', *Intell Data Anal* **6**(5), 429–449.

Källman, P., Ågren, A. & Brahme, A. (1992), 'Tumour and normal tissue responses to fractionated non-uniform dose delivery', *Int J Radiat Biol* **62**(2), 249–262.

Kałużny, J., Wierzbicka, M., Nogala, H., Milecki, P. & Kopeć, T. (2014), 'Radiotherapy induced xerostomia: Mechanisms, diagnostics, prevention and treatment–evidence based up to 2013', *Otolaryngologia polska* **68**(1), 1–14.

Kang, J., Schwartz, R., Flickinger, J. & Beriwal, S. (2015), 'Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective', *Int J Radiat Oncol Biol Phys* **93**(5), 1127–1135.

Khan, F. M. & Gibbons, J. P. (2014), *Khan's The Physics of Radiation Therapy*, 5th edn, Lippincott Williams & Wilkins, Philadelphia, PA, USA.

Kohavi, R. & John, G. (1997), 'Wrappers for feature subset selection', *Artif Intell* **97**(97), 273–324.

Krzanowski, W. & Hand, D. (1997), 'Assessing error rate estimators: the leave-one-out method reconsidered', *Aust N Z J Stat* **39**(1), 35–46.

Kubat, M. & Matwin, S. (1997), Addressing the curse of imbalanced training sets: one-sided selection, *in* 'Proc Int Conf Mach Learn', pp. 179–186.

Kutcher, G. J. & Burman, C. (1989), 'Calculation of complication probability factors for non-uniform normal tissue irradiation: the effective volume method.', *Int J Radiat Oncol Biol Phys* **16**(6), 1623–1630.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G. P. M., Granton, P., Zegers, C. M. L., Gillies, R., Boellard, R., Dekker, A. & Aerts, H. J. W. L. (2012), 'Radiomics: extracting more information from medical images using advanced feature analysis.', *Eur J Cancer* **48**(4), 441–446.

Lasko, T. a., Bhagwat, J. G., Zou, K. H. & Ohno-Machado, L. (2005), 'The use of receiver operating characteristic curves in biomedical informatics', *J Biomed Inform* **38**(5), 404–415.

Laurikkala, J. (2001), Improving identification of difficult small classes by balancing class distribution, *in* 'Artif Intell Med', pp. 63–66.

Lee, T.-F., Chao, P. J., Ting, H. M., Chang, L., Huang, Y. J., Wu, J. M., Wang, H. Y., Horng, M. F., Chang, C. M., Lan, J. H., Huang, Y. Y., Fang, F. M. & Leung, S. W. (2014), 'Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer', *PLoS One* **9**(2), e89700.

Lee, T.-F. & Fang, F. M. (2013), 'Quantitative analysis of normal tissue effects in the clinic (QUANTEC) guideline validation using quality of life questionnaire datasets for parotid gland constraints to avoid causing xerostomia during head-and-neck radiotherapy', *Radiother Oncol* **106**(3), 352–358.

Lee, T.-F., Liou, M.-H., Ting, H.-M., Chang, L., Lee, H.-Y., Leung, S. W., Huang, C.-J. & Chao, P.-J. (2015), 'Patient- and therapy-related factors associated with the incidence of xerostomia in nasopharyngeal carcinoma patients receiving parotid-sparing helical tomotherapy', *Sci Rep* **5**, 13165.

Lemaitre, G., Nogueira, F. & Aridas, C. K. (2017), 'Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning', *J Mach Learn Res* **18**(17), 1–5.

Leung, S. W. & Lee, T.-F. (2013), 'Treatment of nasopharyngeal carcinoma by tomotherapy: five-year experience', *Radiat Oncol* **8**(1), 107.

Livsey, J. E., Routledge, J., Burns, M., Swindell, R., Davidson, S. E., Cowan, R. A., Logue, J. P. & Wylie, J. P. (2002), 'Scoring of treatment-related late effects in prostate cancer', *Radiother Oncol* **65**(2), 109–121.

