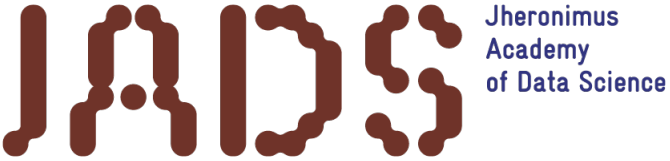


Data Science & Entrepreneurship

Deep Learning analysis of multi-sequence brain MRI for the prediction of survival in patients with glioma brain tumor using Logistic-Hazard

Master Thesis

Assil Jwair
2031096

Supervisors:
prof. dr. E.O. Postma
dr. R.S. Eijgelaar
dr. E. van der Poel

Final version

Den Bosch, Dec 2020

Preface

I was engaged in researching and writing this thesis from February to December 2020. It has been written to fulfill the graduation requirements of the Data Science & Entrepreneurship Master at the Jheronimus Academy of Data Science (JADS). I undertook my intership at the Neurosurgery department of the VU University Medical Center (VUmc). My research question was formulated together with my first supervisor prof. dr. Eric O. Postma and my on-site (VUmc) supervisor dr. Roelant S. Eijgelaar. The study was challenging, but after extensive research, I was able to successfully answer the research question. My second supervisor, dr. Egge van der Poel, guided me in the right direction for Chapter 5 (Social Entrepreneurship) and helped me to understand the social implications of AI in health care.

I would like to express my sincere gratitude to prof. dr. Eric Postma, dr. Roelant Eijgelaar and dr. Egge van der Poel for guiding me throughout my research. I would also like to thank everyone of the PICTURE research group, without their insights I would not have been able to conduct this research. I would like to thank dr. Philip de Witt Hamer, dr. Emilie Dronkers and Britt Rademaker who took the time to express their opinions about the use of AI in healthcare.

To conclude, I also benefited from brain storming ideas with my friends and family. I would like to thank Kimberley Boersma for all her love and support throughout my research.

I hope you enjoy your reading.

Assil Jwair

Abstract

Background: High-grade gliomas such as glioblastomas are the most deadly brain tumors with short overall survival time. Treatment planning and delivery of glioblastoma patients heavily relies on magnetic resonance imaging (MRI). Recent developments have shown that artificial neural networks can extract complex information from images. This has sparked the interest of using deep learning to predict survival times using MRI. Training a neural network to predict survival requires a differentiable loss function that can handle censored subjects (with an unknown time of death). Past studies have used the binary cross entropy loss, however this resulted in unreliable survival estimates.

Aim: To evaluate the feasibility of predicting individual survival curves using pre-operative MRI of glioblastoma patients by incorporating the Logistic-Hazard method in a 3D residual neural network (3D ResNet) model.

Methods: The Logistic-Hazard method uses a modified negative-log likelihood instead of a binary cross entropy loss function to predict reliable survival estimates over time. In this study we trained a 3D ResNet with Logistic-Hazard loss on multi-sequence MRI to predict the survival curves of glioblastoma patients. The performance of this model was compared to a Fully Connected Neural Network (FCNN) based on clinical variables of the same patients. The calibration of the models was evaluated using the inverse probability of censoring weighted Brier score (IPCW Brier score). The discriminative power was assessed using the concordance index (C-index).

Results: Both models were able to predict survival curves of glioblastoma patients. The model using 3D MRI achieved an IPCW Brier score of 0.131 compared to 0.117 of the model using clinical data. The C-indexes were respectively 0.621 and 0.697.

Conclusion: Using a 3D ResNet and Logistic-Hazard loss, we were able to predict the overall survival of glioblastoma patients using only pre-operative MRI. However, the calibration and discriminative power of the clinical model remained higher than the model using MR imaging. In conclusion, the Logistic-Hazard method can be used to predict well calibrated survival curves using either clinical variables or 3D MRI of glioblastoma patients.

Contents

Chapter 1

Introduction

According to the World Health Organization, high-grade gliomas such as glioblastomas are the most deadly brain tumors with short overall survival (OS) time (Wesseling & Capper, 2018). The 5 year relative survival rate for glioblastoma is 9% for patients aged 45-54 and 6% for patients aged 55-64 (Ostrom et al., 2019). Furthermore, the median survival does not exceed 14 months (Ostrom et al., 2019). Survival analyis is often used to study the correlation between covariates (inputs) of each patient and their survival time (Miller Jr, 2011). Accurate predictions of survival time for patients diagnosed with glioblastomas can help treatment planning and outcome prediction (S. Liu, Zheng, Feng & Li, 2017; Feng, Tustison, Patel & Meyer, 2020). In past studies, researchers have used survival analysis to estimate the survival time of glioblastoma patients based on both 3D MRI scans and clinical variables (Aerts et al., 2014; Lao et al., 2017; Nie, Zhang, Adeli, Liu & Shen, 2016; Hao, Kim, Mallavarapu, Oh & Kang, 2018).

Gliomas are the most common brain tumors and originate in the glial cells (Sun, Zhang & Luo, 2018). These type of brain tumors contain various histological sub-regions such as a necrotic core and a non-enhancing tumor core. Previous studies have found that the prognosis of patients with glioblastoma depends on several variables including age (Soffietti & Chio, 1989), extent of tumor resection (Ammirati, Vick, Youlian, Ivan & Mikhael, 1987), location of tumor (Gehan & Walker, 1977), chemotherapy (Walker et al., 1980) and neurological symptoms such as epilepsy (Scott & Gibberd, 1980). However, patients who receive the same type of therapy often have differences in survival rates (Hammoud, Sawaya, Shi, Thall & Leeds, 1996). These differences are thought to be related to the biological behaviour of brain tumors (Hammoud et al., 1996). Magnetic Resonance Imaging (MRI) is often used to portray this biological behaviour of gliomas. A study conducted by Hammoud et al. (1996) found, using MRI, that necrosis, enhancement and edema are prognostic variables of survival in glioblastoma patients. However, despite extensive research, the 5-year survival rate has remained almost the same for glioblastoma patients for the last 30 years (McLendon & Halperin, 2003).

With the increasing size of medical datasets, there has been an increasing interest in using machine learning techniques to predict patient survival times. Most recent studies have used artificial neural networks (ANNs) to predict survival time using the binary cross-entropy (BCE) loss function (Ren et al., 2019; Zadeh & Schmid, 2020). However, the importance of dealing with censored data is often overlooked. In survival analysis, censoring happens when the event (death) time is not observed (see Section 1.2.1) (Miller Jr, 2011). Moreover, using the BCE loss function in survival analysis has been shown to lead to a heavy bias in the predicted survival times (Gensheimer & Narasimhan, 2019; Zadeh & Schmid, 2020). This is especially problematic in health care, where researchers want to predict survival curves per patient. It is therefore important to explore alternative (better) loss functions that lead to less bias in predicted survival times.

Over the last decades, numerous amounts of statistical methods for analyzing survival data have

been developed. Most of these statistical methods, such as Cox proportional hazard (CPH), focus on continuous-time models rather than discrete-time models (Kvamme & Borgan, 2019b). In continuous survival time data, death can occur at any given time. However, in discrete survival data, survival is categorized into discrete time points. The most used survival model is the CPH model (Gensheimer & Narasimhan, 2019). However, it assumes proportional hazards: if a patient's risk of death at some time point is twice as high as that of another patient, then at later time points the risk of death remains twice as high. For models consisting of hundreds of patients, the violation of proportional hazards was, in past studies, not always possible to demonstrate (Gensheimer & Narasimhan, 2019). Moreover, the proportional hazards assumption was found to not be a realistic assumption for clinical situations (Gensheimer & Narasimhan, 2019). In most clinical situations, not all covariates have a constant effect on the risk of death over time. For example, the progression of a disease such as glioblastoma affects the survival of patients. The effects of a covariate such as tumor volume or age on the survival time of a patient may grow more pronounced over time and may vary between patients.

Since the rapid development of ANNs in the last couple of years, new methods that naturally deal with non-proportional hazards have been developed to predict survival over time (Gensheimer & Narasimhan, 2019; Kvamme & Borgan, 2019b). Previous studies have adopted the CPH model to neural networks. However, in the CPH model the partial likelihood for each individual also depends on the output of all individuals with longer survival times (Gensheimer & Narasimhan, 2019). This raises an issue with the use of stochastic gradient descent (SGD), because with SGD only a small number of patients are visible to the model at a time (Gensheimer & Narasimhan, 2019). A possible solution to this is using the entire dataset for each gradient step, but this slows down convergence and could result in the model getting stuck in a local minimum (Gensheimer & Narasimhan, 2019). Another, better, approach is to use a discrete time model which is fully parametric. In a discrete time model, for each discrete time point the probability of the event (death) happening given that the patient has survived at least to the start of the time point is estimated (Gensheimer & Narasimhan, 2019; Kvamme & Borgan, 2019b). Moreover, discrete time models are able to deal with non-proportional hazards (Gensheimer & Narasimhan, 2019; Kvamme & Borgan, 2019b). This is especially important because of the increasing sizes of medical datasets. In this study, a relatively large dataset is used of over a thousand patients.

To the best of our knowledge, Lee, Zame, Yoon and van der Schaar (2018) created the first ANN (DeepHit) which was capable of predicting survival using a discrete-time model. Gensheimer and Narasimhan (2019) also created a discrete-time survival model called Nnet-survival. Nnet-survival is also known as Logistic-Hazard and, in short, optimizes the survival likelihood. Instead of the widely used BCE loss function, the authors used a custom negative log-likelihood (NLL) loss function to optimize their model. A study conducted by Zadeh and Schmid (2020) found that the NLL loss function results in better calibrated survival prediction. Moreover, the model can naturally deal with non-proportional hazards and performs well on large datasets (Gensheimer & Narasimhan, 2019). The authors of DeepHit and Logistic-Hazard tested the models on various datasets (Lee et al., 2018; Gensheimer & Narasimhan, 2019; Kvamme & Borgan, 2019b). However, to the best of our knowledge, neither the DeepHit nor the Logistic-Hazard method have ever been used to predict survival using 3D MR images of glioblastoma patients. Kvamme and Borgan (2019b) explored discrete-time survival prediction with ANNs. The study compared aforementioned methods DeepHit and Logistic-Hazard and found Logistic-Hazard outperforming DeepHit (Gensheimer & Narasimhan, 2019; Lee et al., 2018; Kvamme & Borgan, 2019b). The authors of the study, Kvamme and Borgan, created a Python library named PyCox that contains implementations of various survival models, including the Logistic-Hazard method (Haavard Kvamme, 2020).

Furthermore, previous studies have explored the correlation between medical images and survival predominantly using radiomics (Aerts et al., 2014; Lao et al., 2017; Nie et al., 2016). Radiomics is a process that extracts quantitative features from radiographic images to construct predictive models (Lao et al., 2017). Radiomics models often have thousands of features. However, most

extracted features such as tumor shape, intensity and texture are handcrafted according to previous medical experiences (Aerts et al., 2014). Furthermore, radiomic features are often shallow and low-level image features. The models are therefore limited and cannot fully characterize the image heterogeneity of 3D MR brain images (Lao et al., 2017). Therefore, it is necessary to include deeper, high-level features. Recently in the medical field, deep learning techniques have been used in classification and regression tasks. Convolutional neural networks (CNN) have been widely used for medical image segmentation and classification (S. Liu et al., 2017; Zhao & Jia, 2016; Pereira, Pinto, Alves & Silva, 2016). CNNs are artificial neural networks with multiple convolutional layers. They use a filtering technique which convolves an image with a kernel, subsequently creating a new image (Guo et al., 2016). They are capable of extracting high level features and, unlike in handcrafted radiomics, feature extraction and correlation are intertwined (Gu et al., 2018; Guo et al., 2016). CNNs can learn various filters which can be used to extract important features that are needed for making accurate predictions (Guo et al., 2016). Instead of extracting quantitative features like radiomics, a CNN builds a deep structure that is capable of discovering high-level features that can better characterize brain tumors (Nie et al., 2016). For these reasons, there has been an increase in popularity of 3D CNNs in the medical domain (Baid et al., 2018). In 2015, ResNet popularized the use of skip-connections and this concept was used by most of the succeeding networks (Khan, Sohail, Zahoora & Qureshi, 2020).

In this study we apply a Logistic-Hazard 3D residual neural network (3D ResNet) model to the PICTURE dataset using the (modified) PyCox library. Our goal is to evaluate the performance of the model on multi-sequence 3D MR images of glioblastoma patients. We use a, for medical standards, large dataset consisting of 1132 patients from 12 different hospitals to ensure generalizability. This study uses a 3D ResNet trained on pre-operative multi-sequence 3D MR images to predict the survival curve of glioblastoma patients. Results will be evaluated using the inverse probability of censoring weighted Brier score (IPCW Brier score) and the time-dependent concordance index evaluated at the event times (C-index). Both metrics measure the accuracy of probabilistic predictions by evaluating a model's calibration and discriminative power (Kvamme & Borgan, 2019a). We hypothesize that the Logistic-Hazard method can be used to create a 3D ResNet model that is trained on multi-sequence 3D brain MR images of glioblastoma patients to predict a survival curve per patient. Furthermore, we hypothesize that a Logistic-Hazard 3D ResNet model trained on multi-sequence 3D brain MR images outperforms a Fully Connected Neural Network (FCNN) trained on clinical variables such as age, gender and tumor volume.

1.1 Related Work

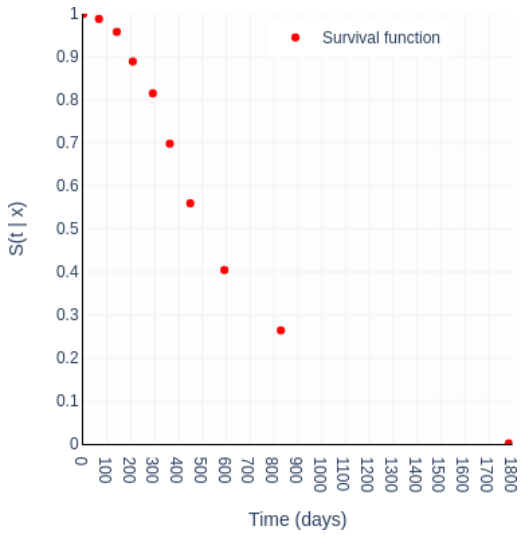
Previous studies have used deep learning techniques to predict OS time rather than survival over time (survival curves). Nie et al. (2016) used 3D deep-feature extraction to classify survival time of patients with glioblastoma in two classes: short OS and long OS (Nie et al., 2016). Chato and Latifi (2017) extracted features from MRI sequences and developed a prediction model that used the extracted features to also predict OS time (Chato & Latifi, 2017). R. Liu et al. (2016) also extracted features from MRI sequences and were able to classify survival time of patients with glioblastoma in short OS and long OS.

However, the studies lacked generalizability due to the small datasets that were used (Nie et al., 2016; Chato & Latifi, 2017). The studies conducted by Nie et al. (2016), Chato and Latifi (2017) and R. Liu et al. (2016) consisted of 68, 163 and 22 patients respectively, all resulting in poor generalizability. Currently, the largest public glioblastoma dataset is the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) 2018 dataset consisting of 542 patients. However, only 346 glioblastoma patients had their OS time documented, making the data set less useful for survival prediction (Menze et al., 2014; Bakas et al., 2018). Another limitation of the dataset is that it contains no censored patients, which is not a realistic situation. Baid et al. (2018) were able to classify survival time of patients with glioblastoma in short OS and long OS, but noted

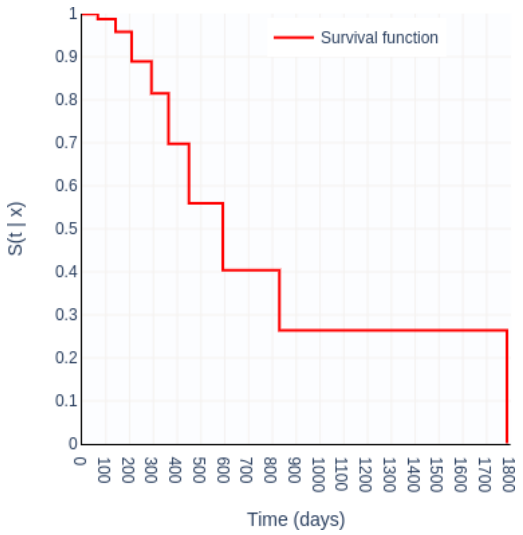
the necessity of large data sets to make accurate survival predictions (Baid et al., 2018).

