Image based feature extraction to improve survival analysis in head and neck cancer

Anton Linnér

Master's thesis
2024:E70

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Abstract

In this thesis we performed a pooled cohort study to investigate the role of radiomics in head and neck cancer prognosis. The aim was to investigate prognostic value for overall survival and cancer recurrence of radiomics combined with previously studied demographic and clinical risk factors. Radiomics features were extracted from the gross tumor volume on a computed tomography captured prior to radiotherapy treatment. Both standard statistical models such as Cox regression, and common machine learning methods such as random survival forest, DeepSurv and DeepHit, were used. The cancer type was constricted to oropharyngeal head and neck cancer due to large amount of missing data in the other head and neck cancer types. Prognostic performance for local recurrence was improved using shape related radiomics (sphericity) and clustering based methods (PCA). In contrast, the results showed no improved performance for overall survival (OS) for any model, where a possible reason might be too few events per covariate or that OS depends mainly on factors not captured by the radiomics data.

These results indicate a role for radiomics in prognostic evaluation, which could prove to be useful treatment decision making and research guidance.

Keywords: Radiomics, Head and neck cancer, oropharynx, Cox regression, DeepSurv, DeepHit, Random survival forest, overall survival, local recurrence

Popular summary

More people get cancer today than ever before, and according to the World Health Organization (WHO) the trend in cancer cases will keep rising in the near future with an increase of 77 % from 2020 to 2050. Despite the increase in cancer cases, cancer mortality seems to be decreasing. One reason that is commonly given for the decrease in mortality is improved treatment methods. Decisions on treatment are made through patient specific information (such as age or overall health status) and tumor specific information (such as cancer type). Tumor specific information can be obtained through biopsy or medical imaging. One way to quantify information in medical images is radiomics, which is a method that is currently not used to inform treatment decision making.

Radiomics extracts information from images with the aim to describe the region of interest in the image (such as the tumor) with great detail. The aim of this thesis was to use survival analysis to evaluate if radiomics could help in predicting survival and recurrence in head and neck cancer, with the goal to potentially use radiomics in treatment decision making in the future.

To do this, both standard statistical methods such as Cox regression, and machine learning methods such as DeepSurv, DeepHit and Random survival forest, were used. These models were used to model the probability of getting the event (either death or cancer recurrence) over time. The machine learning methods were only used to model death probability.

The results showed that neither standard statistical models or machine learning models could utilize the radiomics to improve the probability predictions for the case of death. Radiomics did seem to provide an improvement for recurrence prediction. There therefore seems to be promise to use the radiomic information to inform on how treatment decisions are made.

Acknowledgments

I want to thank my supervisors André Haraldsson and Anna Lindgren for giving me their time and guidance during this thesis. I also want to thank my family and friends, with a special thanks to Musti Kadhim.

Contents

Chapter 1

Introduction

Cancer cases are expected to increase by 77 % from 2020 to 2050 (World Health Organization 2024). Despite the expected increase of cancer incidence, cancer mortality appears decreasing in many developed countries (Hashim et al. 2016, Siegel, Giaquinto, and Jemal 2024), where the three main reasons presented are: declines in smoking, early detection and improved treatments. This suggests the importance of cancer treatment and research.

There are four common ways to treat cancer: radiotherapy, chemotherapy, immunotherapy and surgery. Treatment modality is based on several patient and tumor specific factors. Patient specific factors can be information of the patients overall health (such as age and diseases unrelated to the cancer), whereas tumor specific factors are factors such as the size and cancer type.

Tumor specific information is often acquired by biopsy and by a doctor through medical imaging such as X-ray computed tomography (CT). Some characteristics of the tumor might be difficult to ascertain with imaging, and often general and subjective descriptions of the tumor are made. Moreover, biopsies might give limited information as tumors can be heterogeneous and a small sample might not characterize the full tumor. Radiomics is a novel method that uses computer algorithms to characterize regions of an image with the goal of describing the details of the region in standardized ways (Aerts et al. 2014). The radiomic features of the image describes complicated patterns and shapes of 2D or 3D images. If radiomic features meaningfully characterize tumor information in a complementary way to biopsy or visual estimation there is potential for them to be used to inform treatment choices and guide future research. One of the difficulties with using radiomic features is that they are often highly correlated with each other. This is an issue that can make it hard to get consistent results when radiomics are used in statistical models.

A branch of statistics called *survival analysis* (Andersen et al. 1993) investigate how factors affect survival outcome. It describes the time-to-event, T, for an event of interest such as death or disease recurrence. It takes into account the problem of censoring, which is when there are only partly known observations. Censoring is the reason why standard regression and statistical models can not be used.

One way to model time-to-event in survival analysis is the Kaplan-Meier estimator (KM-estimator) which was introduced in 1958 (Kaplan and Meier 1958) and is a cornerstone to survival analysis that is still used today. The KM-estimator is a nonparametric univariate method which is easy to implement and straightforward to interpret. In 1972 David Cox developed a regression method that could do multivariate modelling (D. R. Cox 1972), which was the next large step in survival analysis. This model is called the Cox proportional hazards model (Cox PH) and is together with the KM-estimator one of the most common statistical methods for time-to-event modelling in the medical field. In 2008 a decision tree-based machine learning method based on the already popular random forest method was introduced, referred to as random survival forest (RSF) (Ishwaran et al. 2008), which is one of the common machine learning approaches today. Some more recent machine learning models are DeepSurv (Katzman et al. 2018) and DeepHit (Lee et al. 2018) which are feed forward neural networks that can use complex relationships between covariates to predict survival outcomes.

1.1 Background and previous research

One common type of cancer is head and neck cancer (H&N cancer) with 890,000 new cases and 450,000 deaths annually worldwide (Global Cancer Observatory 2023), which accounts for roughly 4.5% of cancer diagnoses and deaths (Barsouk et al. 2023). An overview for common H&N cancer locations can be seen in Figure 1.1.

For oropharyngeal H&N cancer specifically (a region within pharynx), the human papillomavirus (HPV) has been shown to be a prognostic marker and might affect treatment recommendations (Patel et al. 2020). It is often measured through the presence of the tumor suppressing protein known as p16. In addition to HPV-p16 status, the size of the tumor has shown promise to be an additional marker in oropharyngeal H&N cancer for both overall survival and local cancer recurrence[1] (Adrian et al. 2022). It was shown that both tumor volume and having a negative p16 status were associated with a worse prognosis, where tumor volume had a more significant negative effect on survival and recurrence prognosis for HPV-p16 positive patients. There has been no conclusion on how radiomics might aid the existing H&N cancer risk factors, such as volume and p16 status for oropharyngeal cancer, when estimating overall survival and local recurrence.

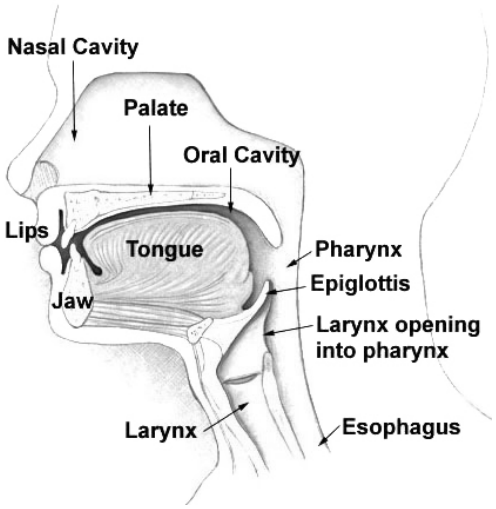

Figure 1.1: Head and neck overview (National Cancer Institute 2023).

The aim of this project was to do a pooled cohort study to investigate how radiomics could improve upon existing models and how they might expand the understanding of the cancer characteristic in H&N cancer. In addition to this, machine learning methods were used in order to see if they better incorporate the information from radiomics in combination with clinical/demographic features and compared to standard statistical methods.

[1] Overall survival refers to death by any cause, and local recurrence refers to recurrence of the cancer close to the original location.

1.2 Data

Two data sets were available for this project. The first data set, ARTSCANIII, was obtained from a prospective randomized trial comparing chemotherapy treatment methods for H&N cancer (Gebre-Medhin et al. 2021). Inclusion criteria were established to ensure some homogeneity of the patient group, such as health conditions unrelated to the cancer, and patients with previous cancer surgery were excluded. All patients were treated with radiotherapy. The second data set was an open data set from patients treated at MAASTRO Clinic, The Netherlands (Wee and Dekker 2019), referred to as H&N1. This data set contains a variety of treatments and has no strict exclusion criteria. ARTSCANIII included 299 patients and H&N1 included 137 H&N patients. To deal with missing values, listwise deletion was implemented, where observations were deleted if they had a missing value in any of the features used in the model.

Chapter 2

Theory

2.1 Survival analysis

Survival analysis is a branch of statistics that aims to model the time, T, to a certain event of interest in a subject. For example, when investigating the time from treatment of a disease to recurrence of the disease. Questions that survival analysis tries to answer are, what is the probability of experiencing the event at a certain point in time? What is the rate of the event given that the subject is event free up until that point? Is it possible to investigate how certain groups and characteristics affect the event rate and event probability? Is it possible to model several mutually exclusive events at the same time?

There are a few key concepts in survival analysis that are important,

- Censoring,
- Survival function and hazard function,
- Competing risks.

2.1.1 Censoring in survival analysis

Survival data must include both the duration that the subject is monitored, and an indicator of an event. If the event did not occur during the monitored time, the duration is recorded but the subject is considered *censored*.

Censoring in survival analysis refers to when certain subjects (e.g. patients) have only a partially observed survival time. More specifically, censoring can occur if the observed survival time is monitored from an unknown starting point or if the time is only known up until a certain time point, t. If the starting point is unknown the censoring is called *left-censoring* and

if the survival time is only known up until a certain time point it is called *right-censoring*. An example of right censoring is a patient who drops out of a study at time t_0 without having experienced an event. It is therefore only possible to know that the patient had *not* experienced the event at time t_0, but after that point the status of the patient is unknown, and considered right censored, demonstrated in Figure 2.1. Left-censoring can be explained in the case of a trial looking at a specific disease, where the interest is modelling the time from onset of disease to death due to the disease. In this case one might not know the true onset time of the disease but rather the time of diagnosis, and the full survival time is partially unknown. In this project, the event start time will be after completed radiotherapy treatment which means only right censoring will be considered.

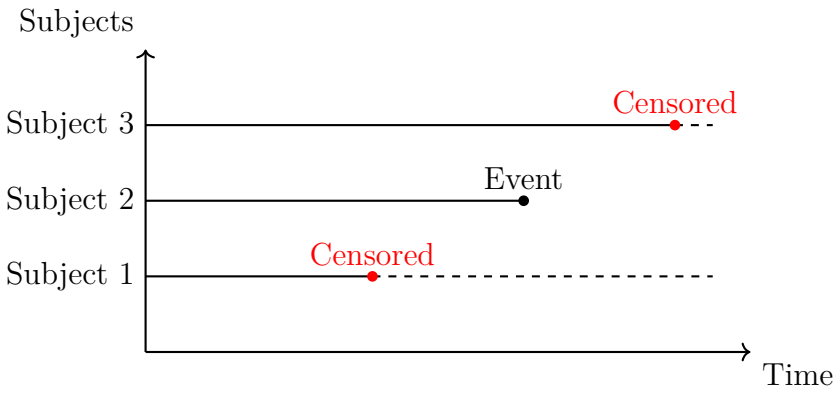

Figure 2.1: Illustration of right censoring

2.1.2 Survival function and hazard function

In survival analysis one commonly wants to characterize the time to event (also called survival time), T, with the survival function which is defined as $S(t) = P(T \geq t)$. The survival function is the probability of observing a survival time greater than some time t. In some instances the cumulative incidence function, CIF, is used. In the case where there is only one event of interest, the CIF is simply 1 - $S(t)$ and defines the probability of observing a survival time less than some time t. In standard survival analysis the survival function is more common to use, but in the context of competing risk (discussed more in section 2.3) the CIF is more commonly used.

Estimating the survival function for different groups or individuals will give insights on their survival probability over time. One of the more common estimators of the survival function is the Kaplan-Meier estimator (KM-

estimator). The KM-estimator uses both censored and uncensored data to estimate the survival function. It does this over time intervals obtained by the ordered observed times for the subjects (including censored subjects), from shortest to longest. This gives a nonparametric estimation of the survival function,

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i}, \tag{2.1}$$

where t_i is the observed survival time of a subject i in the risk set R_i. The risk set, R_i, includes the index of all censored or uncensored subjects that has not experienced the event yet at time t_i. The number of subjects at risk at time t_i is denoted n_i and the number of people for which the event occurred at time t_i is denoted d_i. The KM-estimator assumes non-informative censoring which is independent of the event of interest and does not provide information about the distribution of survival times. To look at an example where this is not the case, we look at the case when our event of interest is recurrence of cancer. If censoring occurs due to the patients worsening health condition (caused by cancer recurrence), then the censoring and event of interest are dependent.

Another important concept in survival analysis is the hazard function, $h(t)$, defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}. \tag{2.2}$$

The hazard function describes the instantaneous rate of an event occurring at time t, given that the event has not happened yet. The cumulative hazard, $H(t)$ is the integral of the hazard function up to time t, and relates to the instantaneous hazard through:

$$H(t) = \int_0^t h(u)\, du. \tag{2.3}$$

The cumulative hazard is related to the survival function in the following way:

$$S(t) = e^{-H(t)}. \tag{2.4}$$

This means that to estimate the survival function, one could estimate the cumulative hazard or hazard function and use equation 2.4 to obtain the survival function.

2.2 Cox proportional hazards model.

The Cox proportional hazards model (Cox PH) is a way to relate the hazard function to a set of covariates through:

$$h(t \mid \mathbf{X}_i) = h_0(t)e^{\mathbf{X}_i\beta}, \tag{2.5}$$

where β is a column vector of coefficients for a subject, i, with the covariates represented by the row vector $\mathbf{X}_i \in \mathbb{R}^p$ where p is the number of covariates. The function $h_0(t)$ is called the baseline hazard function and is the hazard function where the covariates are 0, or in some cases the mean or median. The baseline hazard function relates to the baseline survival function through:

$$S_0(t) = \exp\left[-\int_0^t h_0(u)\,du\right] \tag{2.6}$$

which relates to the survival function by:

$$S(t \mid \mathbf{X}_i) = S_0(t)^{\exp(\mathbf{X}_i\beta)}. \tag{2.7}$$

The baseline cumulative hazard function can be estimated using the Breslow estimator (Hosmer, Lemeshow, and May 2008, p. 175):

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R_i} \exp(X_j\beta)}. \tag{2.8}$$

One can then get the baseline survival function estimate through equation 2.4. Once the β coefficients are obtained, the baseline survival function estimate can be used to calculate the full survival function through equation 2.7.

The part of equation 2.5 that does not depend on time, $\exp(\mathbf{X}_i\beta)$, relates the covariates to the overall hazard function. The exponent, $\mathbf{X}_i\beta$, describes the relative risk for a subject i and is often described as the risk function, $r(\mathbf{X}_i) = \mathbf{X}_i\beta$. For two subjects with covariates $\mathbf{X}_i$ and $\mathbf{X}_j$ and $r(\mathbf{X}_i) > r(\mathbf{X}_j)$, the subject i will have a higher probability to experience the event compared to the subject j.

The Cox proportional hazard model is sometimes referred to as a semi-parametric model since the baseline hazard function, $h_0(t)$, does not assume a parametric form and is estimated nonparametrically. This is in contrast to the $\exp(\mathbf{X}_i\beta)$ which is parametric with the coefficients β.

Hazard ratio and proportional hazard

An assumption of the Cox proportional hazards model is that it assumes proportional hazards. To explain this assumption, we start with explaining

a concept called the *hazard ratio*. In the Cox model, hazard ratio is the ratio between the hazard of two observations having two different values of a covariate (commonly a difference of 1) while the other covariates are constant. This is a common method to interpret the effects of covariates on the hazard function. In the univariate case for a subject with one covariate, X_j, the hazard ratio for the covariate increasing from 0 to 1 is shown below:

$$\frac{h(t|X_j = 1)}{h(t|X_j = 0)} = \frac{h_0(t)\exp(\beta \cdot 1)}{h_0(t)\exp(\beta \cdot 0)} = \exp(\beta) \tag{2.9}$$

This can be extended to multiple covariates by holding all other covariates constant except for the covariate of interest. The hazard ratio describes how much the instantaneous rate of the event increases with an increase of one step of the covariate. The ratio between the hazard functions of two different hazards is equal to a constant that does not depend on time and is therefore *proportional*. If the ratios would depend on time, the hazards would no longer be proportional and the assumption would not hold. This also means that the information of covariates are assumed to be obtained at the time of gathering them, and their effect stays the same independent of how much time has passed.

Likelihood coefficient estimation

To estimate the β-coefficients a partial likelihood function was introduced in 1975, (D. Cox 1975), and is shown below. Since the baseline hazard function is not used in this expression, it is called the *partial* likelihood function. If there are no tied observation times for individuals that experienced the event of interest, the partial likelihood function, given observations $\mathbf{X}$, is given by:

$$\mathcal{L}(\beta \mid \mathbf{X}) = \prod_{i:c_i=1} \frac{\exp(\mathbf{X}_i\beta)}{\sum_{j \in R_i} \exp(\mathbf{X}_j\beta)}, \tag{2.10}$$

where c_i is an indicator for if the event happened ($c_i = 1$) or the subject was censored ($c_i = 0$). The possibility of ties makes it difficult to assess who is in the risk set. If two subjects experience the event at the same time it is ambiguous as to who will be considered in the other persons risk set. For a set D_i containing the subjects with ties at time t_i, the tied corrected likelihood function is given by:

$$\mathcal{L}(\beta \mid \mathbf{X}) = \prod_{j:c_j=1} \frac{\prod_{i \in D_j} \exp(\mathbf{X}_i\beta)}{\prod_{\ell=0}^{m_j-1} \left[\sum_{i \in R_j} \exp(\mathbf{X}_i\beta) - \frac{\ell}{m_j} \sum_{i \in D_j} \exp(\mathbf{X}_i\beta)\right]}, \tag{2.11}$$

where $m_j = \mid D_j \mid$.

To check the significance of the model, or the individual coefficients, either Walds test or the likelihood ratio test is commonly used (Hosmer, Lemeshow, and May 2008, p. 77).

2.3 Competing risk

Up until now, only a single event of interest has been considered. A common complication is when there are competing risks that are mutually exclusive to the event of interest, see Figure 2.2. To ground the concept of competing risk in an example, the example of cancer recurrence and death from any cause can be used. If the subject dies from any cause (related or unrelated to the cancer), it is no longer possible to get cancer recurrence. Death is therefore a competing risk to cancer recurrence and will have to be considered when looking at survival time modeling for the event of interest. An initial thought might be to treat subjects who suffer a competing risk before the event of interest as censored. This is common when obtaining the hazard rate of that specific cause. However, the relation between the hazard function for a specific cause and the survival function for that cause is no longer straightforward.

One option is to use the *cause specific* cumulative incidence function (CIF_k)

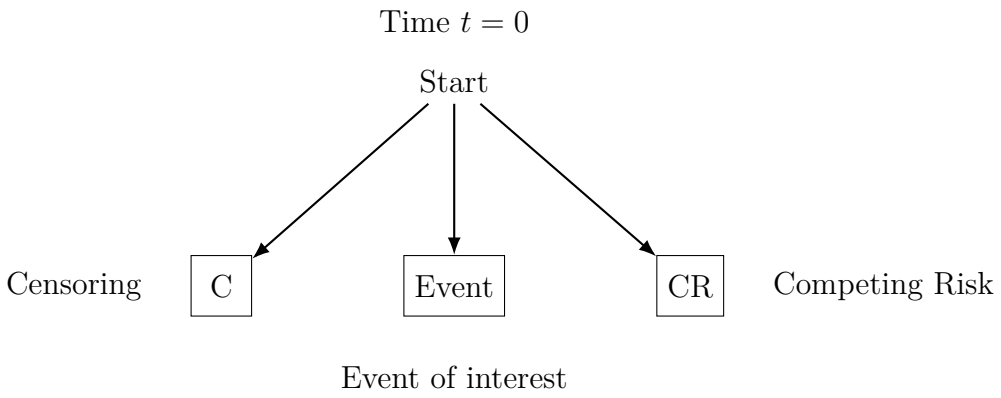

Figure 2.2: Illustrating the possible event outcomes in the presence of competing risks (CR). The event of interest can not occur if the competing risk occurs.

in the presence of competing risks. This is the probability of the event occurring before a specific time for that specific cause, $\text{CIF}_k(t) = \Pr(T \leq t, D = k)$. In a competing risk free setting, the CIF is just equal to 1 - $S(t)$,

which means it would be possible to use the KM-estimator to estimate it. This is however not possible in the presence of competing risks.

The cause specific hazard function is shown in equation 2.12. This is the instantaneous rate of experiencing the event of cause k at a specific time, given that you have not experienced any event.

$$h_k(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t} \tag{2.12}$$

Using the cause specific hazard function, we can define the cause specific cumulative hazard function as:

$$H_k(t) = \int_0^t h_k(u)\, du. \tag{2.13}$$

Using $H_k(t)$, an expression for the overall survival function can be obtained through the following expression:

$$S(t) = \exp\left(-\sum_{i=1}^{k} H_k(t)\right). \tag{2.14}$$

Using equation 2.14, we can define the cause specific CIF as:

$$\mathrm{CIF}_k(t) = \int_0^t h_k(u) S(u)\, du. \tag{2.15}$$

As can be seen, the cause specific CIF_k depends on all causes through the overall survival function, equation 2.14.

2.3.1 Cause specific Cox regression model

The cause specific Cox regression model is similar to the standard Cox regression model, with the key difference that it looks at the hazard function for only one specific cause and ignoring the rest of the causes,

$$h_k(t \mid \mathbf{X}_i) = h_{k0}(t) \exp(\mathbf{X}_i \beta_k), \tag{2.16}$$

where k is the cause of interest. The coefficient estimation is the same as in the standard Cox regression model, where one deals with each cause separately.

2.4 Machine learning

Due to the complexity that censoring adds to the statistics of survival analysis, there needs to be modifications to standard machine learning techniques before they are applicable to survival data. In this section, some basic concepts in machine learning will be explained that are deemed to be of importance in this study. The models modified for survival analysis will then be presented.

Specifically, two machine learning concepts are important:

- Decision trees
- Multi-layer perceptrons (MLPs)

Decision trees

Decision trees take a set of covariates, $\mathbf{X}$, as input and divides the data points based on the values of the covariates, as can be seen in Figure 2.3. First, the value of the covariates for subject i are used as input to a node called the root node. The root node is a node that starts the decision tree and therefore every data point will go through it. In the root node a decision is made that decides which node the subject will be passed on to based on the value of the covariate. The decision is made by setting a threshold for the covariate which dictates the path of the subject down the decision tree. The internal nodes, called decision nodes, act in the same way as the root node. The final nodes are called leaf nodes, which is where the data ends up after passing the root nodes and the decision nodes. This is the final classification of the data based on the covariates.

The decision on which covariate to split, and what the threshold should be that is used to split them, is based on increasing the homogeneity after

each split, which can be done through statistical tests comparing the characteristics of the groups after a split. Many different thresholds are tried, and the split with the best result is chosen.

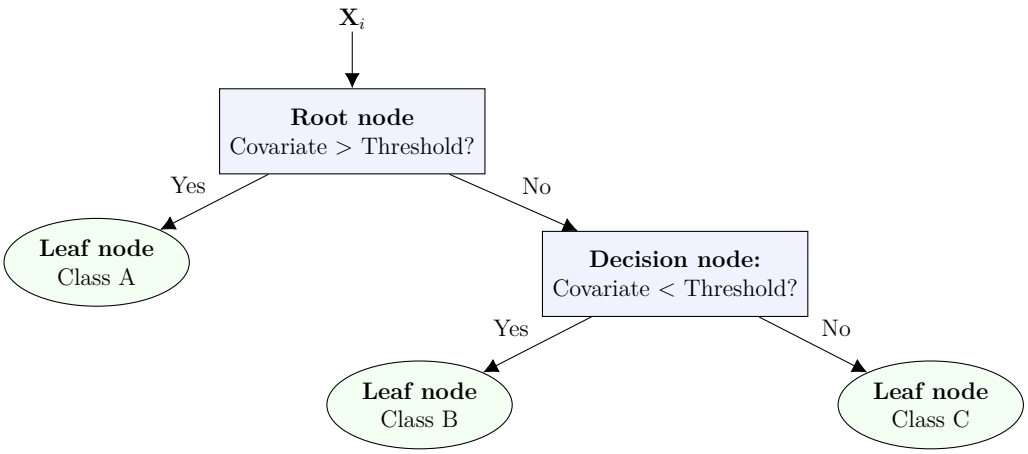

Figure 2.3: Example of decision tree structure.

Multi-layer perceptrons

Multi-layer perceptrons is a class of neural networks that are based on layers of nodes, where each node does some transformation of the data used as input. It starts with the covariates being passed to the input layer, with one node per covariate. This acts like a placeholder for the data and no transformation or modification is done to the data in this layer of nodes. The data is then passed to a layer called a **hidden** layer. When the data is passed to the hidden layer, they are scaled by weights, where each connection has their own weight. The number of nodes in the hidden layer is a parameter to be decided and does not have to match the number of nodes in the input layer. The number of hidden layers is another parameter that needs to be decided on, where more layers can find more abstract patterns in the data but usually need more data points. From the hidden layer(s), the nodes are connected to the output layer. These connections are also attached to a weight, transforming the value. These weights determine the importance of the data points and their transformations. The output layer is the prediction of the network and is compared to the known value that is associated with the input covariates, $\mathbf{X}$. This is evaluated by an error function that the network tries to minimize through gradient based methods. These calculate the gradient of the loss function with respect to the weights and then adjusts the weight in the direction of lower error. The steps the weights adjust is

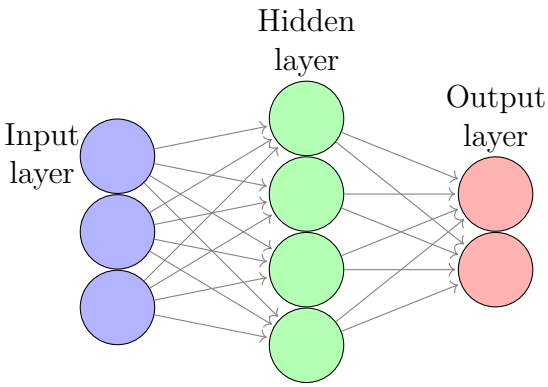

Figure 2.4: Standard MLP structure.

often referred to as the learning rate. In addition to the connections between nodes having weights, each node has an activation function which transforms the value in a non-linear way.

2.4.1 DeepSurv

The DeepSurv proportional hazards neural network is a network which aims to find the risk function $r(\mathbf{X}_i)$ for subject i. The network finds interactions and transformations of covariates that are otherwise hard to capture in standard proportional hazards models. Just like the Cox proportional hazard model discussed above, the DeepSurv network assumes proportional hazards (Katzman et al. 2018) and does not allow for time varying-covariates. It has a multi-layer perception structure which has a single node as output, see Figure 2.5, which represents the risk function. The loss function is the negative log likelihood of equation 2.10 with and added regularization parameter,

$$\mathcal{L}(\theta \mid \mathbf{X}_i) := -\frac{1}{N} \sum_{i:c_i=1} \left(\hat{r}_\theta(\mathbf{X}_i) - \log \sum_{j \in R_i} e^{\hat{r}_\theta(\mathbf{X}_j)} \right) + \lambda \|\theta\|^2 \tag{2.17}$$

Here θ are the weights used by the network and λ is a regularization parameter, which is a way to combat overfitting and to limit the number of features included.

Once the risk function is obtained, it is possible to get a risk score for a set of parameters $\mathbf{X}_i$ for subject i. Once the risk function is found it is possible to obtain the survival function by estimating the baseline hazard non-parametrically and then following the same steps described in the Cox regression section.

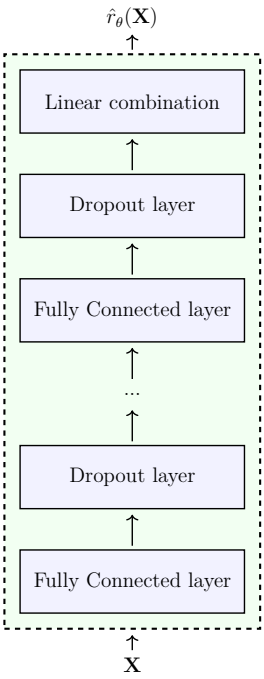

Figure 2.5: DeepSurv MLP model structure.

2.4.2 DeepHit

DeepHit is a neural network that has a similar structure to DeepSurv, in the sense that they are both MLPs, shown in Figure 2.6. However, where DeepSurv estimates the risk function, DeepHit estimates the survival function directly. This is done by first discretizing the time into discrete time points, $t \in \mathcal{T}$ where $\mathcal{T} = \{0, \ldots, T_{\max}\}$. Here T_{max} is the time horizon of interest, which is manually chosen based on the interest of the study. DeepHit tries to estimate the survival probability at certain time points, $P(T = t \,|\mathbf{X}_i)$. The main difference from DeepSurv is the output layer, which is a softmax layer instead of a linear layer, seen in Figure 2.6. The softmax output layer is a probability distribution for a subject with covariates $\mathbf{X}_i$, represented as $\mathbf{y} = [\mathrm{y}_1 \quad \mathrm{y}_2 \quad \ldots \quad \mathrm{y}_{T_{max}}]$. Here y_t is the estimated probability that a subject will experience the event at time point t, given some covariates $\mathbf{X}_i$.

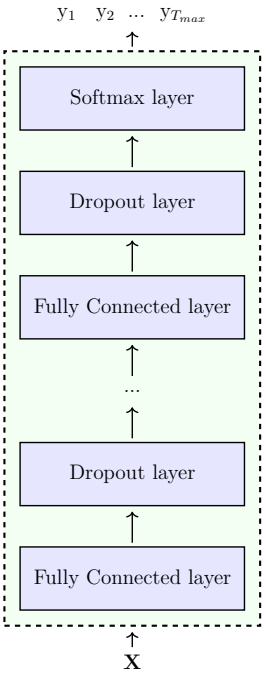

Figure 2.6: DeepHit MLP model structure

Since DeepHit does not estimate the risk function, it cannot use the same loss function as DeepSurv. Instead it uses a loss function that consists of functions that focus on different aspects of the predictions. The first loss function, for N subjects, is given by:

$$\mathcal{L}_1(\theta \mid \mathbf{X}) = -\sum_{i=1}^{N} [\mathbf{1}(c_i \neq 0) \cdot \log(\hat{y}_{t_i}) + \mathbf{1}(c_i = 0) \cdot \log\left(\hat{S}(t_i \mid \mathbf{X}_i)\right)], \tag{2.18}$$

where $\mathbf{1}$ is the indicator function and $\hat{S}(t_i \mid \mathbf{X}_i)$ is estimated by $\hat{S}(t_i \mid \mathbf{X}_i) = 1 - \sum_{t \in \mathcal{T}} y_t$. This evaluates the error made by the prediction of probability of the model. The second loss function looks at how well the model ranks different subjects based on their risk score:

$$\mathcal{L}_2(\theta \mid \mathbf{X}) = \sum_{i \neq j} c_i \mathbf{1}\{t_i < t_i\} \exp\left(\frac{\hat{S}(t_i \mid \mathbf{X}_i) - \hat{S}(t_i \mid \mathbf{X}_j)}{\sigma}\right) \tag{2.19}$$

where σ is a parameter that can be chosen to scale how much an incorrectly ranked pair of subjects will affect the loss function. The total loss

function is:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_1 + (1-\alpha)\mathcal{L}_2, \tag{2.20}$$

where α is a parameter that can be selected to further fine-tune how much each loss function should contribute to the total loss function. Once the probabilities are estimated, a linear interpolation between the time points can be done in order for a more accurate prediction (Kvamme and Borgan 2019).

2.4.3 Random survival forest

In standard random forest, many decision trees are trained simultaneously, hence the name random *forest*. The predictions are the ensemble mean of these trees, which means that each tree makes a prediction independently, and the final prediction is the mean of all the predictions from each tree. Each tree is only allowed a subset of the covariates and data points to train on, which is assigned randomly and is why it is called *random* forest. For each tree, a bootstrap sample of the observations is given for it to use as training data. A bootstrap sample is when observations are drawn from the original set with replacement. That means that the bootstrap sample can have the same observation several times. Typically, a bootstrap sample is as large as the original sample.

At each leaf node, the KM-estimator will be used to estimate the survival function for the observations in that leaf node, and the splits at each root and decision node are based on how well the split divides the subjects into groups with different survival functions. This is done by evaluating the KM-estimator for each group after a split, and performing a log-rank test to determine how different their survival functions are (Hosmer, Lemeshow, and May 2008, p. 77).

2.5 Model evaluation

Once the models are obtained, there needs to be metrics to evaluate their performance. Here two common metrics will be discussed, the concordance index and the Brier score.

2.5.1 Concordance

The concordance index for both the competing risk-free (standard) and competing risk case is explained below.

In general, concordance index ranks individuals based on their estimated risk of experiencing an event.

Standard survival analysis

The Frank Harrell concordance index (Harrell et al. 1982) is an accuracy measures that ranks individuals based on a risk score, $M(\mathbf{X}_i)$, for an individual with covariates $\mathbf{X}_i$. The risk score can be obtained from the estimated risk function, cumulative hazard function or survival probability. The risk score should be constructed such that a higher risk score means a higher risk of experiencing the event. To compare the risk scores of a pair of subjects where there are one or more censored subjects the ranking needs some considerations, since some pairs will be *non-comparable*. The pairs are therefore divided into two subsets:

- Comparable pairs. This is a pair where either:
  - i) Both experienced the event of interest,
  - ii) One subject experienced the event and the other subject is censored but with a shorter observed time than the subject that experienced the event.
- Non-comparable pairs. This is a pair where either:
  - i) Both individuals are censored
  - ii) One is censored and the other subject experienced the event with an event time earlier than the censored subject.

A comparable pair, each with covariates $\mathbf{X}_i$ and $\mathbf{X}_j$ and observed event times t_i and t_j, where $M(\mathbf{X}_i) < M(\mathbf{X}_j)$, are said to be concordant if $t_i > t_j$. If $t_j > t_i$ they are said to be discordant. This means that if the subject with high risk scores experiences the event sooner than the subject with low risk scores, the pair is said to be concordant. The concordance, C, is then:

$$C = \frac{\#Concordant\ pairs}{\#Comparable\ pairs} \tag{2.21}$$

A score of $C = 1$ means that the model ranks the pairs perfectly, and a score of $C = 0.5$ means that the model has no ability to rank the pairs correctly.

Competing risks

In the presence of competing risks, the same definition as in the standard case can be used with some modifications. If one considers the case where there is only one competing risk where the event indicator $k \in \{0,1,2\}$ indicates if the subject was censored ($k = 0$), had the event of interest ($k = 1$) or had the competing risk (k=2), the following addition of comparable pairs can be made: both of the subjects experienced an event where neither was censored, and not both experienced the competing risk. For a pair with individuals with covariates, event indicators and event times ($\mathbf{X}_i$, k_i, t_i) and ($\mathbf{X}_j$, k_j,t_j) respectively, where $M(\mathbf{X}_i) > M(\mathbf{X}_j)$ for k=1, the pair is said to be concordant if:

- $t_j < t_i$, $k_j = 2$,
- $t_j > t_i$, $k_j = 2$.

The reasoning for this, outlined in (Wolbers et al. 2014), is that $M(\mathbf{X}_i)$ is the risk of experiencing the event of interest $k = 1$. If a subject experiences the competing risk, then the event of interest cannot occur and the subject should be estimated to have a lower risk of experiencing the event of interest.

2.5.2 Brier score

The Brier score is a prediction error metric which uses the mean squared error of the estimated survival probability and the outcome (censored or event of interest). It also takes into account the censoring distribution in order to correct for bias introduced due to the censoring.

$$\mathrm{BS}^c(t) = \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}(t_i \leq t, c_i = 1)\frac{(0 - \hat{S}(t|\mathbf{X}_i))^2}{\hat{G}(t_i)} + \mathbf{1}(t_i > t)\frac{(1 - \hat{S}(t|\mathbf{X}_i))^2}{\hat{G}(t)}\Big] \tag{2.22}$$

Here n is the number of subjects, t_i is the observed time point of the censored or uncensored subject and c_i is an indicator for if the subject is censored or not. To adjust for censoring probability, $\hat{G}(t_i)$ is introduced as weighting where $\hat{G}(t_i)$ is the probability of being censoring *free*. If $\hat{G}(t_i) = 0.1$ for a specific time point t_i, it represents a high censoring risk at that time point. The subject will represent $1/0.1 = 10$ subjects to account for the possible censored individuals. This is called inverse-probability-of-censoring weighting (IPCW).

2.6 Principal Component Analysis (PCA)

One way to handle high dimensional feature spaces that suffer from multicollinearity is through principal component analysis (PCA). For a data set $\mathbf{X}$ with n rows and p columns, where each row is an observation and each column is a feature, PCA can be done by:

1. Standardize the data set.
2. Calculate the covariance matrix between the features.
3. Calculate the eigenvectors and eigenvalues for the covariance matrix
4. Transform the data using the eigenvectors.

The motivation is that the eigenvectors are the directions of highest variance which assumes to contain the most information.

The first step is to standardize the data since PCA is sensitive to differences in variance,

$$Z_{ip} = \frac{X_{ip} - \mu_p}{\sigma_p}, \tag{2.23}$$

where μ_p is the mean of feature $\mathbf{X}_p$, and σ_p is the standard deviation of the feature $\mathbf{X}_p$, estimated from the data set $\mathbf{X}$. $\mathbf{Z}$ is now the standardized form of $\mathbf{X}$.

The covariance matrix can be estimated through:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{Z}^T\mathbf{Z}. \tag{2.24}$$

The second step is to obtain, the eigenvectors, $\mathbf{u}$, and eigenvalues λ of the covariance matrix. The eigenvectors $\mathbf{u}$ are the principal components and the eigenvalues describe how much of the variance is explained in each eigenvector. The eigenvectors are the directions of axes with maximum variance. The eigenvectors are ordered by their eigenvalue, where a higher eigenvalue represents an eigenvector that explains more variance in the data. Since there are as many eigenvectors (and eigenvalues) as features, one wants to choose only a subset of eigenvectors. Let $\mathbf{P}$ be a matrix with the eigenvectors as columns, one can then transform the original standardized data to a new data set through the following transformation:

$$\mathbf{T} = \mathbf{ZP}. \tag{2.25}$$

$\mathbf{T}$ is now the new data set with as many features (columns) as eigenvectors chosen. If the variance of the data can be described by only a few principal components there can be a great dimensionality reduction.

2.7 Mutual information

Mutual information (MI) is a type of association between random variables which detects any relation between them, in contrast to correlation that only measures linear dependence. For two random variables X and Y, MI estimates the difference between the joint probability of (X, Y) and their marginal distributions through the equation:

$$I(X,Y) = E\left(\ln \frac{p(x,y)}{p(x)p(y)}\right) = \sum_x \sum_y p(x,y)[\ln p(x,y) - \ln p(x)p(y)] \quad (2.26)$$

As can be seen, if the random variables are independent, $p(x,y) = p(x)p(y)$ results in $I(X,Y) = 0$. Equation 2.26 can be estimated from data by estimating it using empirical distribution functions. This means that continuous variables need to be discretized, where a common bin width is Freedman-Diaconis rule:

$$h = 2\frac{\text{IQR}}{n^{1/3}} \quad (2.27)$$

where IQR is the interquartile range and n is the number of observations. The number of bins, k, is then calculated through:

$$k = \left\lceil \frac{\text{range of data}}{h} \right\rceil \quad (2.28)$$

Unlike correlation, it gives no indication as to in which direction the variables are related, only that a relation exists.

Chapter 3

Methodology

3.1 Data description, selection and pre-processing

3.1.1 Feature description

The features of the two data sets, ARTSCANIII and H&N1, can be divided into two categories, clinical/demographic features and radiomic features. The full feature set of the data sets will not be shown here, but important clinical/demographic and radiomic features will be described.

Clinical and demographic features

Important clinical and demographic features are collected in Table 3.1. These features can describe different clinical aspects and characteristics of the cancer, but also demographic information about the patient (such as age).

Table 3.1: Data characteristics for ARTSCANIII and H&N1, described as 'N (percentage)' unless otherwise specified.

Feature	**ARTSCANIII (N=299)**	**H&N1 (N=137)**
Age, years		
Median (range)	61 (33-77)	61 (44-83)
Missing	1 (<1)	0 (0)
Sex		
Male	239 (80)	111 (81)
Female	59 (20)	26 (19)
Missing	1 (<1)	0 (0)
Primary tumor site		
Oropharynx	253 (85)	88 (64)
Oral cavity	16 (5)	0 (0)
Larynx	12 (4)	49 (36)
Hypopharynx	17 (6)	0 (0)
Missing	1 (<1)	0 (0)
T stage		
T1	43 (14)	35 (26)
T2	115 (38)	32 (23)
T3	56 (19)	24 (18)
T4	84 (28)	46 (34)
Missing	1 (<1)	0 (0)
p16 status		
Positive	237 (79)	23 (17)
Negative	51 (17)	58 (42)
Missing	11 (4)	56 (41)
Survival status		
Censored	240 (80)	63 (46)
Not censored	54 (18)	74 (54)
Missing	5 (2)	0 (0)
Survival time (days)		
Median (range)	1052 (49-2121)	2778 (48-4789)
Missing	1 (<1)	0 (0)

Continued on next page

Table 3.1 continued from previous page

Feature	**ARTSCANIII (N=299)**	**H&N1 (N=137)**
Recurrence status		
Censored	253	113
Not censored	46 (51)	24 (48)
Missing	1 (<1)	0 (0)
Recurrence time (days)		
Median (range)	864 (64-1959)	1142 (36-3200)
Missing	1 (<1)	0 (0)

Radiomic features

The radiomic features are calculated from the PET/CT of the patient. The tumor image region that the features are calculated from are drawn by an oncologist. The radiomic are then calculated using the `PyRadiomics` package in python (Griethuysen et al. 2017). These features can be divided into 6 subgroups where each group tries to describe a different aspect of the image of the tumor. These subgroups are summarized in Table 3.2. In each subgroup there are several features, each described in detail in (Griethuysen et al. 2017) and a list of the 102 features used in this project is show in Table B.1.

Table 3.2: Overview of radiomic features by subgroup.

Subgroup Name	**Description**
Shape	This subgroup aims to characterize the shape of the tumor in both 2D and 3D space. Examples are the volume of the tumor and flatness of the tumor.
First Order	First order statistics of voxels[1] within the region of interest. Examples are the mean, median, and maximum voxel values.
Gray Level Co-occurrence Matrix (GLCM)	This describes the second order joint probability statistics of pairwise pixel values with a certain distance to each other (immediate neighbors are used here). This relates to the small scale structure of the image.
Gray Level Size Zone (GLSZM)	Describes the gray level zones in an image. Gray level zones are zones of connected voxels with the same gray level intensity.
Gray Level Run Length Matrix (GLRLM)	Describes the number of runs in a certain direction of voxel with the same gray level. This quantifies the number of consecutive gray levels in a certain direction.
Gray Level Dependence Matrix (GLDM)	Describes gray level dependencies in an image. Dependencies are defined as the number of voxels that are connected within a specified range that are dependent on a center voxel.

[1]Voxels are pixels in 3 dimensions.

There were no missing radiomic values for the patients in the H&N1 data set, but 7 patients (2 %) in the ARTSCANIII had missing values. Since the complete case method was implemented, these patients could only be used in radiomic-free models.

3.1.2 Pre-processing

It is common practice to scale features that are on different scales and magnitudes when using machine learning methods. Since the radiomics were on different scales, they were scaled using Min-Max scaling from the package `scikit-learn`. The scaling is mainly of importance on the gradient descent based models in this project (DeepHit and DeepSurv).

The T-stage feature is a categorical variable that has values from 1 to 4, where 4 indicates a tumor that is large and has spread to surrounding tissue and 1 indicates a less intrusive size of the tumor. To limit the number of levels used for this factor in the models, T-stage was refactored to small (T-stage = 1 or 2) or large (T-stage = 3 or 4). This turns the T-stage feature from a factor with four levels into a binary factor with two levels.

3.1.3 Patient selection

As can be seen in Table 3.1, ARTSCANIII had four specific H&N cancer diagnoses: hypopharynx, larynx, oral cavity and oropharynx, whereas the H&N1 data set only had larynx and oropharynx. Therefore, a first step of patient selection was to select patients with oropharynx or larynx cancer, since these two cancer types were common to both cohorts.

The p16 status in the H&N1 data set had 41 % missing values. However, most of the missing values of p16 status were with patients in larynx cancer, where only one out of 49 larynx cancer patients had recorded p16 status. Only eight out of 88 of the oropharyngeal cancer patients had missing values for p16 status in the H&N1 data set. Since p16 is of clinical importance for the most prevalent cancer diagnosis, oropharynx, it was therefore decided to only select oropharyngeal cancer patients from the two cohorts in our study. Additionally, since no patients in the ARTSCANIII had treatment by cancer surgery, only patients who did not have cancer surgery were selected.

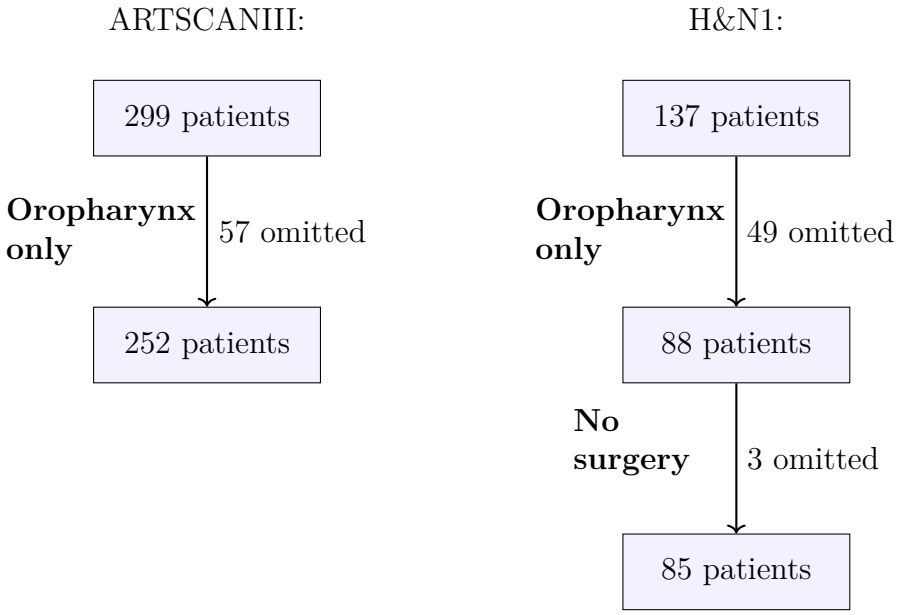

Figure 3.1: Patient selection from the ARTSCANIII and H&N1.

3.2 Data split and feature selection

3.2.1 Train and test split

ARTSCANIII was used as the main training set for the models due to its strict selection criteria for the included patients. The data was split into a test set and a training set. The test set consisted of 15 % of the ARTSCANIII set, and the training set consisted of the remaining 85 %. The training set was used to fit the models, and the models were then evaluated on the test set which acted as an out-of-sample data set. The H&N1 data set was used as an additional test set for evaluation from an external cohort to see how

well the models generalized to other cohorts. This finally results in a training set of 205 patients, and two test sets of 37 patients (ARTSCANIII test set) and 85 patients (H&N1 test set).

3.2.2 Feature selection

An important aspect of feature selection is deciding on the number of features. In survival analysis, the number of features to use can be related to the number of uncensored events in the data set. There is no exact rule for the choice of number of features, but a common standard is to have 10 events per feature included (Ogundimu, Altman, and Collins 2016). For neural networks, there might be more or less events needed, but we will use the same standard for all models. In the training set there are 205 observations and 27 events of death and 27 events of recurrence. An upper limit is chosen to be 3 covariates which is 9 events per feature for both standard survival analysis using machine learning.

There were different feature selection methods based on which of the features were looked at (clinical/demographic or radiomic). For the choice of demographic and clinical predictors, manual selection was implemented based on prior research or relevancy. T-stage and p16 status have shown to be relevant factors for both local recurrence and overall survival in H&N oropharyngeal cancer (Patel et al. 2020; Adrian et al. 2022). These were chosen to be used to represent the reference feature set, which will be denoted as the *demographic* feature from now.

For the choice of radiomic features, two filtering approaches were implemented, *full filtering* and *subgroup filtering*, see Figure 3.2. For the filtering methods, the first thing that is done is to retrieve the observed survival times for the uncensored events (for both overall survival, OS, and local recurrence, REC). This means that these filtered observed event times are only for when an event was actually observed, not for censored events. For the full filtering method, the mutual information is then calculated between the filtered observed event times, and the radiomics. The features with the highest value of mutual information are then chosen to be candidate features to use. In the next step, the mutual information is calculated between *one* of the demographic features chosen, and the radiomic features. Only radiomics that are below the 30th percentile are considered. The final feature set are the features with the highest MI with the event of interest which simultaneously had a low MI with one of the relevant reference features. Since two reference features were chosen and the covariate limit for all limits was chosen to be 3, only one radiomic will be included in addition to the reference covariates.

Subgroup filtering is done by also filtering out the observed event times that are not censored. The radiomics are then divided into the subgroups described in Table 3.2 and the mutual information between each subgroup and the observed event times is calculated. In the next step, one feature is selected from each subgroup based on which feature had the highest mutual information in that specific subgroup. The final feature set with candidate features is therefore one feature from each subgroup. The feature chosen is the highest MI-feature that is not in the same subgroup as any of the features in full filtering sets.

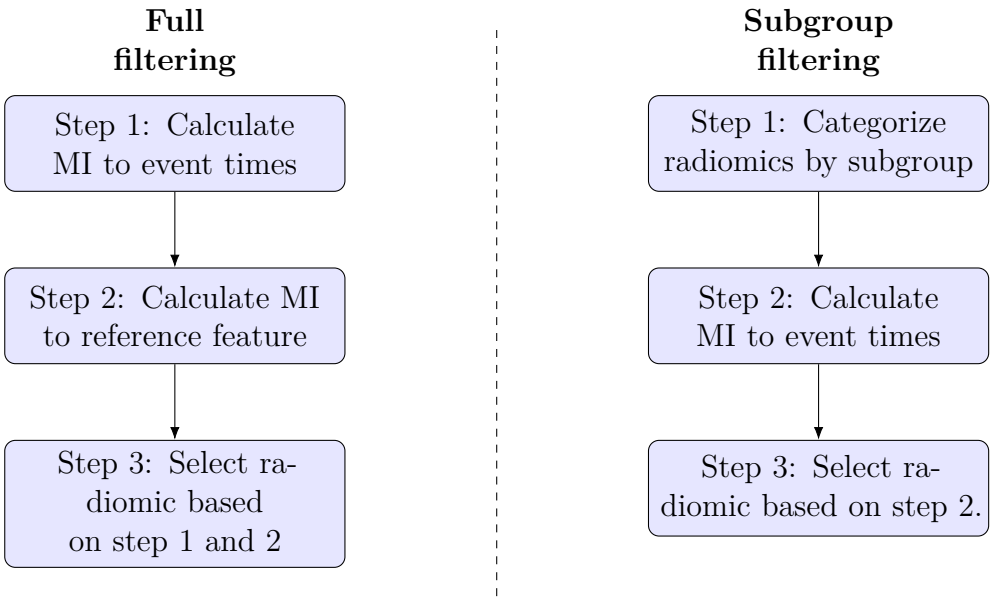

Figure 3.2: Overview of the two filtering methods used in feature selection.

In addition to the filtering steps, PCA will be implemented where two cases will be looked at:

- Reference model features + 1 PC feature
- Only PC features (3 are chosen due to this being the limit of covariates used in models).

There will be 5 feature sets in the end, that will be compared to the demographic set: 2 feature sets from full filtering (with low MI against p16 status and T-stage), one feature set from subgroup filtering, and the two PCA feature sets described above.

Motivation for full filtering method

The motivation for the full filtering method is that we want to find covariates with high association with the event of interest. But it is also of interest to have covariates with low association with prevalent, already used, clinical covariates. This method might however suffer the problem of multicollinearity if more than one feature is chosen, but since only one radiomic was chosen this will not be an issue.

Motivation for subgroup filtering method

Since the radiomic features are already divided into subgroups that aim to capture different aspects of the tumor, it might be useful to utilize these subgroups as a basis of feature selection. This might provide features with different explanatory characteristics.

3.3 Model training and evaluation

Once the feature sets are chosen through the different methods explained in the feature selection method, the models need to be developed. There are in total five models to be developed, where four are models that are concerned with the overall survival probability and one is concerned with local recurrence in the presence of competing risks (death from any cause). The models are described in theory but repeated here in Table 3.3.

Table 3.3: Summary of survival analysis models used.

Model	Application
Standard survival analysis	
Cox regression	Standard approach for hazard modeling
DeepSurv	Neural network-based, estimates risk function.
DeepHit	Neural network-based, estimates survival function
Random survival forest (RSF)	Decision tree-based
Presence of competing risks	
Cause specific Cox regression	Handles competing risks for recurrence analysis

To fit the Cox regression model, the package `survival` in R was used, using the function `coxph`. The `pycox` package was used to fit the DeepSurv and DeepHit models, and the `scikit-survival` package was used to fit the random survival forest model.

3.3.1 Hyperparameter optimization

For the machine learning models, some parameters need to be chosen manually or through optimization. These parameters are commonly denoted as hyperparameters, where the parameters considered are shown in Table 3.4. The optimization was done using the package `optuna` in python. In order to avoid overfitting, k-fold cross validation was used. This validation method divides the data into k folds, and trains the data on k-1 folds and validates it on the remaining fold using an appropriate metric. In our case, the mean of the Brier score over the first four consecutive years of observations was used as a metric. When the model was optimized, it ran a k-fold cross validation each time to obtain a score used in the optuna optimization process. The batch size and dropout for DeepHit was chosen to be the same as chosen for DeepSurv, in order to minimize optimization time. For both DeepSurv and DeepHit, 2 hidden layers with 32 nodes were used, with two dropout layers.

3.3.2 Accuracy evaluation

For each model, the concordance is calculated and presented to present how well the models ranked the individuals in the test sets based on their risk scores. In addition to this, the 1-4 year Brier scores are presented for each Cox regression model and cause specific Cox regression model, the β-coefficients are presented in addition to their respective statistical relevance through the p-value. Standard deviations for the concordance and Brier score were obtained through bootstrapping.

Table 3.4: Hyperparameters chosen to optimize.

Model	Hyperparameter	Description
RSF	Number of decision trees	Controls the complexity and performance of the model.
	Minimum samples leaf	Minimum number of subjects required to be a leaf node.
	Minimum samples split	Minimum number of samples required to split an internal node. Impacts how detailed the learned patterns are.
DeepSurv	Batch size	Number of samples before weights are updated.
	Dropout	Fraction of connections between layers of nodes that are omitted in order to generalize the model.
	Learning rate	Step size taken to reach the minimum in the loss function for the model.
DeepHit	Learning rate	Just as in DeepSurv, used for gradient based optimization.
	σ	Adjusts the punishment in the loss function related to accurately ranking the subjects
	α	Adjusts how much the model should consider ranking relative to prediction error in the loss function.

Chapter 4

Results

4.1 Overview

In this section each model will be presented separately to see how the features improve each model.

4.2 Feature sets

Two feature sets were obtained from full filtering, one with a radiomic feature with low MI with T-stage and one with p16 status. In addition to this, one with subgroup filtering was used. Two feature sets were obtained using PCA, one where the first principal component was used in addition to the reference features and one where only three principal components were used.

The feature sets are shown below in Table 4.1. The feature Max is a radiomic feature which is the maximum gray level intensity in the image region. The feature SDE is Small Dependence Emphasis which is a measure of the dependence in the region, which relates to how homogeneous the region is. A high value of SDE means that there is a low amount of dependencies and relates to less homogeneous textures. Flatness is a measure of how flat the object is, where a low value means a more flat-shaped object. For the local recurrence radiomics, IV is the Inverse Variance is a measure of variance for the joint probability distribution of the gray levels of voxel pairs. The inverse in the name stands for which gray levels are considered. RLN refers to how many similar run lengths (neighboring voxels with same gray levels) there are in the region where a lower value means more similar run lengths. Sphericity is a measure on how sphere-like the region is where a higher value means a more sphere-like region.

Table 4.1: Feature sets for overall survival and local recurrence.

Overall Survival	
Feature Set	**Features**
Demographic	T-stage p16 status
Filtered all 1	T-stage p16 status Max
Filtered all 2	T-stage p16 status SDE
Filtered subgroup	T-stage p16 status Flatness
Demo + PC1	T-stage p16 status PC1
PC Only	PC1 PC2 PC3

Local Recurrence	
Feature Set	**Features**
Demographic	T-stage p16 status
Filtered all 1	T-stage p16 status IV
Filtered all 2	T-stage p16 status RLN
Filtered subgroup	T-stage p16 status Spherecity
Demo + PC1	T-stage p16 status PC1
PC Only	PC1 PC2 PC3

4.3 Overall survival

Below the evaluation metrics for Cox PH, DeepSurv, DeepHit and RSF are presented in addition to a summary of the β-coefficients and their p-values for all feature sets. The standard error for the brier score for each model can be seen in Appendix A.

4.3.1 Cox proportional hazards

Table 4.2 shows the β-coefficients and p-values for the covariates in each feature set. The covariates in the reference set (demographic) both have p-values < 0.05. No other set has a significant covariate at a 5 % level except for the PCA Only set, where the first principal component is significant. In the Demo + PC1 set, the first principal component is close to significant at the 0.05 % level, but T-stage is less significant.

Table 4.2: Cox PH feature set comparison.

Feature set	Covariate	β (Std error)	p-value
Demographic	T-stage	1.112 (0.444)	0.0123
	p16 positive	-1.769 (0.403)	1.15e-05
Filter all 1	T-stage	1.128 (0.508)	0.0264
	p16 positive	-1.772 (0.405)	1.21e-05
	Max	-0.018 (0.268)	0.9469
Filter all 1	T-stage	1.118 (0.448)	0.0125
	p16 positive	-1.763 (0.407)	1.44e-05
	SDE	0.0243 (0.220)	0.9118
Filter all subgroup	T-stage	1.142 (0.457)	0.0125
	p16 positive	-1.746 (0.411)	2.14e-05
	Flatness	-0.066 (0.224)	0.7683
Demo + PC1	T-stage	0.785 (0.484)	0.1046
	p16 positive	-1.778 (0.408)	1.31e-05
	PC1	0.057 (0.032)	0.0787
PC only	T-stage	0.077 (0.032)	0.0162
	p16 positive	0.035 (0.028)	0.2197
	PC1	-0.038 (0.079)	0.6227

Figure 4.1 and 4.2 show the concordance and Brier score for the Cox PH model. Both of these metrics are worse for the H&N1 cohort, except for the concordance of the PCA Only set.

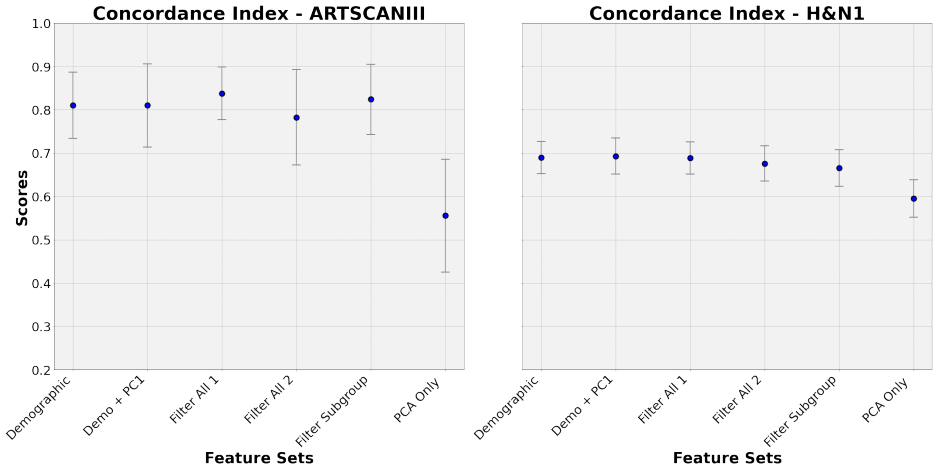

Figure 4.1: Concordance for the Cox PH model in the ARTSCANIII and H&N1 cohorts.

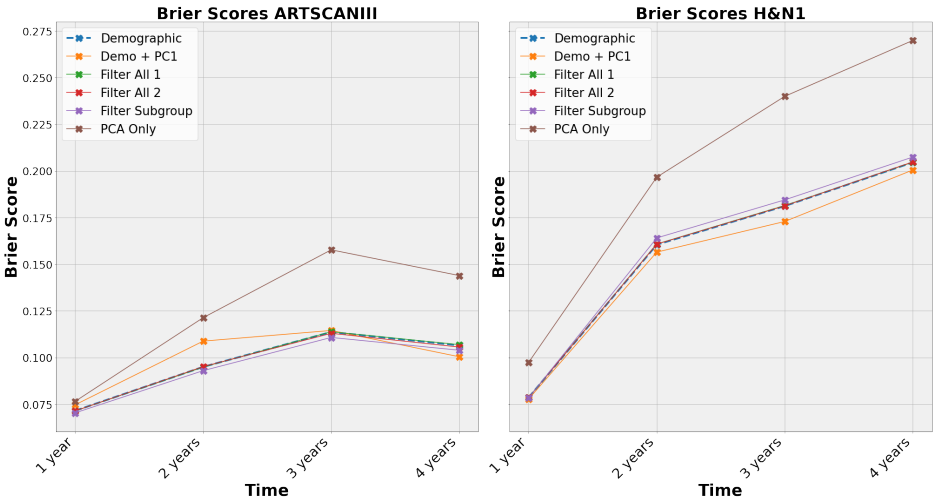

Figure 4.2: Brier score for the Cox PH model in the ARTSCANIII and H&N1 cohorts.

4.3.2 DeepSurv

Table 4.3 shows the hyperparameters for DeepSurv. In general, the batch size was stable across feature sets. The dropout is higher for all radiomic feature sets compared to the demographic set, with the exception of the Demo + PC1 set. The learning rate is not consistent across feature sets.

Table 4.3: Hyperparameters across feature sets.

	Batch Size	**Dropout**	**Learning rate**
Demographic	92	0.17	0.18
Filter All 1	100	0.31	0.28
Filter All 2	89	0.24	0.03
Filter All Subgroups	100	0.35	0.27
Demo + PC1	94	0.14	0.002
PCA Only	96	0.24	0.012

Figure 4.3 and 4.4 show the concordance and Brier score for the DeepSurv model. In both cohorts, the feature sets involving PCA transformations have the lowest concordance score, where they are at or below 0.5 concordance. The concordance in the ARTSCANIII cohort is in general higher than in the H&N1 cohort. The concordance is otherwise comparable between the sets, and the Brier scores are either equal to, or worse than, the demographic set for all feature sets.

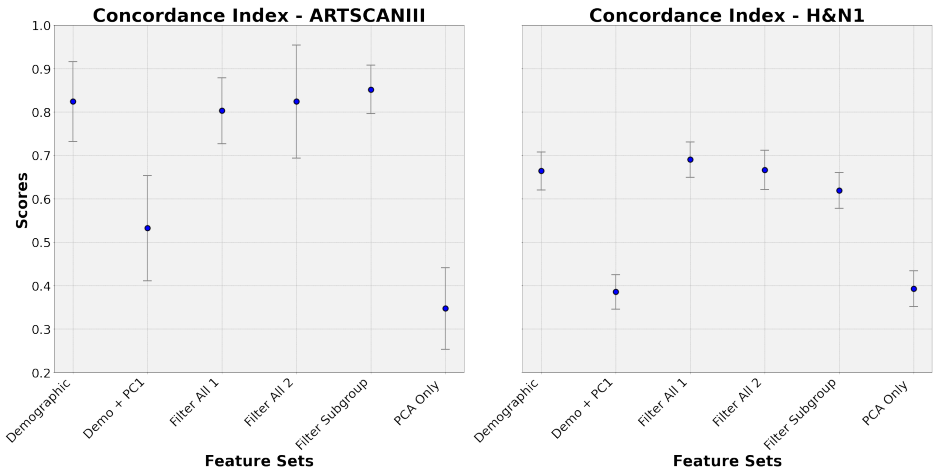

Figure 4.3: Concordance for the DeepSurv in the ARTSCANIII and H&N1 cohorts.

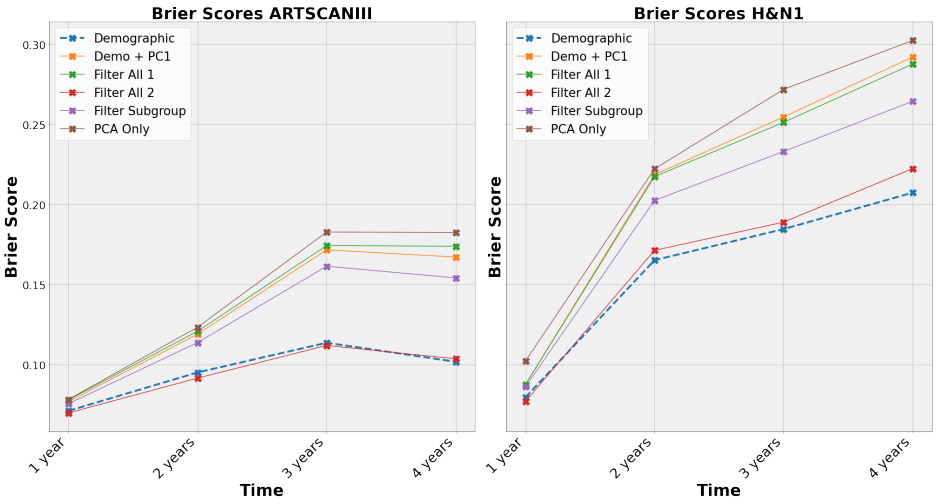

Figure 4.4: Brier score for the DeepSurv in the ARTSCANIII and H&N1 cohorts.

4.3.3 DeepHit

Table 4.4 shows the hyperparameters for the DeepHit model across all feature sets. The α-parameter is above 0.5 for all sets except for Filter All Subgroups. For all sets except for PCA Only, σ is below 0.5. It is lowest for the demographic reference set. The learning rate is consistent across sets.

Table 4.4: Hyperparameters across feature sets.

	α	σ	**Learning rate**
Demographic	0.84	0.17	0.012
Filter All 1	0.69	0.41	0.005
Filter All 2	0.85	0.32	0.007
Filter All Subgroups	0.45	0.47	0.004
Demo + PCI	0.74	0.24	0.012
PCA Only	0.75	0.66	0.001

Figure 4.5 and 4.6 show the concordance and Brier score for the DeepHit model. For the ARTSCANIII set, the PCA-related sets have the lowest concordance. All Brier scores are comparable or worse than the demographic set.

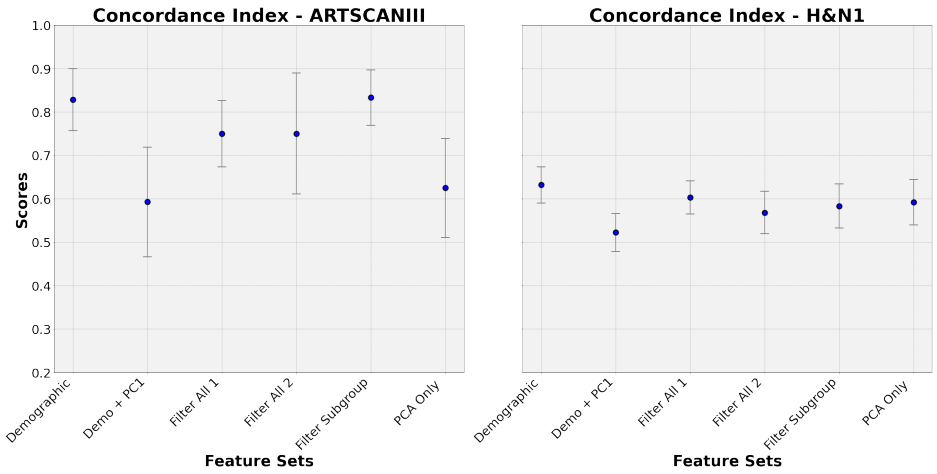

Figure 4.5: Concordance for the DeepHit in the ARTSCANIII and H&N1 cohorts.

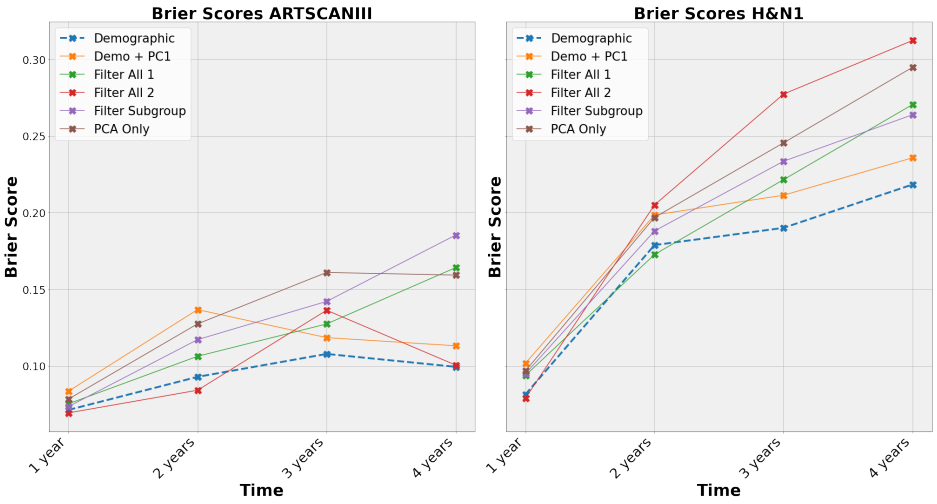

Figure 4.6: Brier score for the DeepHit in the ARTSCANIII and H&N1 cohorts.

4.3.4 Random survival forest (RSF)

Table 4.5 shows the hyperparameter results for the RSF model. Both minimum samples per leaf and per split are low. The number of trees chosen for each set is not consistent across sets.

Table 4.5: Hyperparameters across feature sets.

	Min samples leaf	Min samples split	Number of trees
Demographic	5	5	157
Filter All 1	7	4	77
Filter All 2	24	2	37
Filter All Subgroups	6	8	159
Demo + PCI	6	6	5
PCA Only	19	3	231

Figure 4.7 and 4.8 show the concordance and Brier score for the RSF model. Both scores are higher in the ARTSCANIII cohort than the H&N1 cohort for all feature sets. The concordance is either comparable to, or worse than, the demographic set. This is also the case for the Brier scores.

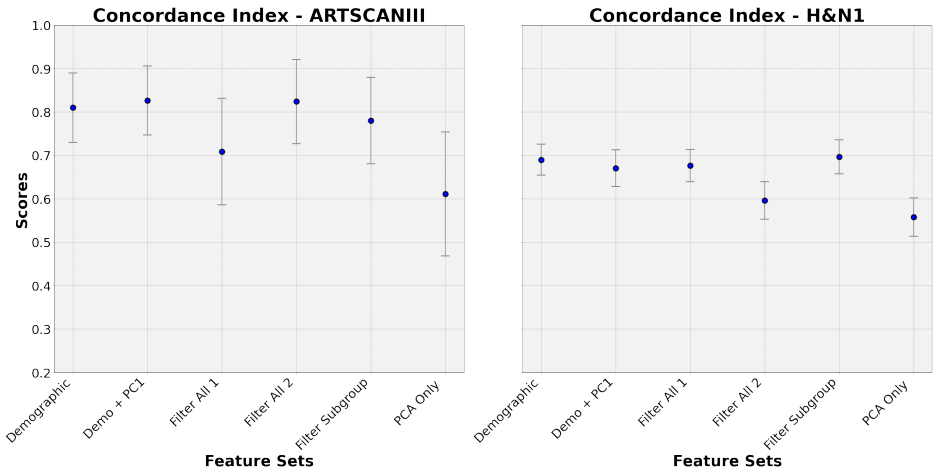

Figure 4.7: Concordance for the RSF model in the ARTSCANIII and H&N1 cohorts.

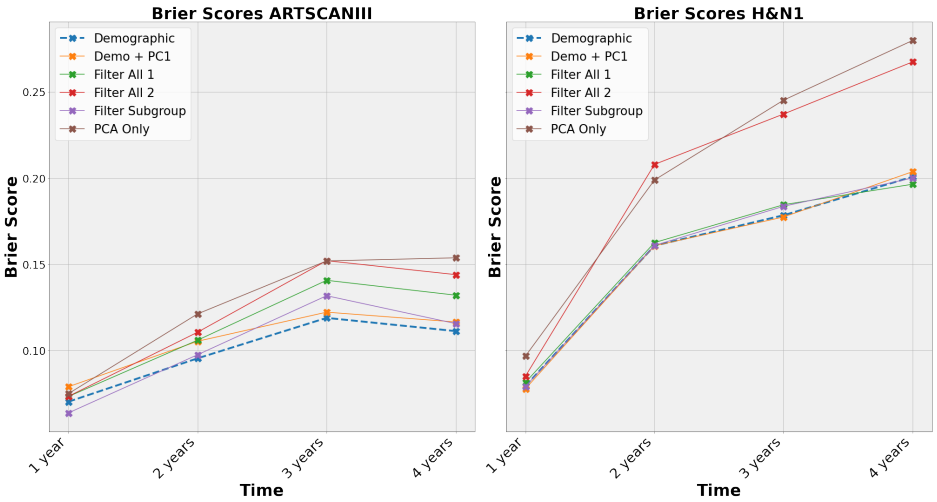

Figure 4.8: Brier scores for the RSF model in the ARTSCANIII and H&N1 cohorts.

4.4 Local recurrence

In this section the cause specific Cox regression model is presented, with a summary of β-coefficients and p-values in addition to the model evaluation metrics.

4.4.1 Cause specific Cox regression

Table 4.6 shows the coefficient values and p-values for each feature set. The PCA Only set has two significant features and one nearly significant feature. The PC1 feature is also significant in the Demo + PC1 set, and changes the magnitude of the T-stage coefficient. Sphericity is also significant in the Filter All Subgroup - set and changes the magnitude and sign of the T-stage covariate, in addition to changing the p-value in the direction of less significance. Figure 4.9 and 4.10 show the concordance and Brier score across

Table 4.6: Cause specific Cox PH feature set Comparison

Feature set	Covariate	β (Std error)	p-value
Demographic	T-stage	0.752 (0.420)	0.0736
	p16 positive	-1.324 (0.451)	0.0033
Filter all 1	T-stage	0.722 (0.426)	0.0264
	p16 positive	-1.382 (0.460)	0.0026
	IV	0.367 (0.207)	0.0768
Filter all 2	T-stage	0.423 (0.447)	0.0717
	p16 positive	-1.295 (0.455)	0.0044
	RLN	0.178 (0.206)	0.3879
Filter all subgroup	T-stage	-0.236 (0.460)	0.6076
	p16 positive	-1.415 (0.444)	0.0014
	Sphericity	-1.319 (0.370)	0.0003
Demo + PC1	T-stage	0.234 (0.456)	0.1046
	p16 positive	-1.382 (0.461)	0.0027
	PC1	0.104 (0.032)	0.0011
PC only	PC1	0.138 (0.034)	6.3e-05
	PC2	0.069 (0.026)	0.0089
	PC3	-0.163 (0.089)	0.0674

all feature sets for the cause specific Cox regression model. The concordance index for the ARTSCANIII seems to be higher than for the H&N1 cohort for all feature sets except for PCA Only set that seems to perform similarly in both cohorts. Most concordance are comparable or worse than the reference

set. All sets except Filter All 1 have a better Brier score, in both cohorts, than the reference group.

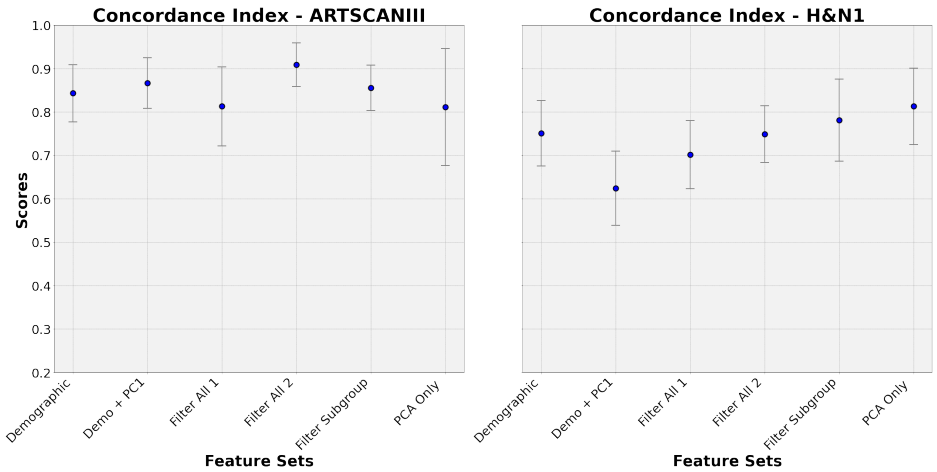

Figure 4.9: Concordance for the cause specific Cox PH model in the ARTSCANIII and H&N1 cohorts.

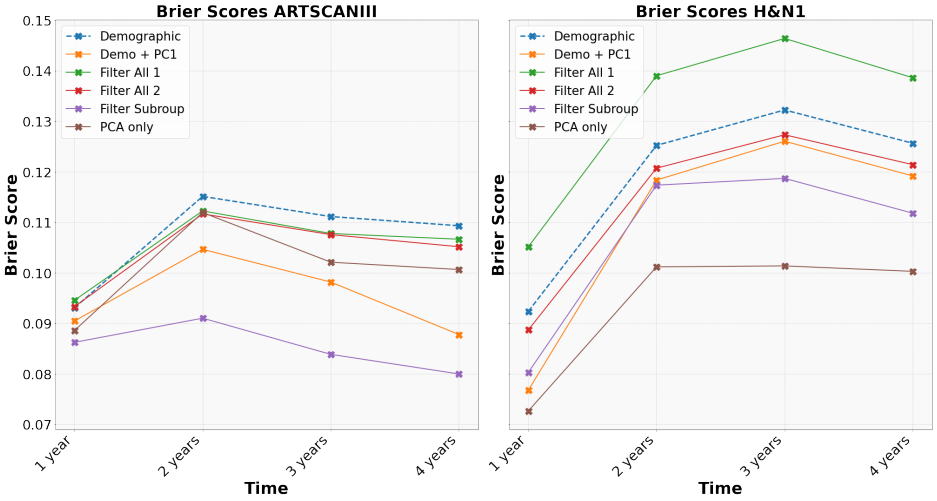

Figure 4.10: Brier score for the cause specific Cox PH model in the ARTSCANIII and H&N1 cohorts.

Chapter 5

Discussion

5.1 Discussion overview

The main goal of the thesis was to see if radiomic features could improve model performance and give insight into H&N cancer. The main part of the discussion will therefore be focused on comparisons between feature sets, but a part of the discussion will also be devoted to model comparison.

5.1.1 Cohort differences

Table 3.1 shows some differences in the T-stage and p16 distribution between the two cohorts. This, in addition to the stricter selection criteria of ARTSCANIII could contribute to the differences in evaluation metric scores between the two cohorts.

5.1.2 Overall survival

The concordance index across the feature sets showed no large improvement using radiomics for the Cox regression model. The only difference can be seen for the feature set using only PCA features, *PCA Only*, which has a worse performance than the reference set, *Demographic*. The results are similar for RSF and for DeepSurv and DeepHit for the concordance score, where the PCA sets seem to perform worse in general. The very low concordance scores for the PCA feature sets in the DeepHit model indicates that the model was not able to model the survival time well. A concordance score lower than 0.5 means that the model ranks the individuals in opposite order of how they experience the event. If a model shows a consistent low concordance, one could use the reverse predictions of the model to rank the subjects. But

more commonly, a very low concordance means that the results are unstable. Table 4.3 shows higher levels of dropout for all feature sets for DeepSurv, with the exception of the feature set using one PC in addition to the reference features. This could indicate some overfitting in the DeepSurv model, and could be one reason for the machine learning models not being able to improve using radiomic features. The Brier scores were equal to, or worse than, the reference set for all models, feature sets and cohorts. These results indicate that radiomics did not yield a meaningful improvement for overall survival.

5.1.3 Local recurrence

For the cause specific Cox regression, the concordance is comparable for the feature sets in the respective cohorts. For the Brier scores, one can see that all feature sets get a better score than the reference model, except for the Filter All 1 - set in the second cohort. Moreover, the PCA Only set gets a better Brier score than the reference set in both cohorts. This set uses neither of the demographic variables and still manages to get comparable or better results. In addition to this, one can see that for the feature set using one principal component in addition to the demographic features alters the magnitude of the coefficient for T-stage by a large margin. As previously mentioned, T-stage relates to the volume of the cancer. Since many radiomic features relate to the shape and size of the cancer it seems reasonable that radiomics might do a better job at describing this characteristic of the cancer. This might especially be true for PCA methods that incorporate information from all radiomic features. The sphericity feature in the subgroup feature set similarly alters the T-stage feature, both in magnitude and significance. Since the sphericity coefficient is negative this means that a larger value (more spherical) means lower risk for recurrence. This has been seen for head and neck cancer in the oral cavity (Tarsitano et al. 2019; Lucchi et al. 2023). One possible reason given is that an uneven surface relates to a more aggressive expanding tumor into adjacent tissue. They also saw that a more spherical tumor seemed to indicate a better prognosis. They saw it in both overall survival and local recurrence, where the effect of sphericity was only checked for local recurrence in this study.

These results indicate that radiomics had a positive impact on the prognostic accuracy for local recurrence. In addition to this, sphericity specifically shows results that have been seen in other H&N - cancer types. Radiomics being more influential for local recurrence than overall survival could be due to recurrence being more directly related to the cancer type compared to overall survival. Many factors that could affect overall survival, such as overall health condition, are not explained by radiomics.

5.1.4 Limitations and sources of error

In other studies, the stability of radiomics for patients has been used as a feature selection method (Aerts et al. 2014). The reasoning for this is that radiomics features that do not yield stable results for the same patient when different images are used are not as reliable as stable radiomics. This has not been taken into account in this thesis, and if the radiomics features used have unstable measurements it might affect the results of the models. In addition to this, how the region is drawn by the oncologist might differ between individuals, which might affect the radiomic values.

Another limitation might be the amount of data and events. If overfitting was a problem for the machine learning models, an increase of data could be a solution. In addition to this, if one wants to include more covariates to see if interactions between radiomics are useful, more data is needed.

One also has to make sure that conclusions drawn from radiomic research have reasonable connections to the epidemiological background of the cancer type. In this thesis, discussions were had with doctors from articles using ARTSCANIII (Adrian et al. 2022), but more thorough discussion and collaboration in study planning could be beneficial to ensure reasonable conclusions are made.

5.1.5 Future work

Future possibilities for the role of radiomics can be examining the role of shape specific radiomics to find consistent features, in collaboration with oncologists to ensure reasonable interpretation. Some safeguards for radiomics suggested (Welch et al. 2019) are:

- using standard software to ensure inter-institutional research. This relates to having standardized feature extraction from imaging,
- making sure to do external cohort testing,
- checking for multicollinearity among radiomics.

A good starting point could be checking the reliability of radiomics across images for the same patient, across institutions and imaging apparatus and professional doing region delineation. This should be done using standard software, such as the `pyrad` packages used in this thesis.

Cancer progression to lymph nodes is used to determine the cancer extent and complexity. In this project only images of the primary tumor were used, but this could be extended to images of lymph nodes to give a more detailed description of cancer progression.

5.1.6 Conclusion

For overall survival, neither standard statistical models such as the Cox PH model, nor the machine learning models RSF, DeepSurv or DeepHit, could see an improvement using radiomic features. The evaluation metrics were better in general for testing done on the ARTSCANIII cohort compared to the H&N1 cohort, which could be due to the greater homogeneity of the ARTSCANIII cohort due to the selection criteria.

For local recurrence, an improvement for the evaluation metrics could be seen when using the shape specific feature sphericity. This has been seen in other research for different H&N cancer locations (Tarsitano et al. 2019; Lucchi et al. 2023) and could indicate a feature of importance. One could see improved evaluation metrics when incorporating PCA features in the model in addition to the reference covariates, which shows the potential for clustering methods to aid in this field. Moreover, one could see that using only PCA features yielded better Brier scores compared to the reference model in both cohorts, suggesting a promising role of information obtained solely from radiomics.

This study shows some promise for using radiomic features for improved prognostic evaluation using non-invasive methods such as imaging. The information gained can then potentially be used to implement more aggressive treatment methods, in addition to pointing to new research areas to improve knowledge about different cancer types.

Bibliography

[1] Gabriel Adrian et al. "Primary tumor volume and prognosis for patients with p16-positive and p16-negative oropharyngeal squamous cell carcinoma treated with radiation therapy". English. In: *Radiation Oncology* 17.1 (Dec. 2022). ISSN: 1748-717X. DOI: 10.1186/s13014-022-02074-7.

[2] Hugo Aerts et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nature communications* 5 (Aug. 2014), p. 4006. DOI: 10.1038/ncomms5006.

[3] Per Kragh Andersen et al. *Statistical Models Based on Counting Processes*. 1st ed. Springer Series in Statistics. Published: 06 December 2012 (eBook), Published: 23 June 1995 (Softcover). Springer New York, NY, 1993, pp. XI, 784. ISBN: 978-0-387-94519-4. DOI: 10.1007/978-1-4612-4348-9. URL: https://doi.org/10.1007/978-1-4612-4348-9.

[4] Adam Barsouk et al. "Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma". In: *Medical Sciences* 11 (June 2023), p. 42. DOI: 10.3390/medsci11020042.

[5] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246. URL: http://www.jstor.org/stable/2985181 (visited on 05/26/2024).

[6] David Cox. "Partial Likelihood". In: *Biometrika* 62 (Aug. 1975), pp. 269–276. DOI: 10.1093/biomet/62.2.269.

[7] Maria Gebre-Medhin et al. "ARTSCAN III: A randomized phase III study comparing chemoradiotherapy with cisplatin versus cetuximab in patients with locoregionally advanced head and neck squamous cell cancer". English. In: *Journal of Clinical Oncology* 39.1 (2021), pp. 38–47. ISSN: 0732-183X. DOI: 10.1200/JCO.20.02072.

[8] Global Cancer Observatory. *Interactive Web-Based Platform for Global Cancer Statistics*. https://gco.iarc.fr. Accessed: 2024-05-22. 2023.

[9] Joost J.M. van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (Oct. 2017), e104–e107. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-17-0339. eprint: https://aacrjournals.org/cancerres/article-pdf/77/21/e104/2934659/e104.pdf. URL: https://doi.org/10.1158/0008-5472.CAN-17-0339.

[10] Jr Harrell Frank E. et al. "Evaluating the Yield of Medical Tests". In: *JAMA* 247.18 (May 1982), pp. 2543–2546. ISSN: 0098-7484. DOI: 10.1001/jama.1982.03320430047030. eprint: https://jamanetwork.com/journals/jama/articlepdf/372568/jama_247_18_030.pdf. URL: https://doi.org/10.1001/jama.1982.03320430047030.

[11] D. Hashim et al. "The global decrease in cancer mortality: Trends and disparities". In: *Annals of Oncology* 27.5 (May 2016), pp. 926–933. DOI: 10.1093/annonc/mdw027. URL: https://doi.org/10.1093/annonc/mdw027.

[12] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis*. John Wiley & Sons, Inc., 2008.

[13] Hemant Ishwaran et al. "Random survival forests". In: *The Annals of Applied Statistics* 2.3 (2008), pp. 841–860. DOI: 10.1214/08-AOAS169. URL: https://doi.org/10.1214/08-AOAS169.

[14] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501452. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452.

[15] Jared L. Katzman et al. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC Medical Research Methodology* 18.1 (Feb. 2018). ISSN: 1471-2288. DOI: 10.1186/s12874-018-0482-1. URL: http://dx.doi.org/10.1186/s12874-018-0482-1.

[16] Håvard Kvamme and Ørnulf Borgan. *Continuous and Discrete-Time Survival Prediction with Neural Networks*. 2019. arXiv: 1910.06724 [stat.ML].

[17] Changhee Lee et al. "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). DOI: 10.1609/aaai.v32i1.11842. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11842.

[18] Elisabetta Lucchi et al. "Pretreatment Tumor Volume and Tumor Sphericity as Prognostic Factors in Patients with Oral Cavity Squamous Cell Carcinoma: A Prospective Clinical Study in 95 Patients". In: *Journal of Personalized Medicine* 13.11 (2023). ISSN: 2075-4426. DOI: 10.3390/jpm13111601. URL: https://www.mdpi.com/2075-4426/13/11/1601.

[19] National Cancer Institute. *Head Neck Overview*. Accessed: May 26, 2024. 2023. URL: https://training.seer.cancer.gov/head-neck/anatomy/overview.html.

[20] Emmanuel O. Ogundimu, Douglas G. Altman, and Gary S. Collins. "Adequate sample size for developing prediction models is not simply related to events per variable". In: *Journal of Clinical Epidemiology* 76 (2016), pp. 175–182. ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2016.02.031. URL: https://www.sciencedirect.com/science/article/pii/S0895435616300117.

[21] Roshal R. Patel et al. "De-intensification of therapy in human papillomavirus associated oropharyngeal cancer: A systematic review of prospective trials". In: *Oral Oncology* 103 (2020), p. 104608. ISSN: 1368-8375. DOI: https://doi.org/10.1016/j.oraloncology.2020.104608. URL: https://www.sciencedirect.com/science/article/pii/S1368837520300440.

[22] Rebecca L. Siegel, Angela N. Giaquinto, and Ahmedin Jemal. "Cancer statistics, 2024". In: *CA: A Cancer Journal for Clinicians* 74.1 (2024), pp. 12–49. DOI: https://doi.org/10.3322/caac.21820. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21820. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21820.

[23] Achille Tarsitano et al. "Pretreatment tumor volume and tumor sphericity as prognostic factors in patients with oral cavity squamous cell carcinoma". In: *Journal of Cranio-Maxillofacial Surgery* 47.3 (2019), pp. 510–515. ISSN: 1010-5182. DOI: https://doi.org/10.1016/j.jcms.2018.12.019. URL: https://www.sciencedirect.com/science/article/pii/S1010518218308837.

[24] L. Wee and A. Dekker. *Data from HEAD-NECK-RADIOMICS-HN1*. Data set. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.8kap372n. 2019.

[25] Mattea L. Welch et al. "Vulnerabilities of radiomic signature development: The need for safeguards". In: *Radiotherapy and Oncology* 130 (2019), pp. 2–9. ISSN: 0167-8140. DOI: https://doi.org/10.1016/j.radonc.2018.10.027. URL: https://www.sciencedirect.com/science/article/pii/S0167814018335515.

[26] Marcel Wolbers et al. "Concordance for prognostic models with competing risks". In: *Biostatistics* 15.3 (Feb. 2014), pp. 526–539. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxt059. eprint: https://academic.oup.com/biostatistics/article-pdf/15/3/526/599536/kxt059.pdf. URL: https://doi.org/10.1093/biostatistics/kxt059.

[27] World Health Organization. *Global Cancer Burden Growing Amidst Mounting Need for Services*. Accessed: 2024-02-26. Feb. 2024. URL: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services (visited on 02/26/2024).

Chapter A

Brier score standard errors

A.1 Overall survival

A.1.1 Cox PH

Cox PH

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.03726	0.03732	0.03946	0.03681
Demo + PC1	0.03919	0.03904	0.04118	0.03648
Filter All 1	0.04078	0.04263	0.04613	0.04255
Filter All 2	0.03569	0.03807	0.04008	0.03781
Filter Subgroup	0.03866	0.03946	0.04093	0.03834
PCA Only	0.04758	0.04886	0.05163	0.04667

Figure A.1: Standard error for the Brier scores in the ARTSCANIII cohort

Cox PH

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.02344	0.01983	0.01802	0.02165
Demo + PC1	0.02261	0.01726	0.01698	0.01983
Filter All 1	0.02685	0.02085	0.02397	0.02314
Filter All 2	0.02202	0.02113	0.02121	0.02229
Filter Subgroup	0.02365	0.01982	0.02266	0.02238
PCA Only	0.03273	0.03479	0.03663	0.03692

Figure A.2: Standard error for the Brier scores in the H&N1 cohort

A.1.2 DeepSurv

DeepSurv

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.03682	0.03887	0.0439	0.03983
Demo + PC1	0.04039	0.04203	0.05572	0.05077
Filter All 1	0.04084	0.04176	0.05515	0.04977
Filter All 2	0.03629	0.03936	0.04445	0.04067
Filter Subgroup	0.04232	0.0408	0.04339	0.03858
PCA Only	0.04012	0.04236	0.06019	0.05605

Figure A.3: Standard error for the Brier scores in the ARTSCANIII cohort

DeepSurv

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.02562	0.02755	0.02746	0.02844
Demo + PC1	0.03056	0.04177	0.04142	0.04258
Filter All 1	0.03038	0.04065	0.03949	0.04014
Filter All 2	0.02372	0.02474	0.02533	0.02782
Filter Subgroup	0.02888	0.03358	0.03325	0.03291
PCA Only	0.03043	0.03546	0.04009	0.03699

Figure A.4: Standard error for the Brier scores in the H&N1 cohort

A.1.3 DeepHit

DeepHit

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.0398	0.03387	0.03149	0.03019
Demo + PC1	0.04536	0.05015	0.03807	0.03753
Filter All 1	0.03298	0.02967	0.02408	0.02895
Filter All 2	0.03883	0.03487	0.04171	0.03297
Filter Subgroup	0.02416	0.02018	0.01873	0.02583
PCA Only	0.04314	0.04737	0.04774	0.04731

Figure A.5: Standard error for the Brier scores in the ARTSCANIII cohort

DeepHit

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.02587	0.02932	0.02413	0.02715
Demo + PC1	0.0294	0.03328	0.03117	0.03288
Filter All 1	0.01563	0.01212	0.01637	0.02114
Filter All 2	0.0237	0.03585	0.03836	0.04106
Filter Subgroup	0.01613	0.01739	0.0152	0.01943
PCA Only	0.03018	0.03755	0.04147	0.04315

Figure A.6: Standard error for the Brier scores in the H&N1 cohort

A.1.4 Random Survival Forest (RSF)

RSF

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.03678	0.03776	0.04063	0.03858
Demo + PC1	0.0407	0.03481	0.03515	0.03668
Filter All 1	0.03842	0.03957	0.04369	0.04135
Filter All 2	0.03875	0.04089	0.045	0.03961
Filter Subgroup	0.03253	0.03451	0.041	0.03679
PCA Only	0.03931	0.04358	0.0435	0.03993

Figure A.7: Standard error for the Brier scores in the ARTSCANIII cohort

RSF

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.02598	0.02056	0.01952	0.01987
Demo + PC1	0.02473	0.01767	0.01648	0.01891
Filter All 1	0.02826	0.02396	0.02307	0.01703
Filter All 2	0.03026	0.03981	0.03837	0.03778
Filter Subgroup	0.02679	0.02338	0.02232	0.01965
PCA Only	0.02899	0.03483	0.03564	0.03425

Figure A.8: Standard error for the Brier scores in the H&N1 cohort

A.2 Local Recurrence

A.2.1 Cause specific Cox regression

Cause Specific Cox

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.0422	0.04017	0.03928	0.03928
Demo + PC1	0.03866	0.03463	0.03372	0.03379
Filter All 1	0.03874	0.03894	0.03811	0.03815
Filter All 2	0.0393	0.04143	0.04044	0.04058
Filter Subroup	0.03609	0.03145	0.03012	0.03057
PCA only	0.03928	0.04295	0.04174	0.04178

Figure A.9: Standard error for the Brier scores in the ARTSCANIII cohort

Cause Specific Cox

Feature set	Year 1	Year 2	Year 3	Year 4
Demographic	0.0172	0.01506	0.01449	0.01479
Demo + PC1	0.01397	0.01674	0.01715	0.01798
Filter All 1	0.02084	0.0176	0.01714	0.01789
Filter All 2	0.01718	0.01545	0.01494	0.01517
Filter Subroup	0.01996	0.02376	0.02475	0.02303
PCA only	0.02245	0.02737	0.02675	0.02679

Figure A.10: Standard error for the Brier scores in the H&N1 cohort

Chapter B

Full radiomic table

Table B.1: Full list of radiomic features, described as 'subgroup_Featurevalue'.

Features	**Features**
firstorder_10Percentile	firstorder_90Percentile
firstorder_Energy	firstorder_Entropy
firstorder_InterquartileRange	firstorder_Kurtosis
firstorder_Maximum	firstorder_Mean
firstorder_MeanAbsoluteDeviation	firstorder_Median
firstorder_Minimum	firstorder_Range
firstorder_RobustMeanAbsoluteDeviation	firstorder_RootMeanSquared
firstorder_Skewness	firstorder_TotalEnergy
firstorder_Uniformity	firstorder_Variance
glcm_Autocorrelation	glcm_ClusterProminence
glcm_ClusterShade	glcm_ClusterTendency
glcm_Contrast	glcm_Correlation
glcm_DifferenceAverage	glcm_DifferenceEntropy
glcm_DifferenceVariance	glcm_Id
glcm_Idm	glcm_Idmn
glcm_Idn	glcm_Imc1
glcm_Imc2	glcm_InverseVariance
glcm_JointAverage	glcm_JointEnergy
glcm_JointEntropy	glcm_MaximumProbability
glcm_SumEntropy	glcm_SumSquares
gldm_DependenceEntropy	gldm_DependenceNonUniformity
gldm_DependenceNonUniformityNormalized	gldm_DependenceVariance
gldm_GrayLevelNonUniformity	gldm_GrayLevelVariance

Feature 1	**Feature 2**
gldm_HighGrayLevelEmphasis	gldm_LargeDependenceEmphasis
gldm_LargeDependenceHighGrayLevelEmphasis	gldm_LargeDependenceLowGrayLevelEmphasis
gldm_LowGrayLevelEmphasis	gldm_SmallDependenceEmphasis
gldm_SmallDependenceHighGrayLevelEmphasis	gldm_SmallDependenceLowGrayLevelEmphasis
glrlm_GrayLevelNonUniformity	glrlm_GrayLevelNonUniformityNormalized
glrlm_GrayLevelVariance	glrlm_HighGrayLevelRunEmphasis
glrlm_LongRunEmphasis	glrlm_LongRunHighGrayLevelEmphasis
glrlm_LongRunLowGrayLevelEmphasis	glrlm_LowGrayLevelRunEmphasis
glrlm_RunEntropy	glrlm_RunLengthNonUniformity
glrlm_RunLengthNonUniformityNormalized	glrlm_RunPercentage
glrlm_RunVariance	glrlm_ShortRunEmphasis
glrlm_ShortRunHighGrayLevelEmphasis	glrlm_ShortRunLowGrayLevelEmphasis
glszm_GrayLevelNonUniformity	glszm_GrayLevelNonUniformityNormalized
glszm_GrayLevelVariance	glszm_HighGrayLevelZoneEmphasis
glszm_LargeAreaEmphasis	glszm_LargeAreaHighGrayLevelEmphasis
glszm_LargeAreaLowGrayLevelEmphasis	glszm_LowGrayLevelZoneEmphasis
glszm_SizeZoneNonUniformity	glszm_SizeZoneNonUniformityNormalized
glszm_SmallAreaEmphasis	glszm_SmallAreaHighGrayLevelEmphasis
glszm_SmallAreaLowGrayLevelEmphasis	glszm_ZoneEntropy
glszm_ZonePercentage	glszm_ZoneVariance
shape_Compactness1	shape_Compactness2
shape_Elongation	shape_Flatness
shape_LeastAxisLength	shape_MajorAxisLength
shape_Maximum2DDiameterColumn	shape_Maximum2DDiameterRow
shape_Maximum2DDiameterSlice	shape_Maximum3DDiameter
shape_MeshVolume	shape_MinorAxisLength
shape_SphericalDisproportion	shape_Sphericity
shape_SurfaceArea	shape_SurfaceVolumeRatio
shape_VoxelVolume	

Master's Theses in Mathematical Sciences 2024:E70
ISSN 1404-6342

LUNFMS-3130-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/