Lowry, R. (2014), One-way analysis of variance for independent samples, *in* 'Concepts and Applications of Inferential Statistics', DOER - Directory of Open Educational Resources, Poughkeepsie, NY, USA.

Luijk, P. V., Pringle, S., Deasy, J. O., Moiseenko, V. V., Faber, H., Hovan, A., Baanstra, M., Laan, H. P. V. D., Kierkels, R. G. J., Schaaf, A. V. D., Witjes, M. J., Schippers, J. M., Brandenburg, S., Langendijk, J. A., Wu, J. & Coppes, R. P. (2015), 'Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer', *Sci Transl Med* **7**(305), 1–8.

Lyman, J. T. (1985), 'Complication probability as assessed from Dose-Volume Histograms', *Radiat Res Suppl* **8**(2), S13–S19.

McKinney, W. (2010), Data structures for statistical computing in Python, *in* 'Proc Python Sci Conf', Vol. 445, pp. 51–56.

Mehta, C. R. & Senchaudhuri, P. (2003), 'Conditional versus unconditional exact tests for comparing two binomials', *Am Stat* **47**, 91–98.

Meirovitz, A., Murdoch-Kinch, C. A., Schipper, M., Pan, C. & Eisbruch, A. (2006), 'Grading xerostomia by physicians or by patients after intensity-modulated radiotherapy of head-and-neck cancer.', *Int J Radiat Oncol Biol Phys* **66**(2), 445–453.

Miah, A. B., Gulliford, S. L., Clark, C. H., Bhide, S. A., Zaidi, S. H., Newbold, K. L., Harrington, K. J. & Nutting, C. M. (2013), 'Dose-response analysis of parotid gland function: what is the best measure of xerostomia?', *Radiother Oncol* **106**(3), 341–345.

Miller, A. B., Hoogstraten, B., Staquet, M. & Winkler, A. (1981), 'Reporting results of cancer treatment', *Cancer* **47**(1), 207–214.

Mohan, R., Mageras, G., Baldwin, B., Brewster, L., Kutcher, G., Leibel, S., Burman, C., Ling, C. & Fuks, Z. (1992), 'Clinically relevant optimization of 3-D conformal treatments', *Med Phys* **19**(4), 933–944.

Molinaro, A. M., Simon, R. & Pfeiffer, R. M. (2005), 'Prediction error estimation: a comparison of resampling methods', *Bioinformatics* **21**(15), 3301–3307.

Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G. & Woodward, M. (2012), 'Risk prediction models: II. External validation, model updating, and impact assessment', *Heart* **98**(9), 691–698.

Munro, T. R. & Gilbert, C. W. (1961), 'The relation between tumour lethal doses and the radiosensitivity of tumour cells.', *Br J Radiol* **34**(400), 246–251.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambridge, MA, USA.

Nanci, A. (2017), *Ten Cate's Oral Histology-E-Book: Development, Structure, and Function*, Elsevier Health Sciences.

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

Nemenyi, P. B. (1963), Distribution-free multiple comparisons, PhD thesis, Princeton University.

Ng, A. Y. (2004), Feature selection, L1 vs. L2 regularization, and rotational invariance, *in* 'Proc Int Conf Mach Learn', p. 78.

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., A'Hern, R. P., Sydenham, M. A., Emson, M. & Hall, E. (2011), 'Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial', *Lancet Oncol* **12**(2), 127–136.

Obuchowski, N. A. & Lieber, M. L. (1998), 'Confidence intervals for the receiver operating characteristic area in studies with small samples', *Acad Radiol* **5**(8), 561–571.

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A. & Moore, J. H. (2017), 'Data-driven advice for applying machine learning to bioinformatics problems', *arXiv preprint arXiv:1708.05070* .