There is little literature that analyzed brain tumor outcome by conditional probability of survival (Lin et al., 2003). Previous studies that created models to predict the survival rate of patients with glioblastoma focused on predicting OS time (R. Liu et al., 2016; Ahmed, Hall, Goldgof, Liu & Gatenby, 2017; Baid et al., 2018; Feng et al., 2020). As mentioned earlier, R. Liu et al. (2016), Ahmed et al. (2017) and Baid et al. (2018) all predicted short term or long term survival using MR images of glioblastoma patients. A disadvantage of such approaches is the loss of information about censored patients and the models would need to be re-trained in order to make predictions for different time points (Gensheimer & Narasimhan, 2019). Moreover, the models did not provide any insight in how much the survival probability declined after a certain amount of months for any given patient.

In this study a survival curve is predicted per patient using a Logistic-Hazard 3D ResNet model. As depicted in Figure 1.1a, the survival curve is defined as a series of declining horizontal steps which approaches the true survival function for a given patient (Goel, Khanna & Kishore, 2010). The survival curve is constructed based on the survival predictions of every discrete time point as depicted in Figure 1.1b. It allows us to show the survival probability over time, as opposed to the widely used OS time which is measured in days or low-survival and high-survival categorical groups. A survival curve shows the probability of survival at a certain time. Thus, it can offer an important estimate to predict survival rates for individual glioblastoma patients. It provides clinicians a more fine grained look at how the survival probability declines over time for a given patient.



(a) For every discrete time point a prediction is made.



(b) Based on the predictions per discrete time point, a step-wise survival curve can be constructed.

Figure 1.1: Example survival curve. The y-axis shows the predicted survival probability.

1.2 Survival Analysis

Time-to-event prediction is a sub field of survival analysis and is concerned with when in the future an event will occur. In this study we are concerned in the event *death* due to glioblastoma. Our goal is to predict when in the future death occurs for a given glioblastoma patient. Our survival data can be partitioned into a number of states at each point in time. When time passes, patients

move between these states. In our study we have two states: whether death occurred or not. In other words, we have a dataset that contains whether a glioblastoma patient passed away and if so when the event happened (in days). We also have patients in the dataset where the event *death* was not observed, which we will discuss next.

1.2.1 Censoring of survival data

When conducting survival analysis, it is important to keep censoring of survival data into account. Censoring occurs when we have information about a patient's death time, but we do not know when exactly the patient died (Jenkins, 2005). In survival analysis there are two types of censoring: right-censoring and left-censoring. Right censored data occurs when the patient leaves the study before the event, death, is observed, the patient is lost to follow-up during the study or a patient withdraws from the study (Goel et al., 2010; Jenkins, 2005). Uncensored patients are patients of whom the death time is observed (known). In Figure 1.2 we can see an example of the difference between right censored patients (white dots) and uncensored patients (black dots). For some patients the event (death) is not observed at the end of the study, these patients are thus right censored. Left-censoring technically occurs when the event, death, has already happened before the patient is included, which is not possible in our case. We will be using the term censoring as short-hand for right censoring.

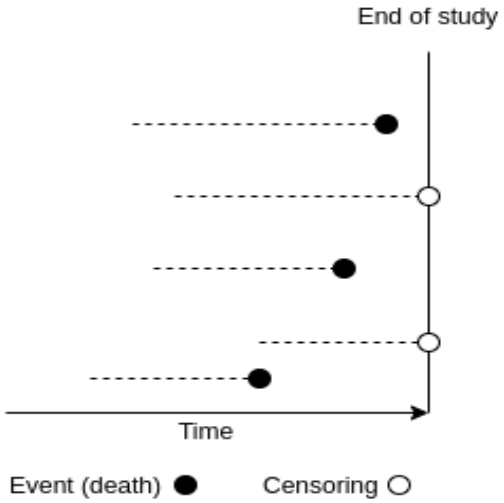


Figure 1.2: Example of right censoring, time is calendar time.

Chapter 2

Methods

2.1 Overview

We built two machine learning models: one trained on clinical data and one trained on 3D MRI data. The model trained on the clinical dataset was developed in order to compare a model trained on categorical and numeric data with a model trained on 3D MRI data. Having these two models allowed us to evaluate the performance of the Logistic-Hazard method on 3D MRI data. First, we will describe the clinical dataset and the FCNN model that was trained on it. Second, we will describe the 3D ResNet model that was trained using the 3D MRI data.

2.2 Dataset

Data from the PICTURE project was used (Visser et al., 2019; R. S. Eijgelaar et al., 2018; R. Eijgelaar et al., 2019). The initial clinical dataset consisted of 1451 patients with a mean age of 61.3 years (±12.6). The dataset consisted of 484 females, 789 males and 178 with an unknown gender. The OS time was available for 1239 patients and the mean survival was 423.8 (±382.9) days. The data was collected in 12 multi-national hospitals: Northwest Clinics, Alkmaar, Netherlands (ALK); Amsterdam University Medical Centers, location VU medical center, Netherlands (AMS); University Medical Center Groningen, Netherlands (GRO); Medical Center Haaglanden, the Hague, Netherlands (HAG); Humanitas Research Hospital, Milano, Italy (MIL); Hôpital Lariboisière, Paris, France (PAR); University of California San Francisco Medical Center, US (SFR); Medical Center Slotervaart, Amsterdam, Netherlands (SLO); St Elisabeth Hospital, Tilburg, Netherlands (TIL); University Medical Center Utrecht, Netherlands (UTR); Medical University Vienna, Austria (VIE); and Isala hospital, Zwolle, Netherlands (ZWO).

2.2.1 Exclusion

First, patients were excluded based on a number of exclusion criteria: gliosarcoma, no clinical variables, no MRI, no usable MRI, infratentorial glioblastoma, no Contrast Enhancing lesion visible on the pre-operative, not included in study, pediatric, the surgery performed within the inclusion period was not the primary surgery performed on this patient, secondary glioblastoma and surgery elsewhere. The study sample flowchart is depicted in Figure 2.2. After the exclusion of these patients, the study sample consisted of 1036 patients. The hospital distribution of the study sample can be seen in Figure 2.1. The filtered clinical dataset consisted of 1036 patients with a mean age of 62.2 years (±12.3) and a mean survival of 408.3 days (±360.8). The dataset consisted of 387 females, 646 males and 3 with an unknown gender.

From the filtered clinical dataset, 1036 patients had pre-operative MRI data. Of the 1036 patients, 1016 patients had a tumor mask and at least one valid MRI sequence as shown in Figure 2.2. Be-

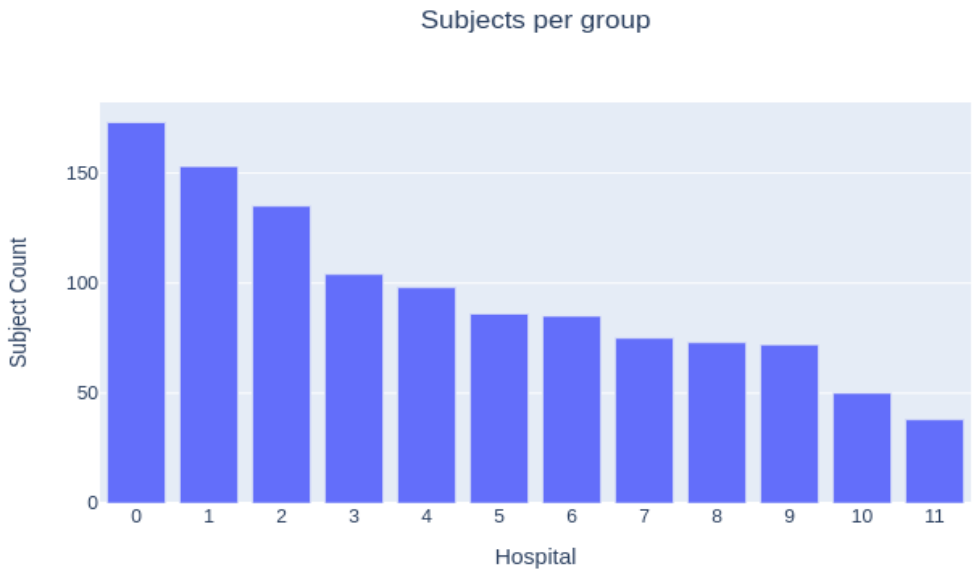


Figure 2.1: Number of patients per hospital after filtering

cause we extracted the tumor regions and created bounding boxes around them, the availability of a tumor mask was a requirement to include patients in this study. The distribution of available pre-operative sequences was as follows: 1016 patients had a T1c (contrast-enhanced T1) scan, 578 had a T1w (T1 weighted image) scan, 504 had a FLAIR (fluid-attenuated inversion-recovery) scan and 450 had a T2w (T2 weighted image) scan. The survival time and event (death) of every patient was known.

The final study thus consisted of 1016 patients with a mean age of 62.1 years (±12.3). The sample consisted of 380 females and 636 males. The mean survival was 410.7 (±361.3) days. The number of missing sequences per hospital are depicted in Table 2.1.

Hospital	FLAIR	T1c	T1w	T2w	T1w, T2w	T1w, FLAIR	T2w, FLAIR	T1w, T2w,FLAIR	Total
0	4	0	1	11	0	0	124	6	146
1	0	0	0	3	2	0	9	130	144
2	0	0	67	1	0	1	1	2	72
3	0	0	0	0	1	13	0	84	98
4	1	0	1	0	0	0	1	3	6
5	19	0	3	1	0	1	1	3	28
6	0	0	5	0	0	3	1	17	26
7	0	0	0	66	2	0	0	0	68
8	0	0	0	0	0	0	1	21	22
9	0	0	1	1	1	0	1	16	20
10	0	0	0	0	0	0	0	5	5
11	0	0	0	0	2	0	0	29	31
Total	24	0	78	83	8	18	139	316	666

Table 2.1: Missing Scans in MRI Dataset

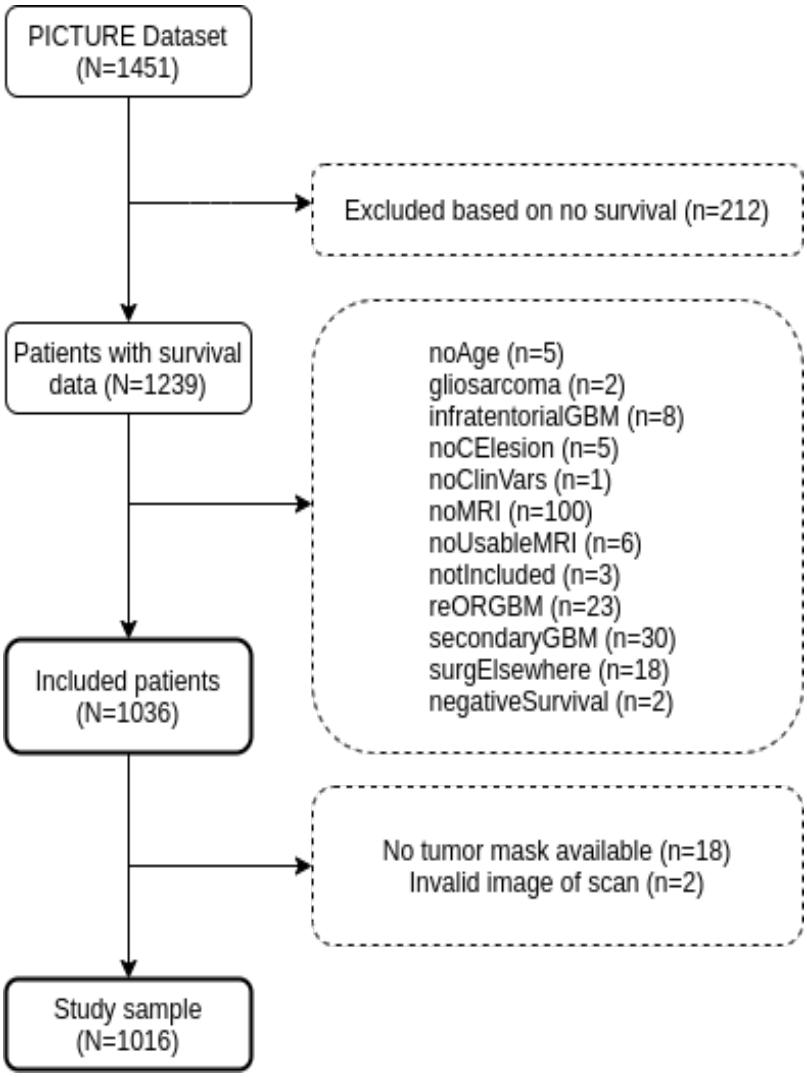


Figure 2.2: Exclusion flow chart

2.3 Preprocessing

2.3.1 Discretization

For the reasons described in Chapter 1, we chose a discrete time method to predict survival times. Logistic-Hazard is a discrete time method and thus required the survival data to be in discrete time. For this reason, we categorized the survival data into fixed discrete time points. Kvamme and Borgan (2019b) studied several discretization methods: one was based on making an equidistant grid and one was to make a grid based on the distribution of the event times. We have used the latter, because it was found to achieve better results on relatively small datasets (Kvamme & Borgan, 2019b). A general trend of event times is obtained by estimating the survival function using the Kaplan-Meier estimator (Kvamme & Borgan, 2019b). This survival function was then used to make a grid from the quantiles as depicted in Figure 2.3. Logistic-Hazard assumes that all events and censoring times occur at the discrete times, so all event times (deaths) that occur in a time interval (between two discrete time points) were moved to the end of that interval. Censored patients were moved to the end of a previous discrete time interval. We discretized the survival data into 10 time points, because smaller discretization grids reduced the number of parameters in a neural network and thus were better for our relatively small training set. The cut off points were set to: $[0, 66, 141, 208, 292, 363, 449, 592, 829, 1785]$ days. The first cut off point, 0 days, consisted of patients who were censored between 0 and 66 days. The second cut off point, 66 days, consisted of patients who died between 0 and 66 days and censored patients that were censored between 66 and 141 days et cetera.

We could illustrate this the best with an example. Suppose that a patient has died at 150 days. He or she thus died during the discrete time interval that ranges from 141 to 208 days. Because all events occur at the discrete times, we know that the patient was alive at 141 days and dead at the following discrete time point of 208 days. Thus, the patients death is observed at 208 days. However, suppose that the patient was censored at 150 days. Then we know that the patient was alive at 141 days but we do not know if the patient was still alive at the following discrete time point of 208 days. Because of this, the patients survival time was then categorized to be observed at 141 days. This means that for every patient we have either specified at which discrete time point their death or censoring was observed.

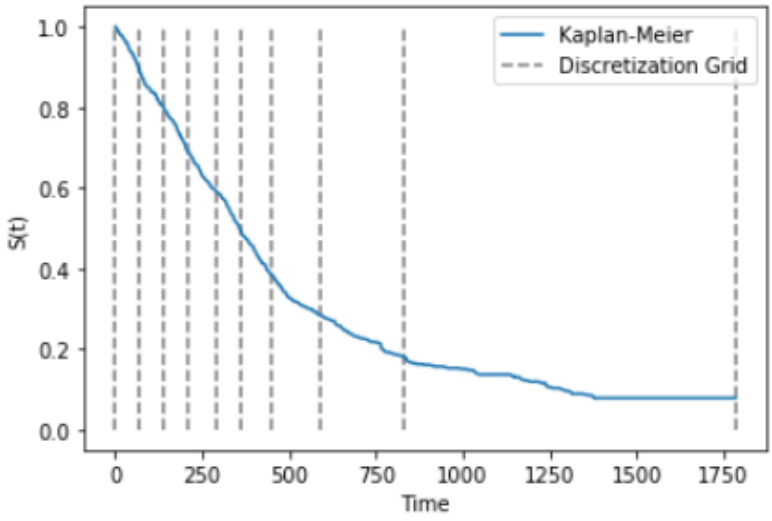


Figure 2.3: Discretization of survival time using quantiles

2.3.2 Clinical dataset cleaning

To make the clinical variables suitable for use with an ANN, we converted the (string) data to numerical data. The variables tumor side (Left/Right), gender (Male/Female), chemotherapy (Yes/No) and event (True/False) were all converted to binary data (0/1). Continuous variables were standardized to a 0 to 1 range. A snippet of the resulting data frame is depicted in Table 2.2.

age	ENTvolML	ENTside	GenderV2	KPSpre	Chemo	SurgeryExtend	surv	DeathObserved
57.1	0.612	1	0	8.0	1	1	734.0	1
55.4	76.784	0	0	6.0	0	1	679.0	1

Table 2.2: Snippet of the clinical dataset (not real data)

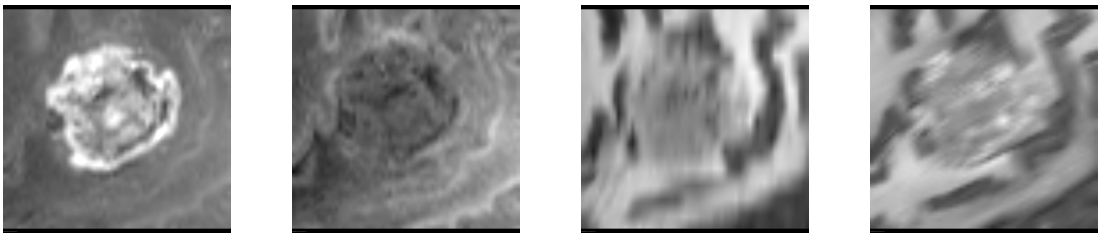
The data frame was then split into train, test and validation data frames. To make the data suitable for use with PyTorch, the data frames were converted to Numpy arrays. ANNs often perform worse if the individual features do not look like standard normally distributed data. Because of this, the continuous variables of the train, test and validation dataset were scaled using the following function:

$$z = (x - u)/s \tag{2.1}$$

Where u is the mean of the samples, and s is the standard deviation of the samples.

2.3.3 MRI sequences

Display of detailed 3D brain images is possible using MRI. There are several MRI sequences such as T1c, T1w and FLAIR. It is known that none of these sequences are able to show the entire extent of a tumor (Banerjee, Mitra, Shankar & Hayashi, 2016). Tumors can have different characteristics in different patients. Radiologists typically describe the tumor core using a T1c MRI, because the tumor boundary is more visible due to the contrast between gray and white matter (Banerjee et al., 2016). The T2w MRI provides better contrast between brain tissue and cerebrospinal fluid (CSF) and is often preferred for describing the edema region (Banerjee et al., 2016). FLAIR can describe both tumor and edema regions, but the edema boundary becomes fuzzy (Banerjee et al., 2016). The differences between the MRI sequences highlights the importance of including them as input for our 3D ResNet model. The different MRI sequences of one (random) patient are depicted in Figure 2.4. These figures are slices of 3D MR images of the whole brain.



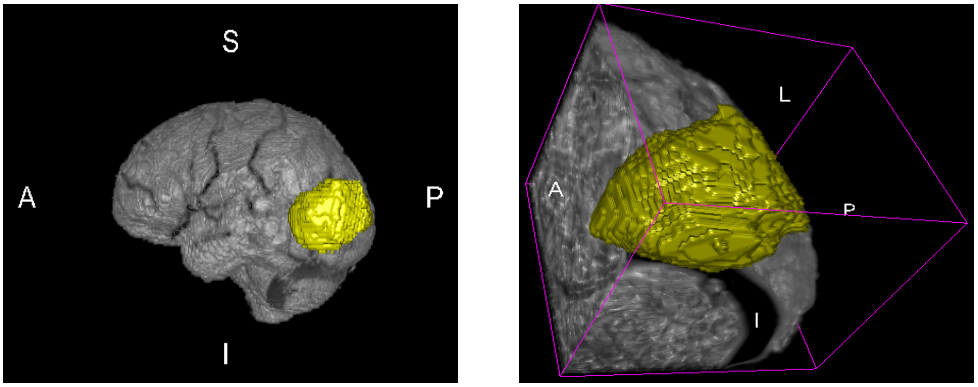
(a) T1c **(b)** T1w **(c)** FLAIR **(d)** T2w

Figure 2.4: MRI sequences

2.3.4 Bounding Box

Due to computational constraints, we trained a 3D ResNet using only the tumor regions as input. This was done by creating a 3D bounding box with the dimensions 64 by 64 by 64 around the

tumor. We chose this size, because some tumors were big and this size was able to capture the tumor of all patients. SciPy's center of mass function calculates the center of mass of the values of an array at labels. Because our tumor mask had binary labels, we were able to use the function to find the tumors center of mass. The output was a set of coordinates which we then used to calculate the edges of the bounding box. However, if the tumor was located too close to the edge of the brain, it could not be centered and thus (zero) padding was used to ensure a centered tumor in the bounding box. The brain of a random patient with the tumor highlighted (a) and the corresponding bounding box (b) are depicted in Figure 2.5.



(a) Tumor location **(b)** T1c bounding box

Figure 2.5: Tumor bounding box of a random patient. Tumor is shown in yellow.

2.3.5 Intensity Normalization

Intensity variations in the testing dataset are hard to capture in the training dataset. These variations stem from the fact that we used MR images of a variety of hospitals which in turn used different MRI scanners and sequence configurations. Because of this, the images often did not have similar intensity ranges. A lack of a standard intensity scale hurts a ResNet's generalizability. Therefore, it was important to limit the spread in the image data using intensity normalization. Intensity normalization is a procedure that maps all image sequences (or modalities) to a standard scale in order to make the same type of sequence of different patients fall within a similar range. A study conducted by (Reinhold, Dewey, Carass & Prince, 2019) evaluated various normalization methods for 3D MR images such as Z-score, Fuzzy C-Means, Gaussian mixture model and WhiteStripe. The code of the normalization methods used in the study were put on GitHub by the authors. Code of the intensity normalization algorithms can be found at https://github.com/jcreinhold/intensity-normalization. We used the code to apply and compare various intensity normalization methods such as Z-score and WhiteStripe. Fuzzy C-Means required a T1w sequence to be available, which we did not have for all patients. The original code can be used to output normalized images as file (Reinhold et al., 2019). However, we wanted to implement normalization on the fly when loading a batch from the dataset to train a neural network. Because of this, we only adapted the functions that are responsible of normalizing an input image and outputting a NumPy array.

Z-score

As depicted in Equation 2.2, Z-score normalization was achieved by determining the mean μ and the standard deviation σ of the intensities of an input modality using the brain mask B (Reinhold et al., 2019). We subtracted the mean from every intensity value and then divided it by the standard deviation to get the Z-score normalized image I as shown in Equation 2.3 (Reinhold et

al., 2019). We chose to use Z-score normalization, because it produced better results than the WhiteStripe normalization. Moreover, Z-score normalization was faster.

$$\mu = \frac{1}{|B|} \sum_{\mathbf{b} \in B} I(\mathbf{b}) \quad \text{and} \quad \sigma = \sqrt{\frac{\sum_{\mathbf{b} \in B} (I(\mathbf{b}) - \mu)^2}{|B| - 1}} \tag{2.2}$$

$$I_{\text{z-score}}(\mathbf{x}) = \frac{I(\mathbf{x}) - \mu}{\sigma}. \tag{2.3}$$

WhiteStripe

WhiteStripe intensity normalization performs a Z-score normalization based on the intensity values of normal appearing white matter (NAWM) (Reinhold et al., 2019). The NAWM is found by smoothing the histogram of the input image and selecting the highest intensity peak (Reinhold et al., 2019). However, we observed no benefit of using WhiteStripe normalization over Z-score normalization.

2.4 PyTorch Dataset and DataLoader

The data loading process is depicted in Figure 2.6. For the MRI data, we created a dataset that was essentially a list of Python objects that contained, for every patient, the path of every MRI sequence, the survival time in days, whether an event happened (0 or 1) and in which discrete time point (group) the patient belonged. An example sample in the dataset had the following format:

```
{'id': 'BMIAXNAT_S34534',
 'T2w': '/home/data/preoperative/BMIAXNAT_S34534/T2w/BMIAXNAT_E18545_T2w.nii.gz',
 'T1c': '/home/data/preoperative/BMIAXNAT_S34534/T1c/BMIAXNAT_E18545_T1c.nii.gz',
 'T1w': '/home/data/preoperative/BMIAXNAT_S34534/T1w/BMIAXNAT_E18545_T1w.nii.gz',
 'ENT': '/home/data/preoperative/BMIAXNAT_S34534/ENT/BMIAXNAT_E18545_ENT.nii.gz',
 'FLR': '/home/data/preoperative/BMIAXNAT_S34534/FLR/BMIAXNAT_E18545_FLR.nii.gz',
 'surv': 29.0,
 'event': 0,
 'group': 0}

Note: the numbers are changed for privacy reasons
```

The clinical dataset was an array of arrays, one for each patient. Each array had a size of seven, one entry for each of the seven variables. Each patient had a tuple as target of the format (y_1, y_2). Where y_1 was the discrete time point (group) the patient belonged to and y_2 the binary event outcome (death).

To convert each sample of the 3D MRI dataset to something PyTorch could handle, we needed the Dataset class of PyTorch. The Dataset class acted as a bridge between our dataset and a PyTorch tensor. The Dataset class applied transformations, augmentations and intensity normalization to the 3D MRI inputs and was able to output tensors containing the data that was used as training, validation or testing data. We made the Dataset class dynamic: the normalization technique could be specified as well as the phase (training, validation or testing) and bounding box size.

At the heart of every PyTorch model lies the DataLoader object. The DataLoader is responsible for loading batches of the PyTorch MRI or clinical dataset to be used as training, validation or testing data.

Because all the clinical data was preprocessed beforehand and converted to the right format for use with the PyTorch DataLoader class, we did not need to create a custom PyTorch Dataset

class for it.

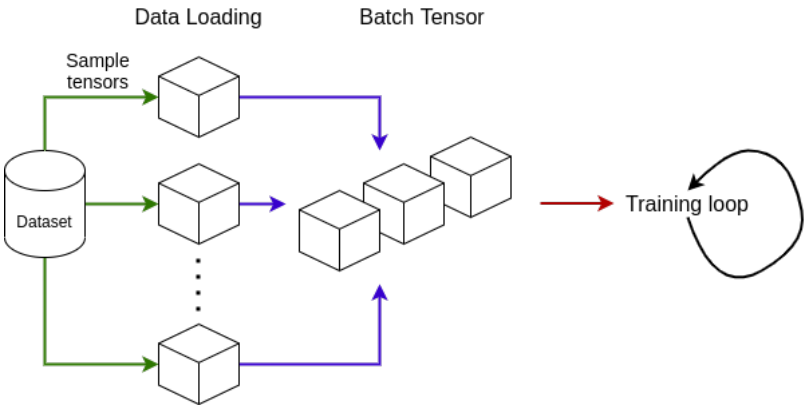


Figure 2.6: Data Loader overview. Green arrows show the role of the PyTorch Dataset class. The purple and red arrows show the role of the DataLoader class. Adapted from: (Antiga, 2020)

2.5 Data split

The proportions of the data split are depicted in Figure 2.7. The training dataset consisted of 80% of the samples, the validation dataset consisted of 10% of the samples and the test dataset consisted of 20% of the samples. Prior to performing the data split, we randomly shuffled the dataset. However, we ensured that the training dataset, validation dataset and testing dataset had the same patients in both the model trained on the 3D MR images, as well as the one trained on the clinical dataset. We also set a fixed random seed to ensure the shuffling resulted in the same datasets for every training iteration.

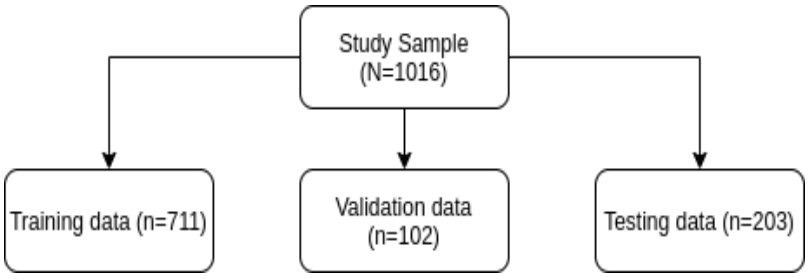


Figure 2.7: Data split proportions

2.6 Loss Function (Logistic Hazard)

A loss function is a function that computes a single numerical value that the learning process will attempt to minimize using gradient descent (Stevens, Antiga & Viehmann, 2020). For this

study, we used the Logistic-Hazard loss function which, to the best of our knowledge, originally was developed by Gensheimer and Narasimhan (2019) and recently slightly modified by Kvamme and Borgan (2019b).

The Logistic-Hazard loss function represents the NLL of our model. The output of our FCNN and 3D ResNet model is an n-dimensional vector where n is the number of discrete time points, which, in this study, was set to 10. Each element of the vector represents the conditional log-odds (logit) of surviving up to that time (Gensheimer & Narasimhan, 2019).

Based on whether the patient is censored or not, the loss is calculated differently. We will describe how the loss is calculated for an uncensored and a censored patient.

2.6.1 Uncensored patients

To compute the loss for patients that were not censored, first we applied binary cross entropy (BCE) with logits which combines a sigmoid layer and the BCE loss in one single class. We will be using an example patient to describe what the loss is and how it is calculated. As input, we have the estimates of our neural network (FCNN or 3D ResNet):

Log-odds: [-0.1154, -0.6996, 0.1101, -0.2162, -0.1485, 0.2301, -0.3744, -0.6813, -0.0525, 0.4338]

As target we have a vector of size 10 (one for each discrete time point) where each element is either a 0 if the patient was alive on that time point or a 1 if the patients death was observed at that time point:

BCE Target: [0, 0, 0, 0, 1, 0, 0, 0, 0, 0]

In this example the patient's death was observed at the discrete time point 292 days. This means the patient died between 208 and 292 days. The reasoning for this was described in Section 2.3.1. To get the conditional survival probabilities, BCE with logits first applies the following sigmoid function to the vector containing the log-odds:

$$S(x) = \frac{1}{1 + \exp(-x)} \tag{2.4}$$

Which results in:

Probabilities: [0.4712, 0.3319, 0.5275, 0.4462, 0.4629, 0.5573, 0.4075, 0.3360, 0.4869, 0.6068]

Then the individual loss (negative log likelihood) is calculated using the following function:

$$\text{Loss} = -(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)) \tag{2.5}$$

Where $\hat{y}_i$ is the i-th scalar value in the probabilities vector and y_i is the corresponding target value. We take the negative of the loss (negative log likelihood), because we use gradient descent to minimize it. Thus, we want high loss values to be associated with bad predictions and low loss values to be associated with accurate predictions. This is the case when we take the negative of the log likelihood. The loss for the example patient for the first discrete time point (index 0), where the event (death) is not observed, is:

$$\text{Loss} = -(0 \cdot \log 0.4712 + (1 - 0) \cdot \log (1 - 0.4712) = 0.6371$$

And for the fifth discrete time point (index 4) where the event (death) is observed:

$$\text{Loss} = -(1 \cdot \log 0.4629 + (1 - 1) \cdot \log (1 - 0.4629) = 0.7702$$

The output of the BCE function is thus a vector of size 10 where each element is the negative log likelihood of a discrete time point:

Negative log likelihood: [0.6371, 0.4033, 0.7497, 0.5909, 0.7702, 0.8148, 0.5234, 0.4094, 0.6672, 0.9334]

The patient's death was observed at the fourth discrete time point (index 3). This means that the total loss for this patient is the sum of the negative log likelihoods of the first 4 indices of above vector:

$$\text{Total loss: } 0.6371 + 0.4033 + 0.7497 + 0.5909 = 2.381$$

2.6.2 Censored patients

The censoring time of all censored patients was known and this was transformed into a vector. This vector represents the discrete time point during which failure (censoring) was observed (Gensheimer & Narasimhan, 2019; Kvamme & Borgan, 2019b). Because there are 10 different discrete time points, there are 10 different possibilities as to when a patient's censoring was observed.

The loss of every censored patient is calculated in a similar manner as described above. However, because the death time of a censored patient is not known, the BCE target is a vector that consists of 10 zeros indicating that the event is not observed at any of the 10 discrete time points:

BCE Target: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

The model' outputs are again first converted to probabilities. The loss is then calculated using Equation 2.5. For uncensored patients, the calculation of the total loss was based on when the patients' death was observed. However, since death was not observed for censored patients, the censoring time is used to determine what the total loss is of a patient. If, for example, a patient's censoring was observed at the third discrete time point (index 2) (the patient was censored between 141 and 208 days), the loss is:

Negative log likelihood: [1.1797, 0.9888, 0.6185, 0.7889, 0.5915, 0.5337, 0.9623, 0.7588, 0.7368, 0.9473]

Because we know that the patient's censoring was observed at the third discrete time point (index 2), the total loss for this patient is the sum of the negative log likelihood of the first 3 indices of above vector:

$$\text{Total loss: } 1.1797 + 0.9888 + 0.6185 = 2.787$$

2.6.3 Total model loss

We calculate the loss for every single patient as described above, sum all the patient losses, take the mean and then the model tries to minimize this mean loss using gradient descent.

2.7 Overfitting and Regularization

One of the most common problems while training an ANN is overfitting. Training a ANN that can generalize well to unseen data is often challenging. When training a model we want to teach it to recognize the general properties of the classes we are interested in. When our model starts to learn specific properties of the training data set, overfitting occurs. For example, in our case, overfitting means that our 3D ResNet model recognizes specific images in our training data set instead of general patterns. As a result, the model starts to lose the ability to generalize to unseen data. This phenomena is often recognized by looking at the training and validation loss. While training a model, we want both the training loss and validation loss to decrease. When the training loss keeps declining and the validation loss stays the same or even increases, the model is overfitting.

2.8 Data Augmentation

There are several ways of improving the generalizability of an ANN. One of the most straightforward way is by increasing the training data set. However, we could not obtain more 3D MR images of glioblastoma patients. Instead, we increased the size of the training set with several data augmentation techniques.

2.8.1 Tumor Center

We added noise to the coordinates of the tumor center. After calculating the tumor center, we randomly moved the center in one of three directions (x, y, z) with 1 to 10 pixels. Because the voxel size was 1mm, this translated to an 1mm to 10mm offset.

2.8.2 Rotation

To make the trained models even more robust and thus reduce overfitting, we applied random rotation to the bounding boxes with a probability of 60% ($p = 0.6$). The bounding boxes were mirrored upside down ($p = 0.3$) or left to right ($p = 0.3$).

2.8.3 Leave sequence out

To make the trained models more robust to missing sequences, we randomly dropped a sequence during training. All patients had a T1c scan available, but not all patients had the rest of the sequences available. Because of this, we never dropped a T1c scan during training. The probability of dropping one of the other sequences was set at 20% ($p = 0.2$).

2.9 Model Evaluation

A predictive model is often evaluated by looking at calibration and discrimination. Calibration shows how closely our predicted probabilities agree with the actual outcome. Discrimination measures our models ability to correctly separate outcome classes. Brier score and the C-index are two of the most common evaluation criteria for survival predictions (Gerds & Schumacher, 2006; Kvamme & Borgan, 2019a). In survival analysis, discriminative power is often evaluated using the C-index while calibration is measured using the IPCW Brier score (Kvamme & Borgan, 2019a). A high C-index (close to one) means a model is able to accurately distinguish between patients who die early and patients who die later. A C-index of 0.5 corresponds to a model doing no better than random guessing. A small IPCW Brier score (close to zero) means a model is able to accurately match predicted and observed death rates (Kvamme & Borgan, 2019a).

2.10 Fully Connected Neural Network

We used a FCNN to predict survival times using clinical variables such as age, gender, tumor volume and chemotherapy (yes/no). Because of our relatively small dataset, to prevent overfitting, the neural network was kept relatively simple with a few hidden layers. We also used 80% dropout ($p = 0.8$) to reduce overfitting. The architecture of the FCNN is depicted in Table 2.3 and was implemented using PyTorch 1.5 (Paszke et al., 2019). As input the FCNN had the seven clinical variables mentioned earlier. As output the FCNN had ten features, one for each discrete time point. Every linear layer was followed by batch normalization and a Rectified Linear Unit (ReLU). Batch normalization speeds up the training process by standardizing the activation's of each input variable per mini-batch (Antiga, 2020). Furthermore, we used the Adam optimizer, which is an adaptive learning rate optimization algorithm (Antiga, 2020). We applied grid search to find the most optimal parameters for the FCNN.

Layer Name	FCNN
Linear_1	in_features=7, out_features=28
Linear_2	in_features=28, out_features=56
Linear_3	in_features=56, out_features=56
Linear_4	in_features=56, out_features=112
Linear_5	in_features=112, out_features=10
dropout (0.8)	

Table 2.3: FCNN Architecture.

2.11 3D ResNet-10

We used a 3D ResNet-10 model to predict survival times using 3D MR images. The network architecture and part of the code that was used was made by Hara, Kataoka and Satoh (2018). We adapted the network to our specific needs. Because of our relatively small dataset, we opted for the least deep ResNet model consisting of 10 layers. The network architecture is depicted in Table 2.4 and was implemented using PyTorch 1.5 (Paszke et al., 2019). Each convolutional layer was followed by batch normalization and a ReLU. The first convolutional layer had as input 4 channels (one for each MRI sequence) and as output 64 filters. Down sampling was performed by conv3_1, conv4_1 and conv5_1 with a stride of two. The dimension of the last fully-connected layer was again set to ten groups/categories because of our discretization of survival times into ten time points. We again used the Adam optimizer.

Layer Name	ResNet-10
conv1	$7 \times 7 \times 7, 64$, stride 1 (T), 2 (XY)
conv2	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 1$
conv3_x	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 1$
conv4_x	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 1$
conv5_x	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 1$
average pool, dropout (0.7), 10-d fc, sigmoid	

Table 2.4: ResNet-10 Architecture. Residual blocks are shown in brackets.

2.12 Feature Maps

Visualizing the features of our 3D ResNet model helped us to understand the learned features. Feature maps are able to show the intermediate representations of an input (3D MR image) in the model. They are generated by applying learned filters to an input image. As the image progresses through the nine 3D convolutional layers of our 3D ResNet model, the details from the images slowly disappear. They look like noise, but there is a pattern in those feature maps which human eyes cannot detect, but a neural network can. This is because the feature maps of the first 3D convolutional layers of our 3D ResNet model retain most of the information present in the input image.

We constructed feature maps based on a multi-sequence 3D MR image. But to show what the model has learned from each sequence, we also constructed feature maps based on one sequence. First, we will describe how we constructed the feature maps based on a whole 3D MR image. Second, we will describe how we constructed the feature maps based on one 3D MRI sequence (T1c).

2.12.1 Feature map based on a whole 3D MR image

After training the 3D ResNet model, we extracted the nine 3D convolutional layers as well as the learned weights of each of the layers. As depicted in Table 2.4, the first convolutional layer has an input of 4 channels (one for each MRI sequence) and an output of 64 filters:

```
Conv3d(4, 64, kernel_size=(7, 7, 7), stride=(1, 2, 2), padding=(3, 3, 3), bias=False)
```

Each filter consisted of 4 channels (one for each MRI sequence) of size [7, 7, 7]:

```
torch.Size([64, 4, 7, 7, 7])
```

By forwarding a multi-sequence (4-channel) 3D MR image through this 3D convolutional layer, dot products between the filters and the local regions of the input were performed. This resulted in a 3D feature map of size [64, 32, 32] for each of the 64 filters:

```
torch.Size([1, 64, 64, 32, 32])
torch.Size([B, C, D, H, W])
```

Where B stands for the batch size, which because we have only one input is 1 and the C stands again for the number of feature maps (one for each filter) which is 64. The D, H and W stand for the depth, height and width of the feature maps respectively. Each of the feature maps can then be visualized to highlight what part of the multi-sequence 3D MR image the 3D ResNet model has learned to recognize and deemed important for survival prediction.

2.12.2 Feature map based on one 3D MRI sequence

To visualize the learned features per MRI sequence, we had to take a more fine grained look at the weights of the first 3D convolutional layer. We first created a completely new 3D convolutional layer, because the 3D ResNet model expects 4 channels (MRI sequences) instead of 1. Thus, we created a layer that expects a one channel (one MRI sequence) input and kept the kernel size, stride and padding the same as the first 3D convolutional layer of our 3D ResNet model:

```
Conv3d(1, 64, kernel_size=(7, 7, 7), stride=(1, 2, 2), padding=(3, 3, 3), bias=False)
```

Then we extracted 1 channel of each of the 64 filters of the first trained 3D convolutional layer of the 3D ResNet model that corresponded to a sequence that we were interested in (eg. T1c). Instead of having 64 filters with each 4 channels, we then had 64 filters with each one channel:

```
torch.Size([64, 1, 7, 7, 7])
```

These filters were learned by the 3D ResNet model by looking at T1c sequences and thus captured the learned features of the T1c sequence. We then set the weights of our newly created 3D convolutional layer to the above extracted 64, 1 channel, 3D filters. We then passed a one sequence 3D MRI input through the new 3D convolutional layer:

```
torch.Size([1, 1, 64, 64, 64])
```

Which resulted in the 64 3D feature maps of the 64 T1c filters:

```
torch.Size([1, 64, 64, 32, 32])
```

Chapter 3

Results

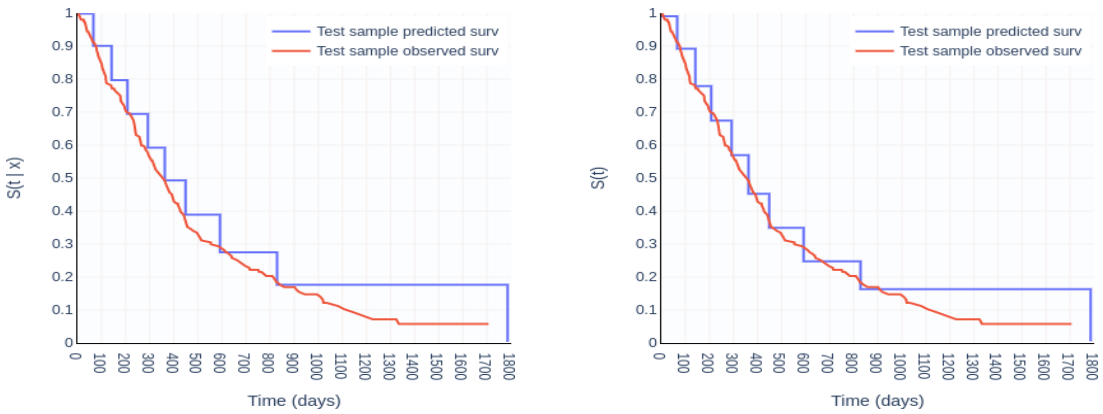
3.1 Overview

First, we will evaluate how well the two trained models are calibrated. This is done by comparing actual survival curves to predicted survival curves of the whole test datasets and by calculating and plotting the Brier score. Secondly, we will evaluate the discriminative power of both trained models by looking at the predicted survival probability for every discrete time point and by calculating the C-index. Third, we will look at the validation loss of both models to understand what factors contribute to the prediction uncertainty of both models. All survival curves shown in this chapter are averaged survival curves of either the whole test dataset or a discrete time point that consists of censored and uncensored patients of a certain time interval. Lastly, we visualize some feature maps to understand what the 3D ResNet model has learned.

3.2 Calibration

To evaluate the calibration of both the FCNN as well as the 3D ResNet model, we plotted and compared the actual survival curve vs. the predicted survival curves. A Kaplan-Meier survival curve was plotted to show actual survival. To obtain the predicted survival curves for the total test dataset, we averaged the predicted survival probability of every discrete time point. In general, a steep survival curve down means that the survival probability of the patients declines fast over time and a more gradual survival curve means that the survival probability of the patients declines more gradually over time. Because we discretized the survival time in this study, the survival curves are plotted as step functions. Here, a longer vertical line down means that the survival probability declined fast between two time points.

Figure 3.1 shows both the actual survival curve and predicted survival curves for the whole test dataset. Figure 3.2a shows the actual survival curve and predicted survival curve of the FCNN model. It can be seen that the calibration is good: the predicted survival curve is very close to the actual survival curve at all discrete time points. Figure 3.2b shows the actual survival curve and predicted survival curve of the 3D ResNet model. Here, it can be seen that the calibration is not as good as the FCNN model. However, the predicted survival curve of the 3D ResNet is close to the actual survival curve at 66, 141, 292 and 449 days.

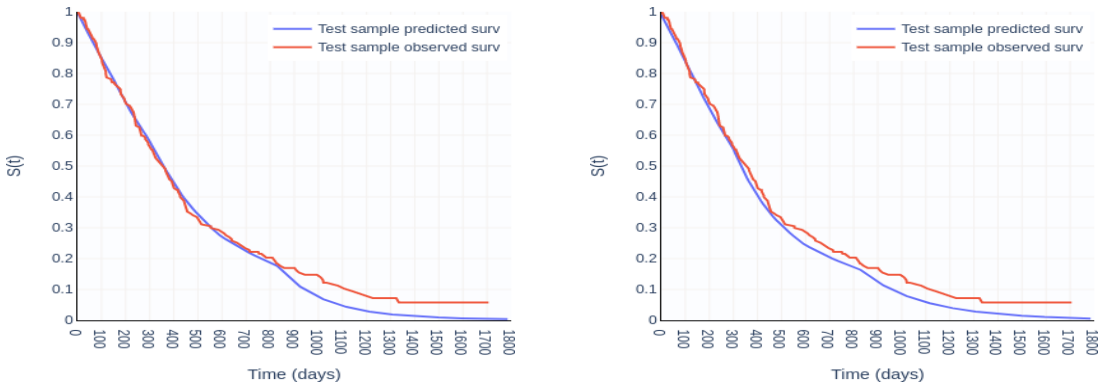


(a) Clinical Dataset model **(b)** 3D MRI Dataset model

Figure 3.1: Predicted survival curves

3.2.1 Interpolation

A limitation of discrete time methods is the lack of survival probabilities between discrete time intervals. One solution to this limitation is the use of interpolation schemes. By assuming constant hazard in each discrete time interval, we can interpolate the discrete time survival curve (Kvamme & Borgan, 2019b). In other words, this interpolation scheme assumes that deaths are spread evenly between two discrete time points. The sharp decline in survival probability at the end of the plots in Figure 3.1 is a result of our discretization. The interpolated survival curves are depicted in Figure 3.2. Up until 829 days the interpolated survival curve of the FCNN model is almost super imposed with the actual survival curve. The interpolated survival curve of the 3D ResNet model deviates more from the actual survival curve, especially after 500 days. The last two discretization points are 829 and 1785. Because censored patients are rounded down and event times are rounded up, we get an unnatural proportion of events at 1785 days. All patients who died between 829 and 1785 days are rounded up to 1785 days, while the patients that were censored between 829 and 1785 days are rounded down to 829 days. As a result, we have 37 patients in this time interval and 36 of them experienced an event (died). However, the true event proportion of individuals that survive past 1400 days is zero in the train dataset. The predictions past 829 days are thus less reliable than those before 829 days. Figure 3.3 shows the FCNN survival predictions of three (test dataset) patients that were censored after 1400 days. We can see that the survival predictions are not accurate. The model predicts a very low survival probability at the censoring time, although they were alive.



(a) Clinical Dataset model **(b)** 3D MRI Dataset model

Figure 3.2: Interpolated Predicted survival curves. The plots again show the FCNN outperforming the 3D ResNet. Both models have inaccurate survival estimates after 829 days.

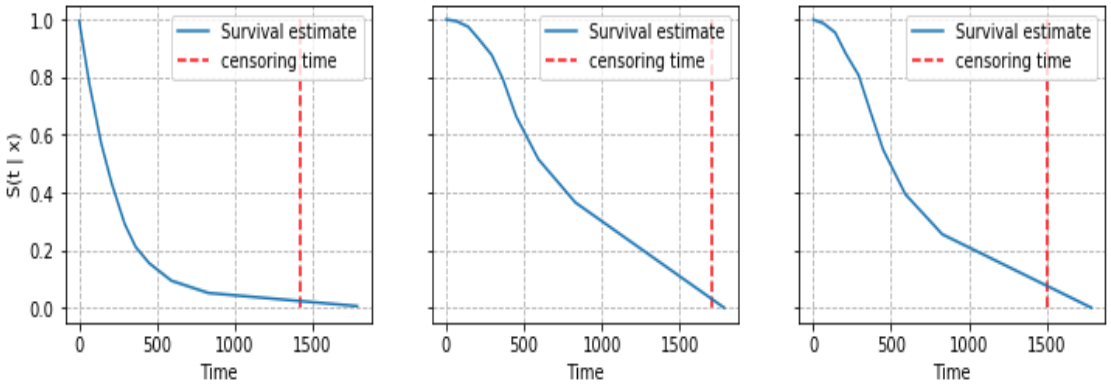


Figure 3.3: FCNN survival estimates of (test dataset) patients censored after 1400 days.

3.2.2 Brier score

Figure 3.4 shows the Brier score over time of both models based on the non-interpolated test dataset predictions. The 3D ResNet model has a higher Brier score at first indicating lesser performance predicting survival in that time interval, but this becomes slightly lower than the FCNN model between 800 and 900 days. This slightly better performance at this time interval is also visible in Figure 3.1. The sharp decline in Brier score at 1785 days is also a result of our discretization as described above. Having more discrete time points between 829 and 1785 days could improve the survival estimates, however we did not have enough patients in our dataset that died in that time interval to justify adding time points. Considering the median survival time of 14 months for glioblastoma patients, estimating survival time beyond 28 months is less important than under 28 months.

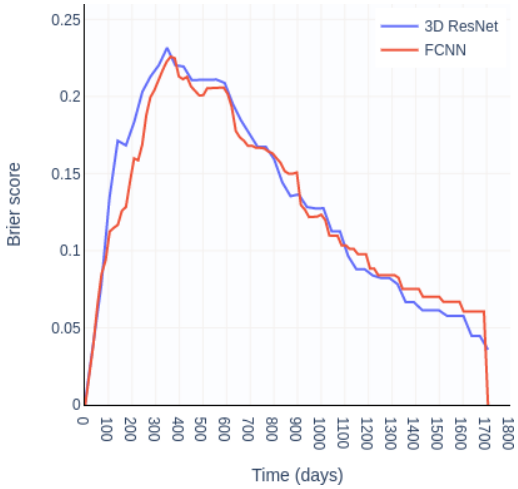


Figure 3.4: Brier score

3.2.3 IPCW Brier score

The IPCW Brier score of the models is depicted in Table 3.1. The 3D ResNet model scored a IPCW Brier score of 0.131 without interpolation and 0.127 with interpolation. The FCNN model scored an IPCW Brier score of 0.117 without interpolation and 0.119 with interpolation.

Interpolation thus slightly improves the 3D ResNet survival estimates, but slightly worsens those of the FCNN. Considering the added benefits of a continuous-time survival curve, this trade off might be considered negligible. The interpolated survival curves allow an even more fine grained look at a patients survival probabilities over time. As explained earlier, an IPCW Brier score close to 0 means a model is able to accurately match predicted and actual survival curves. Because the FCNN model has lower IPCW Brier scores, it is, compared to the 3D ResNet model, able to more accurately match predicted and actual survival times. This verifies our earlier findings.

Model	C-index	C-index interpolated	IPCW	IPCW interpolated
3D ResNet-10	0.621	0.626	0.131	0.127
FCNN	0.697	0.696	0.117	0.119

Table 3.1: C-indexes and IPCW Brier scores for both non-interpolated and interpolated test dataset predictions.

3.3 Discriminative power

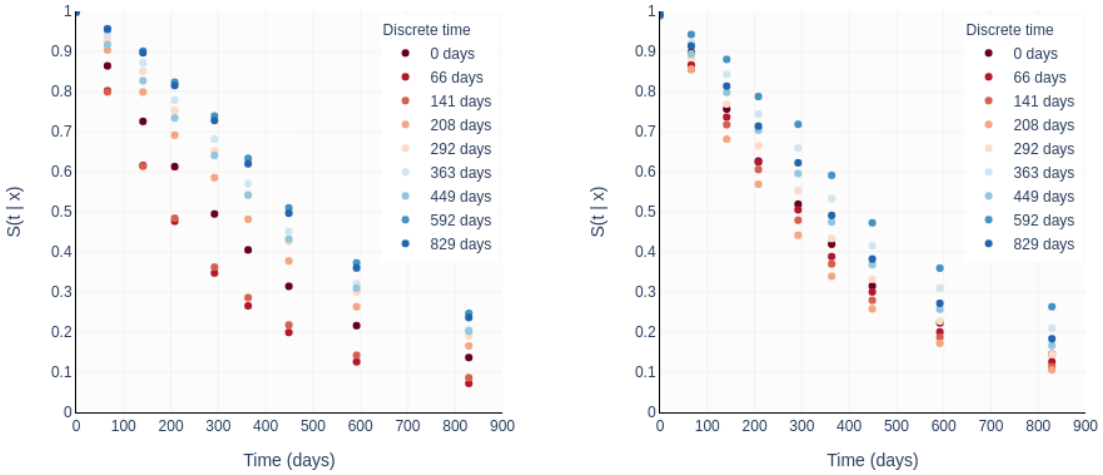
To evaluate the discriminative power of both the FCNN as well as the 3D ResNet model, we calculated the C-index. The C-index of the models is depicted in Figure 3.1. The 3D ResNet model scored a C-index of 0.621. The FCNN model scored a C-index of 0.697. The FCNN model is thus able to more accurately distinguish between patients who die early and patients who die later. Furthermore, we plotted the predicted average survival curves in Figure 3.5. The figure shows the survival probability over time for patients belonging to a given discrete time point. Here, a steep step down means that the survival probability of patients belonging to that discrete time point declines fast over time. Furthermore, more gradual steps down means that the survival probability of patients belonging to that discrete time point declines more gradually over time.

In Figure 3.5 the discrete time point 66 days consists of patients that either died between 0 and 66 days or were censored between 66 and 141 days. The discrete time point 141 days consists of patients that died between 66 and 141 days or were censored between 141 and 208 days et cetera. Since discrete time points 0 and 1785 days are the first and last point respectively, time point 0 days consists of only censored patients and time point 1785 days consists of only uncensored patients. Figure 3.5 thus shows the predicted survival probabilities for patients belonging to a certain discrete time point. In a perfect model, the survival probability for patients that died between 0 and 66 days should decline fast between 0 and 66 days and, if all patients died, the survival probability at 66 days should be 0% ($p = 0.0$). However, if there are censored patients (patients who have not died) this probability should be higher because some patients are still alive. Similarly, for discrete time point 592 days, survival probability should decrease sharply between 592 and 829 days. The predicted survival probabilities should accurately depict this described behaviour. For visualization purposes and because the predicted survival probabilities beyond 829 days are inaccurate as shown in Section 3.2.1, we will look at the predicted survival probabilities up to 829 days.

The predicted average survival probabilities for patients belonging to every discrete time point are depicted in Figure 3.5. We chose to show the survival predictions as markers instead of survival curves, because the survival curves would overlap each other and make the figure unclear. Figure 3.5a shows the average predicted survival probabilities for the model trained on the clinical dataset and Figure 3.5b shows the average predicted survival probabilities for the model trained on the 3D MRI dataset.

As mentioned earlier, in a perfect model, the survival probability declines the fastest for patients belonging to discrete time point 66 days and the slowest for patients belonging to discrete time point 829 days. If we look at the survival probability at 292 days for the 3D ResNet model,

we can see that patients that died between 141 and 208 days or were censored between 208 and 292 days have the lowest predicted survival probability. Ideally, this would be the patients that died between 0 and 66 days. Naturally, those patients should have a lower survival probability at 292 days than patients that died between 141 and 208 days. However, if we would look at this from a low vs. high survival perspective, it is clear that for patients with (actual) high survival times the survival probabilities are higher than for patients with low survival times. Although the order of the survival probabilities may not be perfect, the figure itself does not show which survival curves are close or far from the truth.



(a) Clinical Dataset model **(b)** 3D MRI Dataset model

Figure 3.5: Average predicted survival probabilities. The y-axis shows the predicted survival probability. Every marker corresponds to a discrete time point. For example, the time point "0 days" consists of patients that were censored between 0 and 66 days. The time point "66 days" consists of patients that died between 0 and 66 days or were censored between 66 and 141 days.

Figure 3.6 shows the actual survival curves for every group in the test dataset. We can see that the predicted survival probabilities in Figure 3.5 are more optimistic compared to the actual survival curves. The survival curve for group 0 has a max and min of 1. This, again, is because censored patients are rounded down, while uncensored patients are rounded up. This means that every patient that died between 0 and 66 days is part of discrete time point 66 days and every patient that is censored between 0 and 66 days is part of discrete time point 0 days. Thus, naturally, discrete time point 0 days has no patients that died.

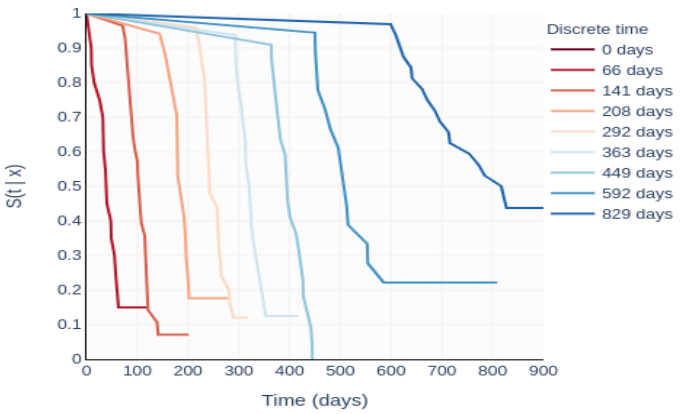
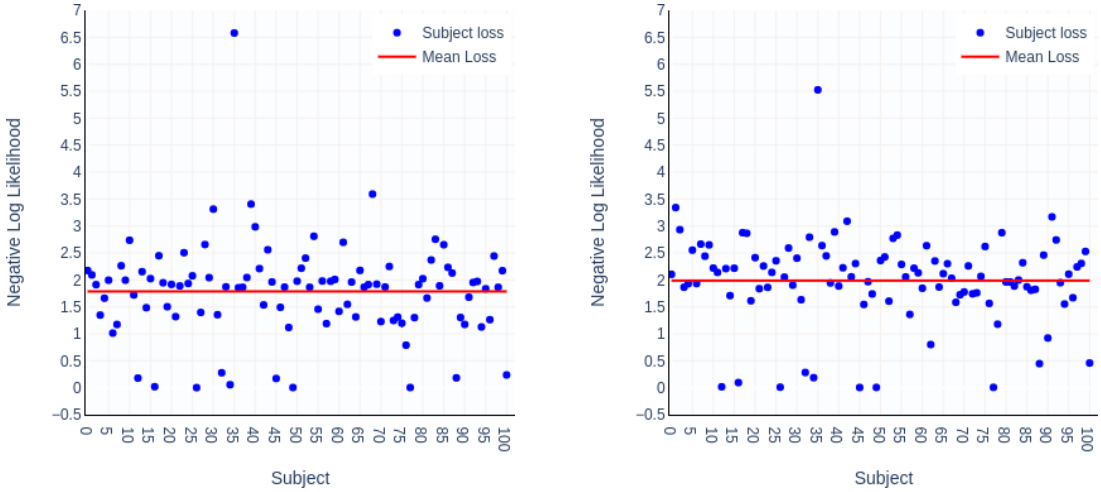


Figure 3.6: Actual survival curve for patients belonging to a certain discrete time point. These are the actual survival curves for patients of every discrete time point. An ideal model would predict the same survival curves. The survival curves of patients that died fast decline fast over time while the survival curves of patients that died late decline slowly over time

3.4 Validation loss distribution

To understand whether the models perform equally well on every patient, the individual loss of every patient in the validation datasets are plotted in Figure 3.7. As explained earlier, a high loss means that the model's predicted survival probabilities are far from the actual survival probabilities, while a low loss means that the model's predicted survival probabilities are close to the actual survival probabilities. See Section 2.6 for a more in depth explanation of how the loss is calculated. Figure 3.7a shows the validation loss spread of the FCNN model. The mean loss is 1.786 (±0.881). Figure 3.7b shows the validation loss spread of the 3D ResNet model. The mean loss is 1.980 (±0.825). Because the FCNN model has a lower mean loss, it is evident that it outperforms the 3D ResNet model.



(a) FCNN model (Clinical Dataset)

(b) 3D ResNet model (3D MRI Dataset)

Figure 3.7: Individual Loss of every patient in validation dataset

At first glance, it is clear that both models are able to predict the survival times of certain patients very well. The 10 patients with the lowest loss are exactly the same in both models and they were

all censored, indicating that both models were accurately able to predict those patients to have a high survival probability at their censoring time. Figure 3.8 shows the survival curves of three of the ten patients and we indeed observe accurate survival probabilities around the censoring times.

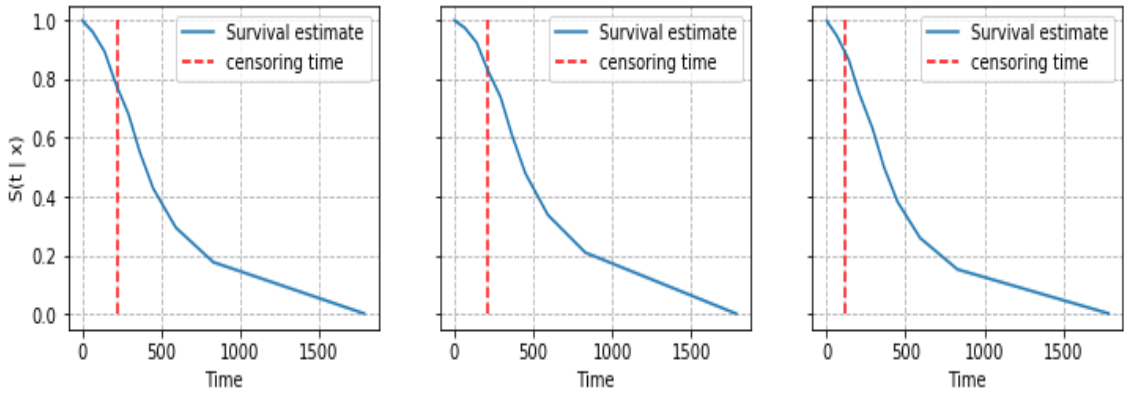


Figure 3.8: FCNN survival estimates of three of the ten (censored) patients with the lowest negative log-likelihood (loss).

Figure 3.9 shows the average loss for patients belonging to a particular discrete time point in the validation dataset for both models. Here, we see much better how the models perform in detail. Both models perform very well on predicting the survival probability of patients belonging to the first two discrete time points (the markers are overlapping for 0 days), indicated by the low loss. However, the 0 days time point only consists of patients that were censored between 0 and 66 days. While the 3D ResNet performs the best on predicting survival for patients with low survival times, the model struggles to accurately predict survival for patients with high survival times. This trend is also the same for the FCNN model. For patients that died between 66 and 141 days, the models perform equally well. In general, the 3D ResNet model is able to make more accurate predictions for patients that died between 208 and 292 days and between 292 and 363 days than the FCNN model. These findings highlight the differences between the models and suggest that combining both models could improve the survival predictions.

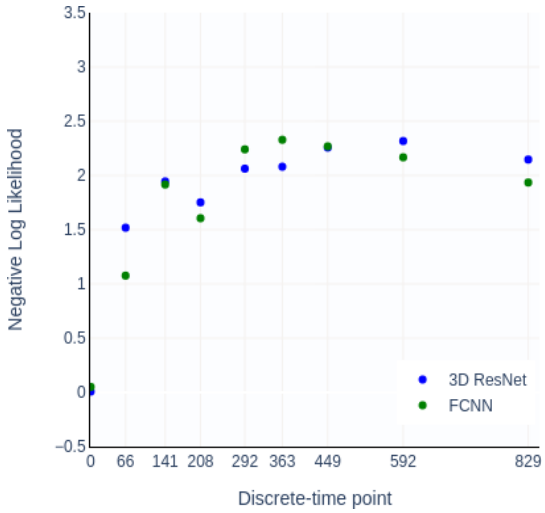
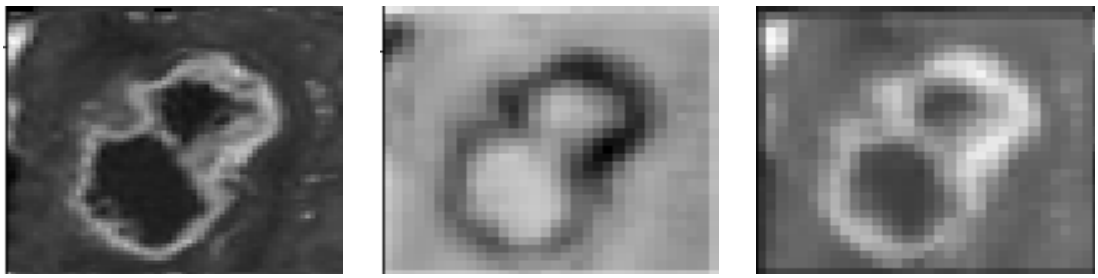


Figure 3.9: Average loss per discrete time point in validation dataset for both models. While on average the FCNN model has a lower loss, the 3D ResNet model outperforms the FCNN model at predicting survival probabilities of patients that belong to certain discrete time points.

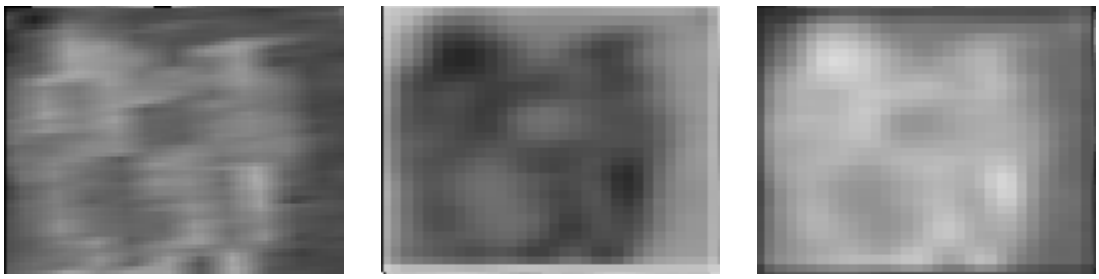
3.5 Feature Maps

As described in Chapter 2.12, visualizing the features of our 3D ResNet model helps us to understand the learned features. They show the intermediate representations of an input 3D MR image in the model. For visualization purposes, we will look at two of the 64 feature maps of the first 3D Convolutional layer of our 3D ResNet model. First, we show the feature maps based on one 3D MRI sequence. We do this for the T1c sequence and the FLR sequence. Second, we show the feature maps based on a multi-sequence 3D MR image.

For a random (test) patient we plotted 2D slices of both the actual input sequence and two corresponding (learned) feature maps. Figure 3.10 shows the feature maps of 2 filters of the first 3D Convolutional layer of the 3D ResNet model. Figure 3.10a shows the T1c modality of the (actual) input image. Figure 3.10b and 3.10c show two corresponding feature maps. Based on the shape and the dark line around the tumor, we can see that the first T1c filter is able to correctly identify the border of a tumor. Based on the dark pixels in the tumor, we can see that the second T1c filter is able to correctly identify the tumor itself without the border. Figure 3.10d shows the FLR modality of the (actual) input image. Figure 3.10e and 3.10f show two corresponding feature maps. The first filter seems to capture the tumor area, while the second filter captures the non-tumor area.



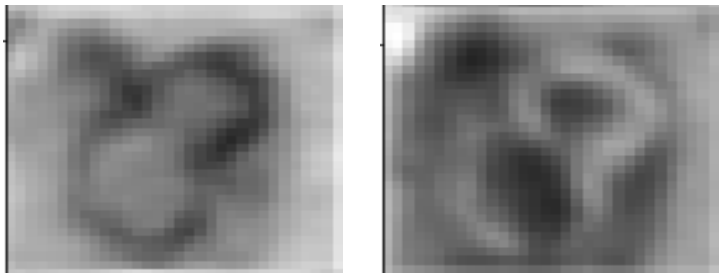
(a) Input T1c modality **(b)** Feature map 1 of T1c **(c)** Feature map 2 of T1c



(d) Input FLR modality **(e)** Feature map 1 of FLR **(f)** Feature map 2 of FLR

Figure 3.10: Two feature maps of the first 3D Convolutional layer of the 3D ResNet model based on T1c and FLR sequences of a random patient

Figure 3.11 shows the feature maps of 2 filters of the first 3D Convolutional layer of the 3D ResNet model based on a whole 3D MR image (4 sequences). The first filter shown in Figure 3.11a is able to identify the border of the tumor and the tumor itself. Figure 3.11b shows the second filter and captures the tumor itself and the area around it which seems to be result of the FLR filter based on the same shape of the dark pixels. Just like in Figure 3.10c, the weights are low for the tumor border. Overall, it seems that the model combined both filters to capture both the tumor itself as well as the adjacent tissue.



(a) Feature map 1 of all sequences **(b)** Feature map 2 of all sequences

Figure 3.11: Two feature maps of the first 3D Convolutional layer of the 3D ResNet model based on 4 MRI sequences of a random patient

Chapter 4

Discussion

We investigated the use of a Logistic-Hazard 3D ResNet model to predict survival curves of glioblastoma patients using multi-sequence 3D brain MR images. We were able to combine both a cutting-edge Deep Learning model (ResNet) and one of the most promising adaptions of survival analysis in machine learning (Logistic-Hazard). The first hypothesis was that the Logistic-Hazard method can be used to create a 3D ResNet model that is able to predict a survival curve per glioblastoma patient based on his/her 3D brain MR images. Both the 3D ResNet model as well as the FCNN were able to predict a survival curve per glioblastoma patient. The 3D ResNet model achieved a C-index of 0.621 and a IPCW Brier score of 0.131. The second hypothesis was that the 3D ResNet model would outperform a FCNN that was trained on clinical variables. However, our FCNN model outperformed the 3D ResNet model with a C-index of 0.697 and a IPCW Brier score of 0.117. Both models were found to be well calibrated as can be seen by the IPCW Brier scores, but lacked some discriminative power.

4.1 Previous studies

4.1.1 Studies that used MRI data

There are, to the best of our knowledge, no studies that have used Deep Learning to predict survival curves of individual glioblastoma patients. Due to this, it is difficult to compare the results of the 3D ResNet model to previously conducted studies. However, we compared our results to the (limited) studies that predicted OS time of glioblastoma patients instead.

Lao et al. (2017) used a Deep Learning based radiomics model to predict survival of glioblastoma patients. The authors categorized survival in two groups: low OS and high OS. This radiomics model achieved a C-index of 0.731. Another study, conducted by Nie et al. (2016), predicted OS time using a 3D Deep Learning model with an accuracy of 89.9%. Chato and Latifi (2017) extracted features from MRI modalities and developed a prediction model that used the extracted features to predict OS time (Chato & Latifi, 2017). The trained model achieved an accuracy of 73% by using a Linear Discriminant classifier (Chato & Latifi, 2017). R. Liu et al. (2016) also extracted features from MRI modalities and were able to classify survival time of patients with glioblastoma in short OS and long OS with a 95% accuracy. Baid et al. (2018) were able to classify survival time of patients with glioblastoma in short OS and long OS with a 63.6% testing accuracy using the BRATS dataset. Z. Liu et al. (2019) used a Deep Attention Network to predict survival of glioblastoma patients using MR images. The model achieved a C-index of 0.68 on unseen data, but was not able to predict a survival curve per patient.

However, contrary to this study, all aforementioned studies used small datasets and lacked multi-center data to validate their model' generalizability. Moreover, the use of a dichotomous outcome prevents a fine-grained look into a patients survival probability over time. For example, Nie et al.

(2016) used a threshold of 22 months to divide the patients into short OS and long OS. Because of this, the model did not provide any information regarding where in those 22 months an event happens. Considering that the median survival for glioblastoma does not exceed 14 months, having more accurate survival predictions in those 22 months is important. The model also did not provide any insight in how much the survival probability declined after a certain amount of months for any given patient. This shortcoming was not present in this study, both the 3D ResNet model as well as the FCNN model were capable of outputting a survival curve over time per patient. This allows clinicians a more fine-grained look at the survival estimates.

4.1.2 Studies that used clinical data

First, we will compare our FCNN model to previous studies that have used an ANN to predict survival. Second, we will compare our FCNN model to previous studies that used more traditional approaches, such as Cox regression, to predict survival.

Hao et al. (2018) developed a deep pathway-based sparse deep neural network named Cox-PASNet to predict survival of glioblastoma patients. The model was able to take both biological pathways and clinical data as input. However, only age was incorporated in the clinical layer of the model. Nonetheless, the model obtained a C-index of 0.635 on the glioblastoma dataset, which is lower than our FCNN model (Hao et al., 2018). Similar to this study, the authors found that age was a strong prognostic factor of glioblastoma (Hao et al., 2018). Furthermore, Yoon et al. (2020) recently developed a Deep Learning model for prediction of OS of glioblastoma patients using both clinical and radiomic data. The model achieved a C-index of 0.693 by using clinical data alone, almost matching our FCNN model (Yoon et al., 2020). However, the model achieved a C-index of 0.768 by incorporating both clinical and radiomic data, suggesting the potential of multi-parametric deep learning models (Yoon et al., 2020).

Datema et al. (2012) predicted the survival probability of newly diagnosed patients with head and neck cancer using a Cox regression model. Again, the authors found age to be the most important predictor of OS. The model achieved a C-index of 0.73 on the Leiden cohort and a C-index of 0.69 on the American cohort, highlighting its generalizability. We were not able to validate the FCNN model using an external dataset, because it requires the dataset to have the same variables as the ones used in this study.

4.2 Strengths

As mentioned earlier in Chapter 1, most previous studies have used radiomic approaches to predict survival times of glioblastoma patients. These studies thus relied on handcrafted features from the region of interest. In this study, we used a Deep Learning approach that has proven performance in capturing image features. Moreover, unlike radiomic approaches, our 3D ResNet does not rely on handcrafted features but on automated feature extraction. Furthermore, we used feature maps to show that the 3D ResNet model learned to identify tumors correctly. By doing so we unfolded part of the black-box of CNNs. However, there is still much work to be done to unfold the rest of the black-box, especially because in the medical domain wrong predictions can be harmful (Samek, Wiegand & Müller, 2017).

Another strength of this study is the use of multi-national hospital MRI data. As mentioned earlier in Chapter 1.1, many studies used small datasets originating from one hospital. We have not tested the performance of our 3D ResNet model on external data. However, we hypothesize that our 3D ResNet model is more generalizable than models of previous studies due to the use of MRI data originating from 12 different multi-national hospitals and the (much) larger dataset that was used compared to previous studies.

Furthermore, we applied Logistic-Hazard; one of the most promising method of survival analysis in machine learning (Zadeh & Schmid, 2020; Kvamme & Borgan, 2019b). It is fast, deals with non-proportional hazards and uses the NLL as loss function instead of BCE (Zadeh & Schmid, 2020; Kvamme & Borgan, 2019b). As explained earlier in Chapter 1, the BCE loss function results in heavy biases of survival predictions and because of these limitations a recent study by Zadeh and Schmid (2020) suggested that the NLL loss function should instead be used for training ANNs that aim to predict survival times. To the best of our knowledge, we are the first to use the Logistic-Hazard method to predict survival times of glioblastoma patients using 3D MR images. Moreover, we are the first to predict survival curves rather than OS for glioblastoma patients using Deep Learning techniques.

4.3 Limitations

Due to computational limitations, we created a bounding box around the brain tumor and used that to predict survival times. This allowed us to train multiple models in a relatively fast way. Due to the creation of bounding boxes, there was some loss of information. Although the 3D ResNet model could have learned that a tumor is on the edge of a brain, the tumors exact location in the brain is lost due to the creation of a bounding box. Moreover, the model was not able to take the surrounding brain tissue into account. The 3D ResNet model might benefit from the additional information retrieved from whole brain 3D MRI images. However, this is a computational limitation, not a limitation of the 3D ResNet model. The model can already handle different input sizes.

In this study, the discretization of the survival times resulted in a step-wise survival curve. Continuous survival models are able to predict a more smooth survival curve. However, by using more time intervals, the step-wise survival curve can become smoother. Furthermore, we interpolated the discrete time survival estimates using an interpolation scheme to produce a smoother survival curve.

Another limitation of the 3D ResNet and the FCNN model is the unknown uncertainty of a prediction. Certainty of a prediction is especially important in health care, because predictions might impact decision making. One way of obtaining prediction uncertainty is by a method called Dropout uncertainty. The use of dropout at inference time can be interpreted as a Bayesian approximation of the Gaussian process (Gal & Ghahramani, 2016). This can be used to determine the uncertainty of any given prediction.

Furthermore, the predictions of 3D CNN models are often hard to explain and the models are a lot of times referred to as black-boxes. However, we illustrated (part of) the decision making process of a CNN using feature maps. Feature maps showed us which part of a given 3D MRI is the most relevant for a particular prediction made by our 3D ResNet model (Dabkowski & Gal, 2017).

As mentioned earlier, intensity variations are hard to capture because hospitals do not always use the same MRI scanners and sequence configurations. Despite the development of more sophisticated normalization techniques, Z-score normalization is found to be the most robust (Carré et al., 2020). However, because we use MR images from a large amount of different hospitals, we hypothesize that the 3D ResNet model could perform better with more image intensity tuning. For example, some (more advanced) normalization techniques require a T1w image to be applied. However, the T1w image was missing for many patients of our MRI dataset and therefore we could not apply these techniques. Future researchers should keep this in mind when collecting MRI data.

Lastly, together with the C-index, the IPCW Brier score is the most common evaluation criteria for survival predictions. However, it can be biased under administrative censoring (Kvamme & Borgan, 2019a). Administrative censoring happens when the censoring times are known for all patients, which is the case in our datasets. To solve this bias, Kvamme and Borgan (2019a)

proposed the administrative Brier score. The authors showed that, under administrative censoring, the administrative Brier score was more able to identify the best survival estimates than the IPCW Brier score. However, that was beyond the scope of this study. Future work that uses datasets where administrative censoring is prevalent can benefit from using the administrative Brier score to identify the best survival estimates.

4.4 Future work

Future work could use whole brain 3D MRI images to train the 3D ResNet model. This might lead to more accurate survival time predictions. Furthermore, larger datasets would allow the use of a finer discretization grid and thus produce smoother survival curves using the Logistic-Hazard method. In the medical domain knowing the uncertainty of a prediction made by any model is a prerequisite for using such methods in the practice. Applying Dropout uncertainty could provide insight in the uncertainty of any given prediction. Another prerequisite of using models like the 3D ResNet model of this study is knowing what image features the model used to predict survival times of a patient. Although feature maps work well for visualizing the first convolutional layers of a CNN model, they can not be used to understand the deeper features the model learned. However, the field of explainable AI is constantly evolving and methods to overcome this limitation are investigated (Samek et al., 2017). With the increasingly large medical datasets, the issue of normalization remains hard to overcome. Over the recent years, new, more advanced 3D MRI normalization techniques have been developed. However, these normalization techniques often require certain sequences to be available. Thus more research is needed to develop a standardized pre-processing pipeline that can handle 3D MR images of different hospitals. We also recommend the use of the administrative Brier score rather than the IPCW Brier score when datasets contain administrative censoring. Lastly, we trained two separate models: one trained on 3D MRI data and one trained on clinical data. However, we hypothesize that the two datasets can complement each other and an ensemble model trained on both datasets simultaneously might produce a model that outperforms both the 3D ResNet as well as the FCNN.

Chapter 5

Social Entrepreneurship

In this study we contributed to the field of survival analysis in health care. More specifically, we provided ways of predicting survival time for glioblastoma patients. Glioblastoma patients often have a short life expectancy and providing them with information regarding their prognosis could help them to plan their lives ahead. We were motivated by the potential societal impact this study has. Gaining a larger understanding of how survival analysis relates to society allowed us to develop new models that are able to predict survival times for glioblastoma patients. As social entrepreneurs we expect the methods described in this paper to cause change in society even after we have moved on to address different societal problems.

5.1 Societal Impact

There have been many technological advances within the field of data science. Not only do we now have more data and more computer processing speed, but also an increasingly larger data science talent pool. Because of this, the development of data science tools within health care is being accelerated and evidently the possible adoption of such tools also has an impact on the society. Despite having many Artificial Intelligence (AI) models that have better diagnostic accuracy than expert clinicians, the models often are deemed to have low credibility and are rarely adopted by health professionals (Shortliffe & Sepúlveda, 2018). To evaluate the (possible) societal impact of implementation of methods such as the ones described in this study, we will dive deeper in the use of data science or more specifically, AI, in healthcare and the implications it has. We will do this from the point of shared decision making (SDM) between a health professional and a patient.

Traditionally, clinicians were expected to make the best clinical decisions and patients were not involved much in the decision-making process (Fritz et al., 2016). However, in the last decade, there has been a rapid increase in SDM training programs for healthcare professionals (Oerlemans, Knippenberg & Olthuis, 2020). SDM or patient-centered care refers to a healthcare approach where the patient's preferences and values are considered in clinical decisions. It is found to boost respect for autonomy, non-maleficence and justice (Oerlemans et al., 2020). Providing prognostic information to a glioblastoma patient can be valuable, because, as mentioned earlier, the 5 year relative survival rate for glioblastoma is very low. Having prognostic information as a patient can thus make a difference when considering treatment options (Lagarde et al., 2008). This is especially true in the case treatment options directly impact important functions of the brain such as speech (Dronkers, Hoesseini, de Boer & Offerman, 2018). The patient then has to consider the trade-off between a long survival time and quality of life (Dronkers et al., 2018). Moreover, with the increasing use of AI in healthcare, patients are increasingly empowered to take care of their own health through the use of health applications (Tran, Riveros & Ravaud, 2019).

However, despite studies finding that SDM improves the patients satisfaction as well as the pa-

tients health outcomes, many physicians still find communicating prognosis to patients difficult (*Artificial Intelligence in Healthcare*, 2020; Dronkers et al., 2018). Some reasons for this include uncertainty about the prognosis or uncertainty about how to communicate the prognosis (Dronkers et al., 2018). Due to the black box nature of many AI models, uncertainty about the prognosis is especially a problem for AI solutions in healthcare (Maddox, Rumsfeld & Payne, 2019). Health professionals that make use of an AI solution need to understand the basis for any advice given (Shortliffe & Sepúlveda, 2018). Furthermore, the solution should be intuitive and simple to learn (Shortliffe & Sepúlveda, 2018). Lastly, because of the societal impact of AI solutions in health care, the scientific foundation for such a model must be strong (Shortliffe & Sepúlveda, 2018).

The adaptation of AI in health care requires research on how best to address some of the main issues, especially when patients' well-being is at stake. One could argue that there also needs to be SDM between a computer and a health professional. For example, studies have found that health professionals are biased toward optimistic predictions which often leads to them overestimating survival time (Dronkers et al., 2018; Verghese, Shah & Harrington, 2018). Using a predictive AI model can mitigate this by providing health professionals a different, well-calibrated, outcome (Verghese et al., 2018). This approach to AI models in healthcare, keeping human intelligence in the loop, has a lot of support among doctors and other health professionals (Shortliffe & Sepúlveda, 2018; Verghese et al., 2018; Stead, 2018). Shortliffe and Sepúlveda (2018) argue that an AI model should assist but not replace a clinician. With the help of AI, certain tasks of a health professional could be externalized, freeing up more time for doctor-patient contact (Shortliffe & Sepúlveda, 2018).

In practice, however, there are a lot of factors that contribute to whether a clinician communicates prognostic predictions retrieved from AI models to patients or not. Based on patient groups, there are differences in communicating prognostic information (Wittenberg-Lyles, Goldsmith, Sanchez-Reilly & Ragan, 2008). For example, glioblastoma patients have a low survival and most of the patients die as a result of the disease (Krex et al., 2007; Wittenberg-Lyles et al., 2008). Furthermore, glioblastoma patients are more likely to have cognitive and physical problems such as speech loss and lack of balance (Halkett, Lobb, Oldham & Nowak, 2010). A study conducted by Chochinov, Tataryn, Wilson, Enns and Lander (2000) found that patients who get accurate estimates of life expectancy are more likely to reach acceptance of terminal illness and discuss preferences for treatment more often. Moreover, studies have found that oncologists that disclose prognosis to glioblastoma patients improve a patients understanding of their expected life expectancy (Diamond et al., 2017). However, despite studies underlying the importance of disclosing prognosis to patients, oncologist often do not share prognosis with their patients (Daugherty & Hlubocky, 2008). Dronkers et al. (2018) found that health professionals often have difficulty in communicating prognosis to patients. This is especially true for glioblastoma patients where the prognosis is often uncertain. In the PICTURE research group there was a general understanding that prognostic predictions from an AI model, such as the one described in this study, should mostly be used by decision makers. However, unclear prognosis is often experienced as frustrating by glioblastoma patients, especially because symptoms can impact their life drastically (Halkett et al., 2010). Because of uncertain predictions, patients are often no longer able to plan their lives ahead (Halkett et al., 2010). Moreover, recent unpublished research has found that less informed oncology patients have higher decisional conflict: they regret their decisions more than informed patients. A glioblastoma patient that is actively involved with the PICTURE research group highlighted the need for communicating prognosis:

The earlier there is an analysis with a clear graph about where the patient (or I) stand at this moment, the better. [Patient]

We explored ways to overcome the aforementioned issues of using AI in health care. A neurosurgeon noted that although the black box nature of an ANN is problematic when it comes to adopting

such methods in healthcare, it is not the most important objection to AI models. Most clinicians see such models as just another source of information that they can use. Thus such models can help in decision making.

Ultimately, it is about the balanced judgment of all information sources by an expert. [Neurosurgeon]

Clinicians often mentioned that communicating prognosis to patients is already a difficult task. Emilie Dronkers has researched the communication between patients and clinicians extensively and noted that a possible solution could be to share (AI) predicted survival curves with patients. In this way, a health professional and a patient could practice shared-decision making together. However, the current platform used by the PICTURE research group is aimed at providing health professionals information that can be used for decision making and does not involve the patient.

One of the reasons of not communicating probabilities such as the ones provided in this study, is that many patients cannot reason in terms of probabilities. Studies have found that there are fundamental difficulties in understanding probabilistic information by patients (Fagerlin, Wang & Ubel, 2005; Fagerlin et al., 2004). However, there are methods to make it less difficult to communicate probabilistic information. Price, Cameron and Butow (2007) studied the influence of graphical display format on quantitative information perception. Graphical formats were found to assist in understanding quantitative information such as probabilities (Price et al., 2007). For example, the authors found that risk information, such as survival time, presented in horizontal format is perceived more accurately than when it is presented in vertical format (Price et al., 2007).

Another issue is that of the amount of information clinical teams must assimilate (Maddox et al., 2019). Many clinicians feel that they cannot keep up with the abundance of new information. Shortliffe and Sepúlveda (2018) argued that health professionals need to understand the basis for any advice given. Maddox et al. (2019) noted that this extra load might lead to higher stress and lower efficiency among clinicians. Furthermore, Emanuel and Wachter (2019) noted that the process of adapting AI in healthcare "will be iterative and messy", highlighting the challenges of behavioral change among health professionals.

It is a big step for many colleagues. [Neurosurgeon]

Although AI solutions can provide physicians additional tools to diagnose diseases and predict health outcomes, they are often seen as a threat to existing jobs (Kabir, 2019). However, most health professionals doubt that machines will replace their human tasks in the foreseeable future (García, Spatharou, Jenkins & Hieronimus, 2020). The goal of AI solutions is thus not to replace or automate health professionals, but rather externalize some of their tasks (Kabir, 2019; Susskind & Susskind, 2015). The human connection between health professionals and patients is vital in providing quality health care and is something that is not easily replaced by AI (Susskind & Susskind, 2015). Instead, AI solutions can address repetitive and largely administrative tasks which now cost as much as 70% of health professionals' time (García et al., 2020). This in turn leads to a higher job satisfaction and possibly lower stress among health professionals (García et al., 2020). Moreover, by externalizing such tasks, health professionals can spend more time looking after patients. For example, a recent study by O'Neill et al. (2020) found that implementing an AI tool to help re-prioritize radiologists' CT work lists helped a hospital to reduce turnaround and wait times. Solutions such as the one described in this study can help health professionals to decide on the best treatment option faster, saving time in the process. AI can augment clinical activities and provide health professionals access to new information that then can be used to improve patient outcomes. The McKinsey Global Institute (MGI) expects that 15 percent of

work hours in healthcare will eventually be automated (García et al., 2020). This is especially important in a sector that faces staff shortages and skill gaps around the globe. According to the WHO, in 2030, there will be a shortfall of 10 million health professionals globally (*Global strategy on human resources for health: Workforce 2030*, n.d.).

The question is thus not whether AI will transform healthcare, but when. AI can bring many opportunities for health care systems to improve patient care and it is time to rethink education and skills of health professionals to grab these opportunities. Many health professionals lack digital skills and the rise of AI in healthcare will require them to be skilled in both medical and data science (Jorgensen, 2019; Sapci & Sapci, 2020). Nowadays, patients can already be informed about their health via their smartphone (Jorgensen, 2019). However, the lack of digital skills among health professionals hampers the development of such e-health applications (Jorgensen, 2019). Studies that educate students in both medicine and data science are lacking, especially in Europe (Jorgensen, 2019). Hospitals and other healthcare institutions need to provide current health professionals tools with which they can continue learning new data science related skills (Rimmer, 2019). By doing so, AI solutions can augment the skills of health professionals. For example, an AI model was found to outperform radiologists in identifying breast cancer (McKinney et al., 2020). However, radiologists need to develop the skills to understand for which patients the AI model should or should not be used (Rimmer, 2019). Moreover, the demand for data scientists and data engineers is high in almost all industries and healthcare organisations need to attract talented people of those fields to pave the road of scaling AI in healthcare (García et al., 2020).

There is a general trend towards more community-based care built on patient empowerment (McDougall, 2019). Shared-decision making has been studied extensively and proven to improve patient outcomes (Dronkers et al., 2018). Nowadays, patients have access to a vast amount of medical knowledge and are increasingly using that information to evaluate options in advance (Aboueid, Liu, Desta, Chaurasia & Ebrahim, 2019). There are already AI solutions that are able to correctly provide diagnoses and the use of such solutions is projected to increase in the future (Imran et al., 2020; García et al., 2020). "Patients are the most under-used resource in healthcare" was first uttered in the 1970s by Warner Slack (deBronkart & Sands, 2018). Slack believed that if data is "implemented wisely and well" it can empower both health professionals and patients (deBronkart & Sands, 2018). Now, 50 years later, this is more true than ever. AI is fundamentally changing how health professionals will do their jobs in the future and AI solutions could prompt health professionals and patients to discuss treatment goals together (McDougall, 2019; García et al., 2020). To empower patients even more, the design and implementation of AI solutions in healthcare should include patients to ensure their needs and preferences are considered.

It's time to reduce paternalism in healthcare. Other industries involve the customer as much as possible. Why should medicine be different? [Adolfo Fernández-Valmayor, Quirónsalud] (García et al., 2020)

To conclude, it is time to reduce paternalism in healthcare and embrace the views of patients. Patients will have more and more access to AI solutions from home and health professionals need to understand how to deal with these developments (Rimmer, 2019; García et al., 2020). AI solutions already outperform physicians in detecting certain cancers and any physician should be able to explain such tools to his/her patients and reach a shared diagnosis (McKinney et al., 2020; Tran et al., 2019). AI has the ability to empower patients to be more informed about their health (Rimmer, 2019). We expect, with the inevitable adoption of AI in healthcare, the need for soft skills will be greater than ever. With automation, repetitive and largely administrative tasks will be a job for machines, leaving more time for health professionals to focus on patients. With this extra focus on patients, we hope shared-decision making takes a central role in the consultation room. We stand at a pivotal moment where we can shape the use of AI in health care and we need to make decisions together with patients to ensure AI solutions empower them as much as possible.

References

Aboueid, S., Liu, R. H., Desta, B. N., Chaurasia, A. & Ebrahim, S. (2019). The use of artificially intelligent self-diagnosing digital platforms by the general public: Scoping review. *JMIR medical informatics*, *7*(2), e13445. 36

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., ... others (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, *5*(1), 1–9. 1, 2, 3

Ahmed, K. B., Hall, L. O., Goldgof, D. B., Liu, R. & Gatenby, R. A. (2017). Fine-tuning convolutional deep features for mri based brain tumor classification. In *Medical imaging 2017: Computer-aided diagnosis* (Vol. 10134, p. 101342E). 4

Ammirati, M., Vick, N., Youlian, L., Ivan, C. & Mikhael, M. (1987). Effect of the extent of surgical resection on survival and quality of life in patients with supratentorial glioblastomas and anaplastic astrocytomas. *Neurosurgery*, *21*(2), 201–206. 1

Antiga, L. (2020). *Deep learning with pytorch*. City: Manning Publications. 13, 17

Artificial intelligence in healthcare. (2020). Elsevier. Retrieved from https://doi.org/10.1016/c2018-0-04097-9 doi: 10.1016/c2018-0-04097-9 34

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M. H., Moiyadi, A., ... Mahajan, A. (2018). Deep learning radiomics algorithm for gliomas (drag) model: a novel approach using 3d unet based deep convolutional neural network for predicting survival in gliomas. In *International miccai brainlesion workshop* (pp. 369–379). 3, 4, 29

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... others (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*. 3

Banerjee, S., Mitra, S., Shankar, B. U. & Hayashi, Y. (2016). A novel gbm saliency detection model using multi-channel mri. *PloS one*, *11*(1), e0146388. 10

Carré, A., Klausner, G., Edjlali, M., Lerousseau, M., Briend-Diop, J., Sun, R., ... others (2020). Standardization of brain mr images across machines and protocols: bridging the gap for mri-based radiomics. *Scientific reports*, *10*(1), 1–15. 31

Chato, L. & Latifi, S. (2017). Machine learning and deep learning techniques to predict overall survival of brain tumor patients using mri images. In *2017 ieee 17th international conference on bioinformatics and bioengineering (bibe)* (pp. 9–14). 3, 29

Chochinov, H. M., Tataryn, D. J., Wilson, K. G., Enns, M. & Lander, S. (2000). Prognostic awareness and the terminally ill. *Psychosomatics*, *41*(6), 500–504. 34

Dabkowski, P. & Gal, Y. (2017). Real time image saliency for black box classifiers. In *Advances in neural information processing systems* (pp. 6967–6976). 31

Datema, F. R., Ferrier, M. B., Vergouwe, Y., Moya, A., Molenaar, J., Piccirillo, J. F. & de Jong, R. J. B. (2012, jul). Update and external validation of a head and neck cancer prognostic model. *Head & Neck*, *35*(9), 1232–1237. Retrieved from https://doi.org/10.1002/hed.23117 doi: 10.1002/hed.23117 30

Daugherty, C. K. & Hlubocky, F. J. (2008). What are terminally ill cancer patients told about their expected deaths? a study of cancer physicians' self-reports of prognosis disclosure. *Journal of Clinical Oncology*, *26*(36), 5988. 34

deBronkart, D. & Sands, D. (2018). Warner slack: "patients are the most underused resource". *British Medical Journal*, *362*. 36

Diamond, E. L., Prigerson, H. G., Correa, D. C., Reiner, A., Panageas, K., Kryza-Lacombe, M., ... others (2017). Prognostic awareness, prognostic communication, and cognitive function in patients with malignant glioma. *Neuro-oncology*, *19*(11), 1532–1541. 34

Dronkers, E. A., Hoesseini, A., de Boer, M. F. & Offerman, M. P. (2018). Communication of prognosis in head and neck cancer patients; a descriptive qualitative analysis. *Oral oncology*, *84*, 76–81. 33, 34, 36

Eijgelaar, R., De Witt Hamer, P. C., Peeters, C. F., Barkhof, F., van Herk, M. & Witte, M. G. (2019). Voxelwise statistical methods to localize practice variation in brain tumor surgery. *PloS one*, *14*(9), e0222939. 6

Eijgelaar, R. S., Bruynzeel, A. M., Lagerwaard, F. J., Müller, D. M., Teunissen, F. R., Barkhof, F., ... Witte, M. G. (2018). Earliest radiological progression in glioblastoma by multidisciplinary consensus review. *Journal of neuro-oncology*, *139*(3), 591–598. 6

Emanuel, E. J. & Wachter, R. M. (2019, jun). Artificial intelligence in health care. *JAMA*, *321*(23), 2281. Retrieved from https://doi.org/10.1001/jama.2019.4914 doi: 10.1001/jama.2019.4914 35

Fagerlin, A., Rovner, D., Stableford, S., Jentoft, C., Wei, J. T. & Holmes-Rovner, M. (2004, may). Patient education materials about the treatment of early-stage prostate cancer: A critical review. *Annals of Internal Medicine*, *140*(9), 721. Retrieved from https://doi.org/10.7326/0003-4819-140-9-200405040-00012 doi: 10.7326/0003-4819-140-9-200405040-00012 35

Fagerlin, A., Wang, C. & Ubel, P. A. (2005, jul). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making*, *25*(4), 398–405. Retrieved from https://doi.org/10.1177/0272989x05278931 doi: 10.1177/0272989x05278931 35

Feng, X., Tustison, N. J., Patel, S. H. & Meyer, C. H. (2020). Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in Computational Neuroscience*, *14*, 25. 1, 4

Fritz, L., Dirven, L., Reijneveld, J., Koekkoek, J., Stiggelbout, A., Pasman, H. & Taphoorn, M. (2016, nov). Advance care planning in glioblastoma patients. *Cancers*, *8*(11), 102. Retrieved from https://doi.org/10.3390/cancers8110102 doi: 10.3390/cancers8110102 33

Gal, Y. & Ghahramani, Z. (2016). *Dropout as a bayesian approximation: Representing model uncertainty in deep learning.* 31

García, J. F., Spatharou, A., Jenkins, J. & Hieronimus, S. (2020, Mar). *Transforming healthcare with ai.* Retrieved from https://eithealth.eu/wp-content/uploads/2020/03/EIT-Health-and-McKinsey_Transforming-Healthcare-with-AI.pdf 35, 36

Gehan, E. A. & Walker, M. D. (1977). Prognostic factors for patients with brain tumors. *Natl Cancer Inst Monogr*, *46*, 189–195. 1

Gensheimer, M. F. & Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, *7*, e6257. 1, 2, 4, 14, 15

Gerds, T. A. & Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, *48*(6), 1029–1040. 16

Global strategy on human resources for health: Workforce 2030. (n.d.). World Health Organization. Retrieved from https://www.who.int/health-topics/health-workforce 36

Goel, M. K., Khanna, P. & Kishore, J. (2010). Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, *1*(4), 274. 4, 5

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... others (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. 3

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48. 3

Haavard Kvamme, N. S., Brian Hart. (2020). *Pycox.* https://github.com/havakv/pycox. GitHub. 2

Halkett, G. K., Lobb, E. A., Oldham, L. & Nowak, A. K. (2010). The information and support needs of patients diagnosed with high grade glioma. *Patient education and counseling*, *79*(1), 112–119. 34

Hammoud, M. A., Sawaya, R., Shi, W., Thall, P. F. & Leeds, N. E. (1996). Prognostic significance

of preoperative mri scans in glioblastoma multiforme. *Journal of neuro-oncology*, *27*(1), 65–73. 1

Hao, J., Kim, Y., Mallavarapu, T., Oh, J. H. & Kang, M. (2018, dec). Cox-PASNet: Pathway-based sparse deep neural network for survival analysis. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. Retrieved from https://doi.org/10.1109/bibm.2018.8621345 doi: 10.1109/bibm.2018.8621345 1, 30

Hara, K., Kataoka, H. & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6546–6555). 17

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, S., Ali, K., ... Nabeel, M. (2020). Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. *arXiv preprint arXiv:2004.01275*. 36

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, *42*, 54–56. 5

Jorgensen, T. (2019). Digital skills. *Where universities matter. Learning & Teaching Paper (EUA)*. 36

Kabir, M. (2019). Does artificial intelligence (ai) constitute an opportunity or a threat to the future of medicine as we know it? *Future healthcare journal*, *6*(3), 190. 35

Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 1–62. 3

Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., ... others (2007). Long-term survival with glioblastoma multiforme. *Brain*, *130*(10), 2596–2606. 34

Kvamme, H. & Borgan, Ø. (2019a). The brier score under administrative censoring: Problems and solutions. *arXiv preprint arXiv:1912.08581*. 3, 16, 31

Kvamme, H. & Borgan, Ø. (2019b). Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*. 2, 9, 14, 15, 21, 31

Lagarde, S. M., Franssen, S. J., van Werven, J. R., Smets, E. M., Tran, T. K., Tilanus, H. W., ... van Lanschot, J. J. B. (2008). Patient preferences for the disclosure of prognosis after esophagectomy for cancer with curative intent. *Annals of surgical oncology*, *15*(11), 3289–3298. 33

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J. & Zhai, G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, *7*(1), 1–8. 1, 2, 3, 29

Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-second aaai conference on artificial intelligence*. 2

Lin, C., Lieu, A., Lee, K., Yang, Y., Kuo, T., Hung, M., ... others (2003). The conditional probabilities of survival in patients with anaplastic astrocytoma or glioblastoma multiforme. *Surgical neurology*, *60*(5), 402–406. 4

Liu, R., Hall, L. O., Goldgof, D. B., Zhou, M., Gatenby, R. A. & Ahmed, K. B. (2016). Exploring deep features from brain tumor magnetic resonance images via transfer learning. In *2016 international joint conference on neural networks (ijcnn)* (pp. 235–242). 3, 4, 29

Liu, S., Zheng, H., Feng, Y. & Li, W. (2017). Prostate cancer diagnosis using deep learning with 3d multiparametric mri. In *Medical imaging 2017: computer-aided diagnosis* (Vol. 10134, p. 1013428). 1, 3

Liu, Z., Sun, Q., Bai, H., Liang, C., Chen, Y. & Li, Z.-C. (2019, sep). 3d deep attention network for survival prediction from magnetic resonance images in glioblastoma. In *2019 IEEE international conference on image processing (ICIP)*. IEEE. Retrieved from https://doi.org/10.1109/icip.2019.8803077 doi: 10.1109/icip.2019.8803077 29

Maddox, T. M., Rumsfeld, J. S. & Payne, P. R. O. (2019, jan). Questions for artificial intelligence in health care. *JAMA*, *321*(1), 31. Retrieved from https://doi.org/10.1001/jama.2018.18932 doi: 10.1001/jama.2018.18932 34, 35

McDougall, R. J. (2019). Computer knows best? the need for value-flexibility in medical ai. *Journal of medical ethics*, *45*(3), 156–160. 36

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... others (2020). International evaluation of an ai system for breast cancer screening. *Nature*, *577*(7788), 89–94. 36

McLendon, R. E. & Halperin, E. C. (2003). Is the long-term survival of patients with intracranial glioblastoma multiforme overstated? *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *98*(8), 1745–1748. 1

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., ... others (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, *34*(10), 1993–2024. 3

Miller Jr, R. G. (2011). *Survival analysis* (Vol. 66). John Wiley & Sons. 1

Nie, D., Zhang, H., Adeli, E., Liu, L. & Shen, D. (2016). 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *International conference on medical image computing and computer-assisted intervention* (pp. 212–220). 1, 2, 3, 29, 30

Oerlemans, A. J., Knippenberg, M. L. & Olthuis, G. J. (2020). Learning shared decision-making in clinical practice. *Patient Education and Counseling*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0738399120305322 doi: https://doi.org/10.1016/j.pec.2020.09.034 33

Ostrom, Q. T., Cioffi, G., Gittleman, H., Patil, N., Waite, K., Kruchko, C. & Barnholtz-Sloan, J. S. (2019). Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2012–2016. *Neuro-oncology*, *21*(Supplement_5), v1–v100. 1

O'Neill, T. J., Xi, Y., Stehel, E., Browning, T., Ng, Y. S., Baker, C. & Peshock, R. M. (2020). Active reprioritization of the reading worklist using artificial intelligence has a beneficial effect on the turnaround time for interpretation of head cts with intracranial hemorrhage. *Radiology: Artificial Intelligence*, e200024. 35

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf 17

Pereira, S., Pinto, A., Alves, V. & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, *35*(5), 1240–1251. 3

Price, M., Cameron, R. & Butow, P. (2007, dec). Communicating risk information: The influence of graphical display format on quantitative information perception—accuracy, comprehension and preferences. *Patient Education and Counseling*, *69*(1-3), 121–128. Retrieved from https://doi.org/10.1016/j.pec.2007.08.006 doi: 10.1016/j.pec.2007.08.006 35

Reinhold, J. C., Dewey, B. E., Carass, A. & Prince, J. L. (2019). Evaluating the impact of intensity normalization on MR image synthesis. In *Medical imaging 2019: Image processing* (Vol. 10949, p. 109493H). 11, 12

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L. & Yu, Y. (2019). Deep recurrent survival analysis. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 4798–4805). 1

Rimmer, A. (2019). *Technology will improve doctors' relationships with patients, says topol review.* British Medical Journal Publishing Group. 36

Samek, W., Wiegand, T. & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*. 30, 32

Sapci, A. H. & Sapci, H. A. (2020). Artificial intelligence education and tools for medical and health informatics students: systematic review. *JMIR Medical Education*, *6*(1), e19285. 36

Scott, G. & Gibberd, F. (1980). Epilepsy and other factors in the prognosis of gliomas. *Acta Neurologica Scandinavica*, *61*(4), 227–239. 1

Shortliffe, E. H. & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama*, *320*(21), 2199–2200. 33, 34, 35

Soffietti, R. & Chio, A. (1989). Prognostic factors in cerebral astrocytic gliomas. 1

Stead, W. W. (2018, sep). Clinical implications and challenges of artificial intelligence and deep learning. *JAMA*, *320*(11), 1107. Retrieved from https://doi.org/10.1001/jama.2018.11029 doi: 10.1001/jama.2018.11029 34

Stevens, E., Antiga, L. & Viehmann, T. (2020). *Deep learning with pytorch*. Manning. 13

Sun, L., Zhang, S. & Luo, L. (2018). Tumor segmentation and survival prediction in glioma with deep learning. In *International miccai brainlesion workshop* (pp. 83–93). 1

Susskind, R. E. & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press, USA. 35

Tran, V.-T., Riveros, C. & Ravaud, P. (2019). Patients' views of wearable devices and ai in healthcare: findings from the compare e-cohort. *NPJ digital medicine*, *2*(1), 1–8. 33, 36

Verghese, A., Shah, N. H. & Harrington, R. A. (2018). What this computer needs is a physician: humanism and artificial intelligence. *Jama*, *319*(1), 19–20. 34

Visser, M., Müller, D., van Duijn, R., Smits, M., Verburg, N., Hendriks, E., ... others (2019). Inter-rater agreement in glioma segmentations on longitudinal mri. *NeuroImage: Clinical*, *22*, 101727. 6

Walker, M. D., Green, S. B., Byar, D. P., Alexander Jr, E., Batzdorf, U., Brooks, W. H., ... others (1980). Randomized comparisons of radiotherapy and nitrosoureas for the treatment of malignant glioma after surgery. *New England Journal of Medicine*, *303*(23), 1323–1329. 1

Wesseling, P. & Capper, D. (2018). Who 2016 classification of gliomas. *Neuropathology and applied neurobiology*, *44*(2), 139–150. 1

Wittenberg-Lyles, E. M., Goldsmith, J., Sanchez-Reilly, S. & Ragan, S. L. (2008). Communicating a terminal prognosis in a palliative care setting: deficiencies in current communication training protocols. *Social science & medicine*, *66*(11), 2356–2365. 34

Yoon, H. G., Cheon, W., Jeong, S. W., Kim, H. S., Kim, K., Nam, H., ... Lim, D. H. (2020, aug). Multi-parametric deep learning model for prediction of overall survival after postoperative concurrent chemoradiotherapy in glioblastoma patients. *Cancers*, *12*(8), 2284. Retrieved from https://doi.org/10.3390/cancers12082284 doi: 10.3390/cancers12082284 30

Zadeh, S. G. & Schmid, M. (2020). Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1, 2, 31

Zhao, L. & Jia, K. (2016). Multiscale cnns for brain tumor segmentation and diagnosis. *Computational and mathematical methods in medicine*, *2016*. 3