Osorio, E. M. V., Hoogeman, M. S., Al-Mamgani, A., Teguh, D. N., Levendag, P. C. & Heijmen, B. J. (2008), 'Local anatomic changes in parotid and submandibular glands during radiotherapy for oropharynx cancer and correlation with dose, studied in detail with nonrigid registration', *Int J Radiat Oncol Biol Phys* **70**(3), 875–882.

Ospina, J. D., Zhu, J., Chira, C., Bossi, A., Delobel, J. B., Beckendorf, V., Dubray, B., Lagrange, J. L., Correa, J. C., Simon, A., Acosta, O. & De Crevoisier, R. (2014), 'Random forests to predict rectal toxicity following prostate cancer radiation therapy', *Int J Radiat Oncol Biol Phys* **89**(5), 1024–1031.

Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M. M., Lambin, P. & Aerts, H. J. (2015), 'Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer', *Front Oncol* **5**, 272.

Pavy, J., Denekamp, J., Letschert, J., Littbrand, B., Mornex, F., Bernier, J., Horiot, J.-c. & Bartelink, M. B. H. (1995), 'Late effects toxicity scoring: the SOMA scale', *Int J Radiat Oncol Biol Phys* **31**(5), 1043–1047.

Pawitan, Y. (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford, United Kingdom.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: machine learning in Python', *J Mach Learn Res* **12**, 2825–2830.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2011), 'pROC: an open-source package for R and S+ to analyze and compare ROC curves.', *BMC Bioinformatics* **12**(1), 77.

Rubin, P., Constine, L. S., Phillips, T. L., Fajardo, L. F. & Wasserman, H. (1995), 'Overview of late effects normal tissues (LENT) scoring system', *Int J Radiat Oncol Biol Phys* **31**(5), 1041–1042.

Sanchez-Nieto, B., Fenwick, J. F., Nahum, A. E. & Dearnaley, D. P. (2001), 'Biological dose surface maps: evaluation of 3D dose data for tubular organs', *Radiother Oncol* **61**(Suppl 1), S52.

Schultheiss, T. E., Orton, C. G. & Peck, R. A. (1983), 'Models in radiotherapy: volume effects.', *Med Phys* **10**(4), 410–415.

Seabold, S. & Perktold, J. (2010), Statsmodels: econometric and statistical modeling with Python, *in* 'Proc Python Sci Conf', Vol. 57, p. 61.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D. & Levy, S. (2005), 'A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis', *Bioinformatics* **21**(5), 631–643.

Sugiyama, M. & Kawanabe, M. (2012), *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*, The MIT Press, Cambridge, MA, USA.

Suit, H. D., Shalek, R. J. & Wette, R. (1965), 'Radiation response of C3H mouse mammary carcinoma evaluated in terms of cellular radiation sensitivity', *Cellular radiation biology* pp. 514–530.

Tang, J., Alelyani, S. & Liu, H. (2014), Feature selection for classification: a review, *in* C. C. Aggarwal, ed., 'Data Classification: Algorithms and Applications', CRC Press, Boca Raton, FL, USA, pp. 37–64.

Tomek, I. (1976), 'Two modifications of CNN', *IEEE Trans Syst Man Cybern* **6**, 769–772.

Trotti, A., Byhardt, R., Stetz, J., Gwede, C., Corn, B., Fu, K., Gunderson, L., McCormick, B., Morris, M., Rich, T., Shipley, W. & Curran, W. (2000), 'Common toxicity criteria: version 2.0. An improved reference for grading the acute effects of cancer treatment: impact on radiotherapy', *Int J Radiat Oncol Biol Phys* **47**(1), 13–47.

Trotti, A., Colevas, A. D., Setser, A. & Basch, E. (2007), 'Patient-reported outcomes and the evolution of adverse event reporting in oncology', *J Clin Oncol* **25**(32), 5121–5127.

Trotti, A., Colevas, a. D., Setser, A., Rusch, V., Jaques, D., Budach, V., Langer, C., Murphy, B., Cumberlin, R., Coleman, C. N. & Rubin, P. (2003), 'CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment', *Semin Radiat Oncol* **13**(3), 176–181.

Tseng, H. H., Luo, Y., Cui, S., Chien, J. T., Ten Haken, R. K. & Naqa, I. E. (2017), 'Deep reinforcement learning for automated radiation adaptation in lung cancer', *Med Phys* **44**(12), 6690–6705.

Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. (2015), 'A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities', *Phys Med Biol* **60**(14), 5471–5496.

Van der Walt, S., Colbert, S. C. & Varoquaux, G. (2011), 'The NumPy array: a structure for efficient numerical computation', *Comput Sci Eng* **13**(2), 22–30.

Van Dijk, L. V., Brouwer, C. L., van der Schaaf, A., Burgerhof, J. G. M., Beukinga, R. J., Langendijk, J. A., Sijtsema, N. M. & Steenbakkers, R. J. H. M. (2017), 'CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva', *Radiother Oncol* **122**(2), 185–191.

Wijers, O. B., Levendag, P. C., Braaksma, M. M. J., Boonzaaijer, M., Visch, L. L. & Schmitz, P. I. M. (2002), 'Patients with head and neck cancer cured by radiation therapy: A survey of the dry mouth syndrome in long-term survivors', *Head Neck* **24**(8), 737–747.

Wilson, D. R. (1972), 'Asymptotic properties of nearest neighbor rules using edited data', *IEEE Trans Syst Man Cybern* **2**(3), 408–421.

Withers, H. R., Taylor, J. M. & Maciejewski, B. (1988), 'Treatment volume and tissue tolerance.', *Int J Radiat Oncol Biol Phys* **14**(4), 751–759.

World Health Organization and others (1979), 'Who handbook for reporting results of cancer treatment'.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J R Stat Soc Series B Stat Methodol* **67**(2), 301–320.

# LIST OF PUBLICATIONS

1. Gabryś, S.H., Buettner, F., Sterzing, F., Hauswald, H., Bangert, M. (2017). **Parotid gland mean dose as a xerostomia predictor in low-dose domains.** Act Oncol 56(9), 1197-1203.[1]

2. Gabryś, S.H., Buettner, F., Sterzing, F., Hauswald, H., Bangert, M. (2018). **Design and selection of machine learning methods using radiomics and dosiomics for NTCP modeling of xerostomia.** Front Oncol 8, 35.[2]

[1] The work presented in this paper is described in chapter 3.
[2] The work presented in this paper is described in chapters 4 and 5.

# ACKNOWLEDGEMENTS

I would like to express my gratitude to

- my supervisor Dr. Mark Bangert for his outstanding support, dedication, and unlimited enthusiasm throughout the whole project,
- my doctoral advisers Prof. Markus Alber and Prof. Wolfgang Schlegel,
- the members of my thesis advisory committee Dr. Florian Buettner and Dr. Florian Sterzing,
- the German Cancer Research Center (DKFZ) and the Helmholtz International Graduate School for Cancer Research for awarding me a stipend and giving an opportunity to pursue my PhD,
- everyone who helped with patient data retrieval from the databases of Heidelberg University Hospital, especially Dr. Henrik Hauswald, Dr. Dieter Oetzel, Dr. Kai Schubert, Dr. Sebastian Klueter, Henning Mescher, and Aleksander Emig,
- Dr. Sarah Gulliford, Prof. Uwe Oelfke, and James Morden for their support during my research visit at the Institute of Cancer Research,
- all members of the E0404 Radiotherapy Optimization group,
- my parents,
- Maria for support and limitless supply of Buchweizen.

# AFFIDAVIT

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema
   *Machine learning using radiomics and dosiomics for normal tissue complication probability modeling of radiation-induced xerostomia*
   handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift