Investigating Brain Cancer Survival with Machine Learning

Callum Biggs O'May

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2019

Abstract

Brain cancer has one of the worst prognoses of any type of cancer, but the factors affecting survival are poorly understood. This project investigates a new hand-curated dataset of patients presenting at a clinic with brain cancer, which includes demographic information, clinical information, genetic tests, tumour characteristics, and treatment information. We utilise machine learning techniques to interrogate the clinical consensus on how these factors influence one another and which are important for survival. We also explore options for how to impute missing data in the dataset. We demonstrate limitations in the Cox proportional hazards model, and show that a random forest model can perform better at survival modelling. Finally we examine whether the treatment protocols are optimal for patient survival.

Acknowledgements

My thanks go to Jacques Fleuriot, for his advice and guidance on the technical aspects of the research; and to Paul Brennan, for his patient explanation of the medical side of our work.

Table of Contents

Chapter 1

Introduction

1.1 Brain cancer

Brain cancer is one of the most deadly types of cancer [1], and the prognosis for the most common type – glioma – is poor[1]. For glioblastomas (GBM), the most common and most dangerous type of glioma, the median survival is only 14.6 months, even when undergoing surgery, chemotherapy and radiotherapy [4].

The Stupp protocol [4], introduced in 2005, defined the current clinical approach. Although subsequent studies appear to demonstrate that its introduction has improved patient survival [5], patient survival times are still extremely varied, and a large portion of this variance is unexplained. Identifying the factors which affect survival could help optimise treatment regimens; allow doctors to tailor treatment better to the patient and deliver better care; and motivate the development of novel treatments. The treatment protocols for many forms of cancer have undergone major improvements in recent years, but for brain cancer has remained relatively unchanged in the past 15 years. This was part of the motivation for this project.

1.2 Survival analysis

Survival analysis is the study of predicting how long a person is likely to live, given some information about the patient [6]. In its narrowest sense it consists of providing a point estimate for this expectation, but it can also involve estimating a *survival function*. This is a function of time which expresses, for each time t, the likelihood that a

[1]This is largely due to the functional importance of the brain, but may also be partly due to high heterogeneity within the tumours making them resistant to treatment [2, 3]

patient will survive until at least time t. We introduce the main techniques used for this in Chapter 2.

1.3 Hypotheses and objectives

This project investigates an unpublished, hand-curated dataset which was collected by Dr Paul Brennan, Senior Clinical Lecturer and Honorary Consultant Neurosurgeon at the University of Edinburgh, who collaborated on this project. The project is exploratory and open-ended, with a goal of extracting any insights possible from the dataset. That being said, the core of the research will be along three main lines of enquiry:

1. Are there dependencies among the features, and if so, what are they? Do these support previous research? Do they impact our view of treatment?

2. Which factors predict survival, and how? How do the factors interact to affect survival? Are the assumptions of the Cox proportional hazards model (see Section 2.2.2) justified? Can we construct a model which performs better than the Cox?

3. Are the treatment decision made by the clinic optimal?

1.4 Thesis structure

Chapter 2 gives a review of previous related work, focusing on survival modelling of brain cancer patients. Chapter 3 consists of a discussion of the dataset, and details of the data imputation (filling in of missing values). Chapter 4 explains the methods and results of this work. Chapter 5 discusses future work and concludes the thesis.

Chapter 2

Background

We first introduce the techniques used for imputation in Section 3.3. Note that we also experimented with neural networks (with a single hidden layer) for imputation, but found that they performed poorly. This was likely due to the small size of the dataset, the few dependencies among features, and the large number of parameters - see the beginning of Section 3.3 for discussion.

2.1 Imputation techniques

K-nearest neighbours *K-nearest neighbours* (KNN) is a simple and very commonly used non-parametric algorithm in machine learning [7]. When predicting the target value or category of a feature vector $\boldsymbol{x}$, we simply look at the k 'nearest' points (according to the some distance metric), and take the (possibly weighted) mean (for regression) or mode (for classification) of the target values of these points. We consider using both unnormalised features and normalised features for the KNN, as using unnormalised features imposes an implicit weighting on the features. One interesting extension might be to either learn this weighting, or design it by hand using clinical expertise.

Linear model We also consider a simple model where for feature vector $\boldsymbol{x} = (x_1, \ldots, x_d)$ and target variable y, we have $y = \sum_{i+1}^{d} a_i x_i$ for some parameters $\{a_i\}_{i \leq d}$. One potential downside (along with the strong assumption of linearity) is that the predictions are unbounded for extreme input values.

Logistic regression Logistic regression can be considered the adaptation of the linear regression model to classification. For a binary target variable, we have

$$P(y = c_0 \mid \boldsymbol{x}) = \sigma(\sum_{i+1}^{d} a_i x_i)$$

where $\sigma(\cdot)$ is the logistic sigmoid function. This can be generalised to a multi-class classification problem either by fitting a 'one-vs-rest' logistic regression for each class and combining them, or with a genuine multinomial logistic regression. See [7] for details of the latter.

2.2 Survival analysis techniques

We now introduce two of the most common approaches to survival analysis, as well as the machine learning models that will be used in this research.

2.2.1 Kaplan-Meier estimation

Kaplan-Meier (KM) estimation is a simple, empirical approach to estimating survival functions. It looks at a cohort (or sub-cohort) of patients and estimates their group survival function; it does not predict individual survival. The construction is straightforward: at each time step, we calculate the ratio of patients alive at the previous step who survived this step. Then the probability of survival up to this step is defined recursively as the probability of survival up to the previous step multiplied by this ratio. That is, letting $S(t)$ be the survival function, N_i be the number of patients alive at the start of time step i, and n_i the number that died during time step i, we have $S(t) = \prod_{i=0}^{t}(1 - \frac{n_i}{N_i})$. We will explore this approach in detail in Section 4.1.

2.2.2 Cox proportional hazards model

The Cox proportional hazards model is a semi-parametric multivariate regression model common in the field of survival analysis [8]. The model is expressed in terms of a *hazard function* which describes the risk to an individual as a function of time. Suppose we have a k-dimensional feature vector $\boldsymbol{x}$, and let $h(t, \boldsymbol{x})$ be the risk to the individual $\boldsymbol{x}$ at time t. Then the model supposes that $h(t, \boldsymbol{x}) = h_0(t)e^{\boldsymbol{x}^T \boldsymbol{\beta}}$ for some vector of parameters $\boldsymbol{\beta}$, where $h_0(t)$ is the *baseline hazard* (varying with time). Thus the hazard at a given time is log-linear in the features. Cox demonstrated in his original paper [8] that

one need not specify the form of the hazard function in order to fit the parameters $\boldsymbol{\beta}$ through a partial likelihood method.

This model makes several strong assumptions, supposing that the features affect the hazard independently and linearly, and also that the multiplicative effect of any feature on the hazard function is constant through time. We will interrogate these assumptions in Chapter 4.

2.2.3 Decision trees and random forests

Decision trees were popularised within the field of machine learning as a learning algorithm in the mid-1980s through the work of Breiman et al. [9] and Quinlan [10]. Both formulations (CART and ID3 resp.) operate in a similar manner: the feature space undergoes an iterative binary partitioning, at the end of which each resultant region is assigned a constant value (or category). This corresponds to the creation of a binary tree where each non-terminal node represents a split, and each data point belongs to exactly one leaf node (which has an assigned value). Decision trees (under another name) were used to model survival as early as 1993 [11].

Although decision trees found much use, they have been shown to be significantly more effective when combined as an ensemble to form a *random forest*, an idea popularised by Breiman [12]. This is effective because averaging over many trees can reduce variance at minimal cost to the bias. Breiman also introduced *bagging* (bootstrap aggregating), which is the process of sampling from a training set with replacement to create new training sets. This is typically used when constructing random forests as it leads to a more diverse set of trees in the forest.

Ishwaran et al. introduced *random survival forests* which are a particular case of random forests able to deal with censored data: data where the final target variable - survival in our case - is not always fully complete [13]. This occurs when, for example, a patient is still alive, so 'time to death' is not known. These have found use in a variety of medical settings, and have been shown to perform better than the Cox model under some circumstances [14, 15, 16].

Random forest hyperparameters

The performance of a random forest is highly dependent on its hyperparameters. Among other things these determine the number of trees, the trees' dimensions, the optimiser, and the loss function. In the original paper [12], Breimen demonstrated a number of

key properties of random forests: (i) They are (relatively) immune to overfitting; (ii) Their strength depends on the strength of the trees in the forest and the dependence between them; and (iii) The performance of the forest is largely independent of the number of features chosen as candidates for each split. These will inform our understanding of hyperparameter choice.

The simplest hyperparameter is the number of trees. This has a major effect on the performance of the model, with the rule of thumb being adding trees reduces the error. Although random forests are resistant to overfitting, other work has shown that random forests can overfit for noisier datasets than Breimen tested on [17]. The main cost of more trees is computational, but this is not of major concern to us as the computational cost is still reasonable for fitting very large forests to a dataset the size of ours.

Another important feature of a random forest is the size and shape of its trees. There are several ways this can be controlled. It is possible to set the maximum allowed depth of a tree, so that if a tree reaches this depth, it will stop partitioning further. Tree size can also be controlled by setting the minimum number of samples required at a leaf node; this stops trees from separately fitting tiny subsets of the dataset. Previous work has shown that it is not always beneficial to grow trees with the smallest possible terminal nodes [18].

One of the defining elements of random forests is the fact that they select a random subset of features as candidate features for each split. The number of features selected as candidates can be controlled. Whether or not to use bootstrapping is another hyperparameter, as is the loss function with which to train.

Hyperparameter optimisation Given the many choices of hyperparameters and the lack of clarity of exactly how these affect model performance, we choose to optimise our hyperparameter selection. We do this through a grid search: for a set of possible hyperparameter choices, we build a model for each possible choice, evaluating their performance (on a held-out validation set) and choosing the set which perform the best. We use the same metric as our test metric (mean square error) to evaluate different settings. We choose the set of candidate hyperparameters by hand, as exhaustively searching the hyperparameter space is not feasible. Although some hand-engineering is taking place here, iterative experimentation ('narrowing down' good options) can relatively conclusively demonstrate that optimal settings have been found. We use a built-in routine in Scikit-learn to *cross validate*: we split the dataset into k portions, and in turn treat each portion as the validation set with the rest of the portions as the

training set. All the tree-based models undergo this procedure.

2.2.4 Extra Trees and AdaBoost

Extra-Trees (Extremely Random Trees) [19] is an algorithm closely related to random forests. However, whereas the trees in a random forest are grown by selecting a random subset of candidate features for each split and then computing the optimal split for each of these candidate features, Extra Trees selects a small number of random split values for each feature and chooses the best from these. This introduces an additional element of randomness to the training procedure which can introduce some bias in the model, but reduces the variance of its predictions as it leads to a more diversified forest. As Breimen described this improves the performance of the forest.

AdaBoost [20] was in fact introduced before Breiman's random forest paper. It was one of the first classifiers which used *boosting*, a method of combining weak models into stronger models. AdaBoost iteratively learns weak learners (in our case decision trees) while motivating its new trees to be effective at predicting the data points which so far have not been predicted well.

2.3 Existing work

Machine learning techniques have seen wide use for a variety of medical tasks including brain cancer classification [21], colorectal cancer modelling [14], location-agnostic cancer survival modelling [22], breast cancer diagnosis, recurrence and survival [23, 24, 25] and oral cancer prognosis [26]. Deep learning in particular has been used extensively in recent years, including for modelling brain cancer survival [1, 2, 16, 37]. However we restrict ourselves to a discussion of the uses of random forests in modelling brain cancer survival for the sake of brevity. They largely started to be applied to the field of survival analysis after Ishwaran's seminal paper [13].

Jain et al. use random survival forests to predict glioblastoma survival [15]. Their focus is on the non-enhancing region (NER) of the tumour: that is, the region which does not activate the constrast agent used in the imaging. They explore the hypothesis that features of the NER may be predictive of survival. The dataset consists of 45 patients, with information derived from magnetic resonance (MR) images, clinical features, and four genetic features. The clinical features include *Karnofsky performance*

score (KPS), which is a measure of the mental performance of the patient (see Section 3.2.1 for details); *extent of resection* (EOR), a measure of the portion of the tumour which is removed during surgery; age; and year of diagnosis (used as a proxy for quality of care). The authors use these to train Cox models and a random forest model and demonstrate that Karnofsky performance score (KPS) and extent of resection (EOR) are predictive of survival. They also show that several features of the NER (specifically whether it crosses the midline and the relative volume of blood in the region) are predictive of survival.

Descriptions of the random forest model and how it was trained are brief. The authors do state that they train a forest of 50,000 trees and require at least 3 samples at each leaf node, but how these values were chosen is not explained, and it is not clear whether these values were validated. The size of the forest in particular is somewhat surprising given the small size of the dataset (both in terms of the number of features and the number of patients).

The authors also do not describe the performance of the random forest in comparison to the Cox model (or indeed any model) which is required if any meaningful conclusions are to be made about the feature importances derived from the model. It is also not clear how the variable importances are calculated, although the authors do note that they use the randomSurvivalForest package in R which appears to use a permutation importance approach [13] (see Section 4.4 for a discussion). They also describe growing a single decision tree and report the features it learns. Given that single decision trees tend to show high variance, the is unlikely to be very meaningful.

The main difference between the dataset used by Jain et al. and the one we investigate in this research is that theirs is much smaller (around 10% the number of patients). Additionally, theirs consists only of the features from the MR imaging, several genomic features, and three clinical features; our dataset includes many demographic features, symptom features, and treatment features among others. Our dataset also includes information on two biological markers: MGMT methylation status and IDH status. MGMT (O[6]-methylguanine-DNA methyltransferase) is an enzyme whose activity can be affected by whether its promoter is methylated, and the state of this methylation is thought to be important for glioblastoma patients [4] – see Section 4.2 for a discussion. The IDH1 gene codes for the creation of an enzyme used in various biological processes. It is often found mutated in patients with glioblastomas, and patients with the mutation have been shown to have a significantly better survival outlook [27] – see the Appendix for further details. However Jain et al. do not include these in

their analysis.

Gittleman et al. investigate survival in glioblastoma patients using similar methods to ours [28]. Their final output is a *nomogram* – a simple graphical tool for estimating the relationship between several variables. They utilise two datasets of combined size $N = 1,354$, and the features included were age, gender, race, Karnofsky performance score, EOR and MGMT methylation.

The authors compared the Cox model, random survival forests, and recursive partitioning analysis (RPA) [29]. Interestingly, in addition to cross-validating the models on the first dataset for testing, they also tested them on the second dataset to evaluate how well they generalise across datasets. They treated survival as categorical, using the models to output probabilities of patient survival at 6, 12, and 12 months – this is in contrast to our research which treats survival as continuous. They found that the Cox model outperformed the random forest.

Their random forest model ranks age as the strongest predictor, followed by MGMT methylation, Karnofsky performance score, gender, and finally EOR. Given that much previous research has demonstrated the significance of extent of resection in predicting survival [30], it is surprising to see that their random forest model ranks EOR as the lowest importance variable. Indeed, their Cox model (which performs better than their RF model) does consider EOR to be a significant predictor, and a stronger predictor than age. This is particularly surprising given that the evidence seems to suggest that EOR is a significant predictor only above a threshold value, an effect which a random forest can model but a Cox cannot.

The authors also report that the RPA approach performs better than the random forest. RPA is an approach which grows decision trees and then chooses the largest tree for which the statistical confidence of the final splitting is above a threshold confidence level (we discuss other examples of this approach shortly). The fact that this method reportedly outperforms a random forest (a combination of many decision trees) is surprising, especially given that their optimal tree has just 6 non-leaf nodes. It is possible that the poor performance of the random forest is a result of the authors' implementation rather than the method itself. The only information about hyperparameters and training is that the forest has 1,000 trees, and that the model considers all variables as candidates at each split. Again there is no indication that they validated these choices, and no justification for the choices is given.

One of the major differences between the research by Gittleman et al. and our

study is the number of features considered – they restrict themselves to only 6 features while as we will see, we consider around 20.

A number of studies have used raw magnetic resonance imaging as input for a survival forest to investigate brain cancer survival. This is a fairly different approach to using clinical data, as the number of features is typically much larger. Chang, Ken et al. looked at patients with recurrent glioblastoma who had been treated with bevacizumab, a chemotherapy drug, and demonstrate that features extracted from MR images can be used to predict survival in glioblastoma patients [31]. Notably they learn a classifier (of 'high survival time' vs 'low survival time') and use this an input to a Cox model. This is in contrast to our research which treats survival as continuous and learns a regressor to explicitly predict survival time. Yang et al. also used random forests trained on features from MRI scans to predict survival, but they were exploring whether *texture features* were predictive of survival [32]. Although the remit of these studies and the approaches required are quite different, they demonstrate that raw MRI images can be used effectively as input for a random forest survival model. Such an approach could be a future direction for our work – see Chapter 5.

As discussed, *recursive partitioning analysis* is an approach which grows a decision tree – in essence it is a random forest with just 1 tree (and a different training approach). This is commonly used in medical literature as a classification or regression model, and a number of authors have used it for survival analysis. Sanai et al. investigate the role of extent of resection in survival for patients with glioblastoma [33]. Interestingly, they are critical of the seminal work by Lacroix which demonstrated that extent of resection is significant, but only if more than 89% of the tumour is removed [30]. The authors note a number of weaknesses in the statistical analysis of the paper (including an approach to defining cutoffs in variables which has been shown to lead to high rates of false positives) and aim to reinvestigate its claims. They conclude that a survival advantage can be seen with EOR greater than around 78%. RPA has also been employed to investigate brain metastases of other primary tumours. Li et al. use it to investigate how EOR affects survival in patients with brain metastases of primary lung tumours [34], and Gaspar et al. use it to demonstrate the importance of Karnofsky performance score and age for survival in patients with brain metastases [35].

Chapter 3

Data preprocessing and imputation

The code for this project was written in Python 3.7.2 [38] using Jupyter Notebooks [39]. We chose to utilise the Pandas library from the SciPy ecosystem [40]. We make use of Pandas and NumPy from the SciPy ecosystem [40], Matplotlib [41], Scikit-learn [42], PyTorch [43], and a survival analysis package, Lifelines [44].

3.1 Dataset exploration and missing data

First, a preliminary exploration of the dataset was carried out. This consists of 1,334 patient records with, notionally, 288 features for each patient. These include demographic features (e.g. sex, ethnicity); patient health features (e.g. alcohol consumption, history of cancer); symptom features (numbered in order of appearance); radiological analysis features (e.g. tumour type, size, location); and treatment features (e.g. type of surgery, radiotherapy dose, chemotherapy dose). Notably age is missing from this dataset - it was removed by the neurosurgeon for confidentiality reasons, although later discussion determined that this was unnecessary. A large number of the features pertain to symptoms or signs (symptoms are as reported by the patient, whereas signs are as observed by the clinicians), and there is some duplication; after removing features with the same value for all patients and the duplicated features, we are left with 156 features.

There is significant heterogeneity in the dataset: only 511 of the 1,334 patients have a glioblastoma (the most common type), with the next most common being metastasis, glioma, and meningioma. Figure 6.3 in the Appendix shows the types of tumour present in the dataset. Discussions with the consultant neurosurgeon determined that these different types have fairly different characteristics, which makes modelling more

difficult.

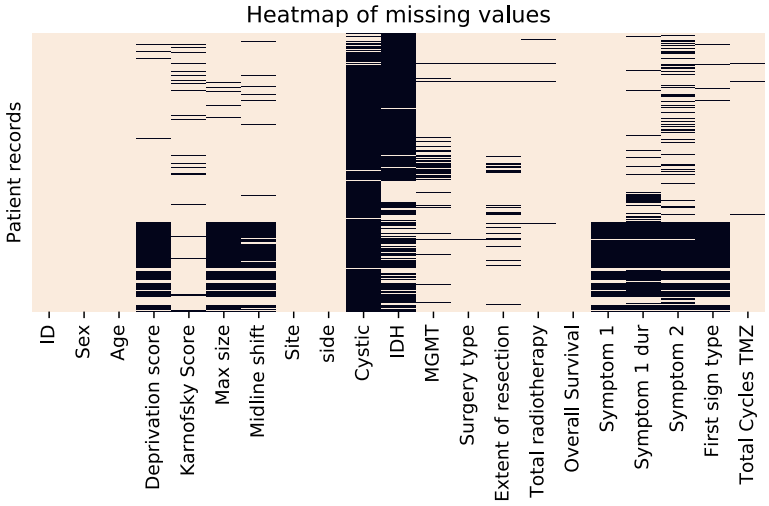
Looking then at the survival field (measured in days), we also see that a significant portion of patients (538) are still alive. In the context of survival modelling this is *censored data*: we do not know the exact survival time of these patients. This can be handled in various ways but requires some care. We also note that most features are not complete for all patients – most are around 70-90% complete (See Figure 6.4 in the Appendix). The key takeaways are that a number of features are missing many values; and almost all features are missing some values. It is important to note that an entry being empty does not necessarily imply that it is *missing*. For example, Symptom 3 (the third symptom the patient presents with) may be correctly empty if they only presented with two symptoms.

Given both the heterogeneity of the dataset and the level of missing data, after further discussion with the neurosurgeon a decision was made to look only at the patients with glioblastoma multiforme. This was supposedly a much more homogenous set of patients, and the neurosurgeon was able to provide a more complete set of records for these patients which had been used for analysis previously.

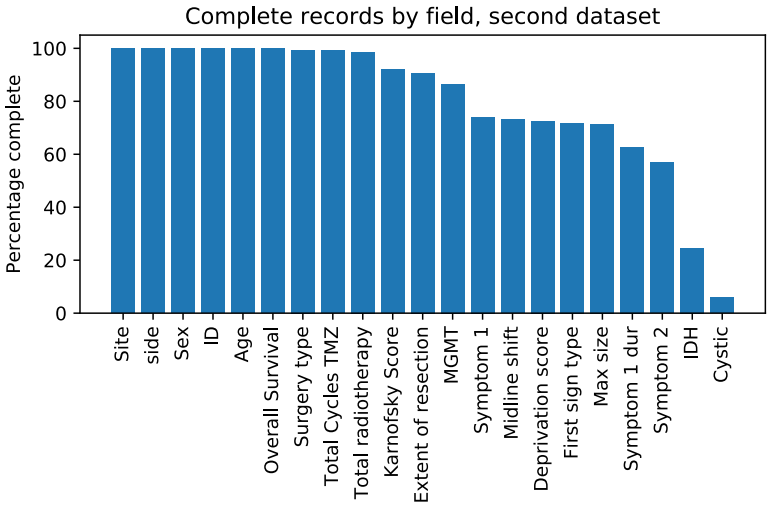
3.1.1 Second dataset

The provided second dataset did not initially include the symptom features, so these had to be imported from the first dataset according to patient ID. Given that each symptom was less complete than the last, we chose to include only the first two symptoms and the first sign. Figure 3.1b shows the by-feature completion for this dataset, where we see that it is much more complete than the first dataset. Figure 3.1a shows a heatmap of the missing values (where black represents a missing value). The patients are ordered chronologically along the y-axis. The specific features on the x-axis are not important currently, and we do not go through them in detail (but see the Glossary in the Appendix for an explanation of all terms if desired); this figure is intended to provide a high-level view of the level of 'completion' of the dataset. We note that there do appear to be patterns in the missing data: there is clearly a section of patients near the bottom which are missing a number of fields, several of which are important. Discussions with the consultant neurosurgeon determined that this was largely due to the data being compiled by different authors at different points – for example, the author who compiled the data since 2015 did not record the symptom features. There are also some features which were not collected from patients: the test for IDH mutation was

not routine until around 2013. Other than these, the neurosurgeon believes that the data missing elsewhere is missing at random, and does not follow any pattern. However there were a number of fields they were surprised to see with missing values, so further investigation would be worthwhile.



(a) Heatmap demonstrating the missing data for the second dataset. Apart from cystic status and IDH, most other fields are significantly more filled in than in the first dataset.



(b) Bar chart of second dataset showing the percentage of records complete for each field.

Figure 3.1: Graphs demonstrating the missing data for the second dataset.

This dataset is significantly smaller, with just 451 patients, all of whom had glioblastoma, and all of whom have passed away. There are 153 features, but a large number of these pertain to the specific details of the chemotherapy protocol. Ignoring these (bar the summary statistic of number of total cycles), we are left with 28 features, plus

the symptom features (of which we choose to include first symptom type and duration, second symptom type, and first sign.

The average age of the patients (at diagnosis) is 63 ± 13 years (where we report error bars of one standard deviation here and going forward). 57% of the patients are male. The most common location for the tumour to be found in is the frontal lobe (173), followed by the temporal lobe (127), and then parietal lobe (79). A significant portion of patients present with tumours in multiple lobes (59). Of the 451 patients, 342 received surgery of some kind; 318 received radiotherapy; and 197 received chemotherapy. This illustrates the typical clinical approach: surgery if possible, after which radiotherapy and chemotherapy [4]. Patients who do not receive surgery are very unlikely to receive other treatment, as they are likely to have been deemed too frail for the benefits to outweigh the risks.

The mean survival time 350 ± 404 days, which reflects the high variance in the dataset – attempting to explain this high variance is one of the goals of this thesis. The median is just 240 days, which is in line with previous estimates [4]. We note that the ratio between the mean and median of $ln(2)$ is exactly as expected for an exponential distribution. Figure 3.2 is a histogram of survival times (having removed extreme outliers for the sake of visualisation). The exponential shape is immediately noticeable, and on the graph is a fitted exponential curve. The next section discusses in finer detail the handling and preprocessing of the data, but we return to discussing the statistics of the dataset in Section 3.3.1.

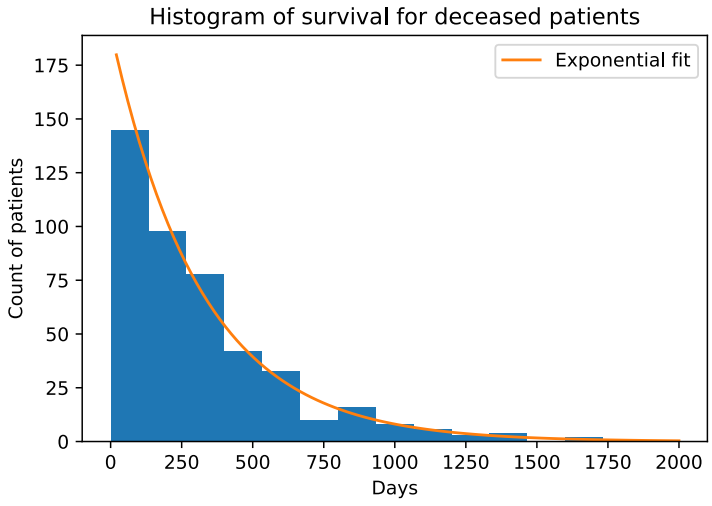


Figure 3.2: Histogram of survival times. Appears to have a roughly exponential shape.

3.2 Preprocessing

In its initial form the data was not suitable for fitting a survival model, so some preprocessing was required. Since it was imported from a CSV file, all entries were initially formatted as strings. However the features were implicitly either categorical (encoded as integers), continuous, or genuinely strings (e.g. symptom descriptions). The provided codebook was used to determine which features belonged to which group.

A number of features had Unknown values for some patients in the dataset – these were set to NaN (i.e. missing) for continuous features, but for categorical features we set this as a new category. The reasoning for this is that there may be some latent pattern in the Unknowns (e.g. more likely when a patient presents to the clinic in urgent need of surgery), and keeping this information might allow our models to discover and utilise this. There were also None values for both continuous and categorical features. Discussions with the consultant neurosurgeon determined that, for the continuous features, None could safely be interpreted as 0. For the categorical features, it was again mapped to a new category.

A number of the models we will use require the feature space to be $\mathbb{R}^k$ for some $k \in \mathbb{N}$. For this reason, categorical features encoded as integers must be mapped to such a space. The standard approach is one-hot encoding (also known as dummy coding), where for a feature with n possible values, the feature is mapped to n new features, each of which take the value of an indicator function for a category. We employ this strategy here. One subtlety is that this will result in linearly dependent columns, creating a singular (and thus non-invertible) feature matrix. This causes immediate problems for linear algebra routines, so we drop one of the columns from the new feature matrix. This is not a loss of information (as long as every model can fit a 'bias') since the dropped category becomes the base case.[1]

We also choose to drop one-hotted columns which have very few (normally < 5) 1s. Very high collinearity can make models much harder to fit, and assuming that the training observations come from the same distribution as the test observations, given that these categories were so rare, the model loses very minimal predictive power by dropping them. During imputation, we also explore whether normalisation improves performance. For the survival modelling, we always normalise.

Finally we also shuffle the data before splitting into train and test sets. There are two closely related reasons for this, both stemming from the fact that our patient

[1] See http://www.statsmodels.org/dev/contrasts.html for a discussion of this approach.

records are stored in chronological order. Firstly, there could be temporal trends in the data; it could be that, for example, the standard of care has improved over the timespan of the dataset, so recent patients have a better expected survival when all features are kept constant. Although the collaborating neurosurgeon considers this unlikely, we choose to eliminate this possibility in case such a change has gone unnoticed. If it were the case, we would not be able to assume that our training set and our test set come from the same distribution, which would mean that the learning of our model may not generalise (to the test set) well. The second reason is that we do not want our models to be able to exploit the temporal ordering within the training or test set (e.g. if the two patients before this survived for a long time, this one is likely to survive for a long time too) – the artificial intelligence field abounds with examples of algorithms learning some artificial structure in the data rather than genuine features of the observations [46].

3.2.1 Karnofsky performance score

There are several features which could be interpreted as continuous or categorical. Karnofsky Performance Score is a measure of performance status – how able the patient is to 'carry on normal activity and to work' – introduced in 1948 [47]. Figure 6.5 in the Appendix shows the definitions in the original paper. It ranges from 0-100, but many authors only assign scores of multiples of 10. This was the case for this dataset in particular, so it could have made sense to treat it either as continuous or as an 11-class variable. Treating it as continuous imposes additional structure on the variable in the form of an ordering (and, in fact, an evenly-spaced matric). Thus the relationship between scores of 90 and 80 is in some sense the same as the relationship between scores of 80 and 70. By treating it as an unordered categorical variable, one loses this understanding. Although the clinicians involved in this dataset never assign values other than multiples of 10, there is still an implicit understanding of the variable as having an ordering, and treating it as a categorical variable loses this. Additionally, treating it as continuous allows for simpler and better analysis; in a model which, for example, reports the weighting/importance of each variable, a single such weighting of the variable is easily understood, whereas a weighting for each possible score is much less easily interpretable.

That being said, this 'linear' understanding of the variable is in fact a constraint on the model. One could imagine the assumption of the even spacing to be violated

(e.g. the difference between 80 and 90 is much less than the difference between 90 and 100), or even, in the most extreme case, the assumption of the ordering (e.g. scores of 40 and 50 are indistinguishable and are assigned at random). To some extent, the former can be accounted for with non-linear models: since they do not assume a linear relationship between the variable and the outcome, it is possible to model such 'uneven' orderings. Some work has been done to explore *ordinal variables* [48] – that is, categorical variables with an ordering imposed – but such an approach was beyond the scope of this project. As a result, it was treated as continuous to preserve this structure, noting that this is in line with other research [33]. A question to answer in future work could be whether this choice affects the performance of the survival models, and how.

This feature is also interesting due to being subjectively evaluated: there is no clinical test for Karnofsky score. As such, one could reasonably suppose that there could be structural differences in the scores allocated by different doctors (e.g. some doctors giving higher scores than others). Unfortunately the information in this dataset did not permit this analysis, but an interesting piece of further work could be to investigate this phenomenon, and possible means of correcting for it.

One other interesting note with regards to this variable is that, at least in the original definition, a score of 100 is defined as 'Normal; no complains; no evidence of disease'. One would imagine, then, that someone who has been referred to an oncology clinic is very unlikely to have this score, since they must have had at least some symptoms for them to go to a doctor and be referred. However, in fact 100 is the second most common score – 103 of the 451 patients (23%) are reported as having a score of 100. This suggests that the definitions have at the very least changed since its introduction, and are likely to be somewhat imprecise. We also see that, although scores as low as 20 are present in the database, scores below 50 are extremely rare, with only 11 such patients.

3.2.2 Extent of resection

The other feature which could have been interpreted as categorical or continuous was extent of resection: the percentage of the tumour that is cut out during surgery. This is encoded categorically in the dataset (100%, 90-99% etc.). According to similar reasoning as with KPS (to keep the ordering embedded in feature representation), we choose to map this to numeric values and treat it as continuous. We also then have the choice of what values to map to. One obvious choice is the midpoint of the range -

i.e. for a patient belonging to the 90-99% category, map to 95%. Another is to map randomly into the range (in fact even then one can choose what distribution to draw from, although uniformly would likely be preferable). However randomly mapping adds a noise element to the feature which may worsen performance later. Although not pursued as part of this project, this could also be an interesting question to answer empirically. We choose to map to the midpoint of each range as previous researchers have chosen to do [33].

3.3 Data imputation

Typically machine learning approaches are more dependent on the size of the dataset than traditional data analysis. This is largely due to the fact that machine learning models tend to have many more parameters than a traditional model, which is part of why they are more flexible and can fit more complex functions/processes. A model with more parameters is likely to be more prone to overfitting, and overfitting is a much bigger problem with smaller datasets [7].

In some fields - for example financial services - there is so much data available that this is unlikely to be a problem. Many datasets number in the millions or billions, and working with them provides the opportunity to fit highly complex models with the confidence that they are unlikely to overfit. Within the field of biomedicine and healthcare, however, datasets tend to be smaller. This is particularly true when studying rare or unusual diseases like brain cancer. The fact of its low incidence rate means it is impossible to collect large volumes of data. As a result, when using machine learning with such datasets one wants to make the most use out of the data that is there.

Missing data can be a problem for many models. Considering, for example, fitting a simple linear univariate model of $y = ax + b$ to data. This model has a closed form solution, obtained with relatively simple matrix operations. However, clearly these cannot be performed when the matrix is missing some values. Similarly, for gradient descent fitting, missing data can be a problem: it will not be possible to evaluate the gradient as a point which is missing some features. Even worse, when it comes to the point of prediction, the model will be useless for any input which is not complete. That being said, some models can handle missing data to some extent. One example is K-nearest neighbours (introduced in Section 2.1), where, when input an observation with some subset of the features complete, it will compare only to those other points with a superset of those features complete. However, depending on the proportion of

missing features (and their importance), this could significantly affect performance.

As demonstrated already, the dataset for this project had relatively significant portions of data missing. Some of this was correctly 'missing' (i.e. should not have had a value), but even after accounting for this there are meaningful gaps in the data.

There are a number of more or less elegant solutions to this problem. The most simple is to ignore all data points (in our case patients) with any fields missing. However, under this approach there is clearly a compounding effect of having many fields with a portion of the entries missing. Looking at the fields we preliminarily wish to investigate (19 fields), there are only 28 patients with every field complete – clearly not enough to do meaningful data analysis on.

A second option is to *impute* the missing data: to fill it in with a reasonable guess as to what it could have been. The obvious benefit of this approach is that it allows one to keep all the information contained in every observation, even when they are missing some features. A typical choice for continuous variables is to 'mean-fill' – that is, for some observation missing feature i, impute the mean of feature i over all observations. The intuition for this approach is fairly clear: it is reasonable given few assumptions that the mean is our best possible guess. Indeed, the sample mean is an unbiased estimator of the expectation (population mean) of a random variable. To think about this another way, if one chooses to fill all missing values (for a given feature) with a constant value, the mean minimises the *square error*. For cases where outliers may significantly affect the mean, the median is another common choice (and in fact this minimises the *absolute error*). For categorical variables, the comparable approach is to fill with the mode (the most common category). We will consider mean-filling (and mode-filling) as our baseline for imputation. Note that we choose to mean-fill as most of our continuous features are constrained so outliers are less of a concern.

A more sophisticated approach might look to individually predict each missing value. We can frame the imputation as a regression problem: for an observation $\boldsymbol{x} = (x_1, \ldots, x_n)$ where x_i is missing, we want to calculate $\mathbb{E}[x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$. Similarly, for categorical features, we can frame it as a classification problem. A key assumption of this approach is that the features are not independent, or else it would not be useful to consider the other features in the imputation. Thus we first explore whether there are dependencies between the features; if not, then we do not expect to be able to outperform the mean/mode-filling baselines.

3.3.1 Feature dependencies

A simple first exposition of the dependencies is to look at the correlations between features. Figure 3.3 shows a matrix of the Pearson correlations for the continuous and binary features (we note that multi-class categorical features require a somewhat different treatment). We see that a number of pairs of features have significant correlations.

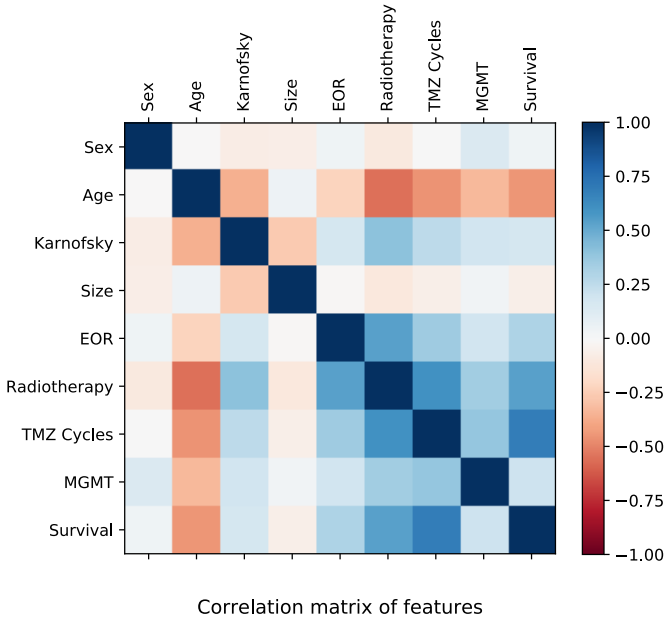


Figure 3.3: Heatmap showing Pearson correlation coefficients between continuous features.

We now explore these dependencies in more detail. This is necessary to understand the data better, but will also be valuable later when interpreting our survival models. One interesting point of focus is the features which are determined by the clinic, since it would be reasonable to suppose that these may well have dependencies with other features: clinicians make decisions consciously and unconsciously dependent on the patient in question.

Preliminarily we can see that, as shown in Figure 3.4d, the older a patient is, the less likely the clinicians are to operate. This is particularly clear for patients over the age of 80. However, the numbers are small for these ages (5 patients ≈ 80, and 1 patient ≈ 90), and all but one had Karnofsky score < 90 – given that the collaborating neurosurgeon has made clear that a low Karnofsky score makes surgery unlikely, this suggests that the decision not to operate could have been based on other factors than

their age. Previous research has shown that, absent other factors, attempting maximal possible resection is worthwhile even in the elderly [49].

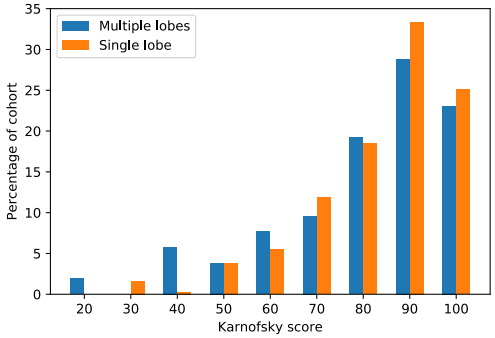
Figure 3.4c shows the relationship between extent of resection and age for all patients who were operated on and successfully had some of the tumour removed, showing a negative correlation[2]. The Pearson correlation coefficient is -0.23, significant at the 1% level. This suggests that, even once the decision has been made to operate, older patients are likely to have less of the tumour removed.

Discussion with the collaborating neurosurgeon made clear that although a surgeon will generally always remove as much as possible, they may take a different approach depending on, for example, the patient's age. There are a number of surgical procedures – operating while awake, the use of pink dye, testing the functionality of the area during surgery – which aid in achieving a higher EOR, and might be more likely to be used on a younger patient. Another factor could be that complications during surgery (e.g. bloody vessels or heart problems from anaesthetic) can lead to reduced extent of resection, and these problems could be more likely in older patients. Figure 3.4b shows that older patients are more likely to have a lower Karnofsky score (likely a result of complicating factors) – a correlation of -0.26, $p < 0.01$.

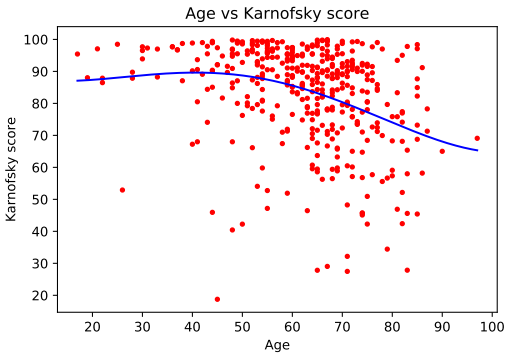
There also seems to be a relationship between EOR and Karnofsky score (Figure 3.4f), and this could well be along the same lines: not a direct causal link, but a correlation. This is in agreement with previous work by Simpson et al. [50], but for our results we do not see a significant p-value (0.16). One explanation for the relationship could be that a lower Karnofsky score could suggest that the tumour is in a more dangerous location, and could be harder to resect. Figure 3.4a seems to show that patients with tumours in multiple lobes are less likely to have a Karnofsky score of 90 or 100. Although the resectability of a tumour is highly dependent on the subtleties of its location, shape, size etc., the consultant neurosurgeon is clear that, all else the same, tumours which cross between multiple lobes are much harder to operate on.

Indeed, Figure 3.4e shows the relationship between EOR and the site of the tumour. It is clear that a significantly larger portion of tumours that are located in multiple regions have 0% EOR: that is, no attempt is made to debulk. The other noticeable trend is that tumours in multiple lobes seem to be less likely to have 90-99% removed. This is not obviously explained. However, the resectability of a tumour is more fine-grained that we can see in the dataset, as it varies significantly for tumours even within a given

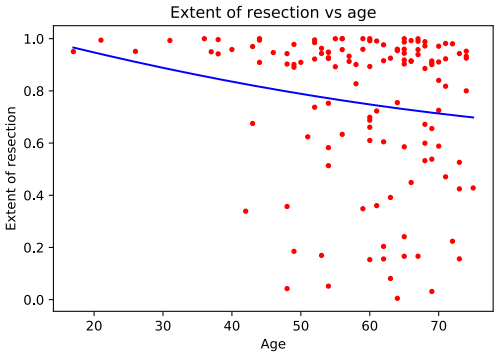
[2]In the dataset, EOR was encoded categorically as either 100%, 90-99%, 50-89% or $< 50\%$. For the sake of these visualisations, we mapped randomly to these ranges. The same has been done for Karnofsky score.



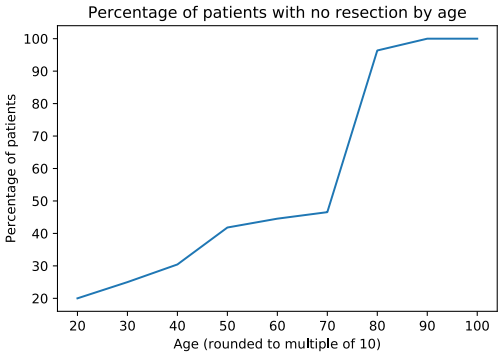
(a) Splitting Karnofsky score by whether the tumour is in a single or multiple lobes.



(b) Relationship between age and Karnofsky score.

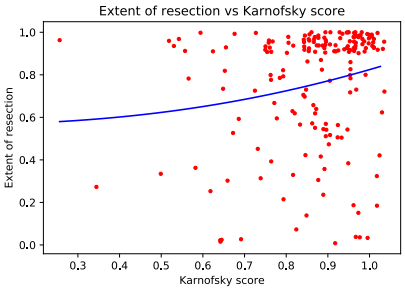


(c) Relationship between extent of resection and age.



(d) Percentage of each age decile with no resection.

(e) Relationship between extent of resection and site of tumour.



(f) Relationship between extent of resection and Karnofsky score.

Figure 3.4: Analysis of relationships between extent of resection and pre-operative factors.

lobe. The work by Simpson et al. [50] seems to suggest that extent of resection is not significantly dependent on the location of the tumour, but this work did not investigate

tumours which cross lobe boundaries.

Contrary to the suggestion of the collaborating neurosurgeon, extent of resection did not appear to vary significantly with the size of the tumour, even when split by location. The is likely due the the large range of resectability within each lobe, and the fact that EOR is dependent more on the finer-grained features of the tumour.

3.3.2 Imputation methodology

We now consider the task of imputation. As previously discussed we establish as our baseline mean-filling (or mode-filling for categorical variables). We recall that we can cast the problem as a supervised regression or classification problem. Suppose we have some N x D feature matrix Φ, where for some $0 < a \leq N$ and $0 < b \leq D$, we have $\Phi_{a,b}$ is missing. Then we can take $\Phi_{i \neq a, j \neq b}$ (that is, the feature matrix with row a and column b removed) as our 'training set', our target vector as $\Phi_{i \neq a, j = b}$ (column b), and our 'test set' (consisting of only a single point) as $\Phi_{i = a, j \neq b}$ (row a). We thus have a typical supervised regression/classification task with training set, target vector, and test set.

This is generalised in the obvious way to several missing data points. We note that, given the proportions of missing data, we must select somewhat carefully which features to use to predict others. For training a prediction model for a given feature, we need to select only the data points with all features completed which we want to use for the prediction. A naïve approach might be to use all the features, and just ignore points which are missing other features too when training/predicting. However the compounding effects of missing features means that we are left with very few datapoints for training, and we miss out on imputing most of those with the feature missing.

We thus must make judgements about which features to use to predict each others. The common-sense approach is to use features which are 'mainly filled' and for which the missing points are not highly correlated (positively or negatively) with the missing points of the feature we are predicting.

Another design choice involves constrained features. A number of continuous features in the database are restricted to certain values: either within a range (e.g. EOR must fall between 0 and 1) or certain discrete sets (e.g. as discussed Karnofsky score is always a multiple of 10 in the dataset). These present two separate challenges of how to deal with each. For the former, it is necessary to restrict the range of the output

to the range of the feature. Not to do so can cause significant problems when fitting models later, where gradients can go to infinity, or fail to converge. If the feature is constrained on an open interval like $(0, 1)$, a simple deterministic transformation can implicitly constrain our outputs: we can put out targets through a logit function to map them monotonically to an unbounded space, fit the models to these transformed targets, and then when testing, de-transform the predictions with a logistic sigmoid function. However, for several of our features, the range is closed (e.g. EOR falls within $[0, 1]$). In this case, there is no deterministic monotonic function with an unbounded range (consider where such a function would map the maximum of the domain). One simple if rather inelegant choice is to 'clamp' the output to the range: if a model outputs a value greater than the maximum of the range, set it to the maximum, and do similarly for outputs less than the minimum. We choose this solution here for ease. Note that clamping in this way will always reduce the error when compared to leaving the outputs unconstrained: since the target cannot fall above the maximum, a prediction which is greater than the maximum must necessarily have a greater error than the maximum itself.

For the second situation, we have a choice as to whether or not to 'snap' our predictions to the discrete set. We choose not to, as previous research appears to show that in some circumstances, rounding values creates more bias [51].

3.3.3 Evaluation

To evaluate the imputation, we must use different metrics for categorical and continuous variables. For continuous variables we choose the standard mean square error as our performance metric; for categorical variables, we choose simple accuracy (what % of predictions are correct). One downside of this metric is that, for highly imbalanced classes, a model can achieve good accuracy without learning much from the data. Consider, for example, a case where the observations come from class A with 95% chance, and class B with 5% chance (negative/positive for some disease is a common example). Then, without learning anything, a classifier can predict class A every time, and achieve 95% accuracy. One solution which we do not attempt here is to rebalance the classes by resampling the dataset. The precise method of resampling is itself a topic of research, and a number of solutions (to avoid overfitting/information loss) have been suggested [52, 53, 54, 55]. Fortunately most of the categorical fields in our dataset are not extremely unbalanced, so this should not be a major problem. The one exception

is IDH (with 104 patients being wildtype, and just 7 being mutated) – we will bear this in mind at our evaluation. At any rate we will be using mode-filling as our baseline, so any outperformance beyond the baseline will be evidence of genuine learning.

To evaluate each technique for each variable, we evaluate on the entire dataset with a 10-fold cross-validation. We also repeat the entire analysis several times for each feature to reduce the variance of the results and remove any chance that the performance was a result of the particular splitting of the dataset. We also cross-validate within each fold to select hyperparameters (k for K-nearest neighbours).

3.3.4 Results and discussion

Figure 3.5 shows the results of the imputation analysis. For the continuous features, KNN on normalised features is normally the best performing. As expected, the normalised KNN outperforms regular KNN – this confirms that the 'weighting' of the features which arises simply from their scaling is not an effective weighting for this task. As mentioned previously, constructing such a weighting could be an interesting extension of this work. We might expect the linear model to perform better in contrast to KNN. However, linear models can make extreme predictions, particularly when input a point with extreme values for which there were not many similar points in the training set. This can lead to very large errors on some predictions.

For some of the features – Karnofsky score and EOR in particular – we see significant outperformance of the baseline. This confirms our previous analysis that these features are reasonably strongly dependent with other features. For some – in particular deprivation score (a measure of the wealth of the area in which a patient lives) – we see the best algorithm barely outperforming the baseline.

For categorical features, logistic regression tends to be the best, but a reasonable number have KNN or mode-filling as the best. This is roughly as one might expect: in some senses, logistic regression is a more *intelligent* algorithm as it actually learns parameters: in this case, feature importances. KNN is a fairly blunt instrument in that it really does not learn anything from the dataset. Although logistic regression is somewhat similar to the linear model, it is less inclined to suffer from the problem of extreme predictions because its loss (i.e. 1 - accuracy) is capped. When interpreted probabilistically this is not the case; it can produce very confident, and wrong, predictions. But when using it as we are, with accuracy as our evaluation metric, the 'worst' it can do on any one input is to get it wrong. This is in contrast to the linear model which can

produce vastly wrong predictions which lead to a high error. For categorical features, as with continuous variables, we see that some imputations significantly outperform the baseline, while others either fail to beat it (IDH). Interestingly, it seems that the symptom and sign features are possible to impute meaningfully more accurately than the baseline. This confirms the hypothesis that these are related in a significant way to the other features.

For the interested reader, an additional piece of analysis regarding this imputation (investigating which features are important for imputing each other) is available in the Appendix.

Target variable	Mean-fill	Linear	KNN	KNN (norm)
Karnofsky score	234.6	249.3	199.7	**185.8**
Extent of resection	0.198	**0.058**	0.130	0.070
Max size	292.9	462.7	286.4	**278.3**
MGMT	285.4	354.2	283.0	**259.6**
Deprivation score	1.72	3.43	1.85	**1.71**
Symptom 1 duration	86.1	115.0	91.4	**80.9**

(a) Mean square error of continuous variable imputation.

Target variable	Mode-fill	Logistic regression	KNN	KNN (norm)
Symptom 1	26.3%	**34.9%**	24.6%	32.0%
Midline shift	29.7%	**42.8%**	28.6%	37.0%
IDH	**83.3%**	77.8%	**83.3%**	83.0%
Cystic	16.7%	**33.3%**	8.3%	10.0%
Symptom 2	10.0%	11.6%	10.0%	**16.0%**
Sign 1	30.5%	**39.8%**	28.1%	31.8%

(b) Accuracy of categorical variable imputation.

Figure 3.5: Tables demonstrating the results of the imputation

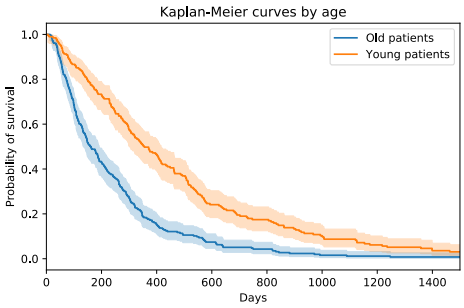
Chapter 4

Survival analysis

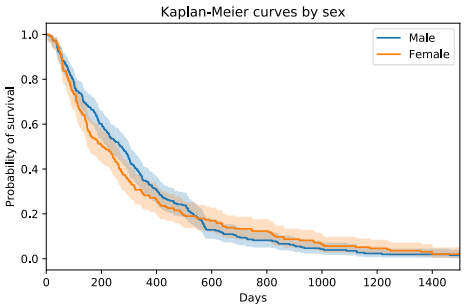
4.1 Kaplan-Meier curves

Kaplan Meier curves (introduced in Section 2.2.1) are particularly useful for illustrating *non-stationary* effects on survival, which otherwise may not be apparent. We first take a look at several Kaplan-Meier graphs to confirm our previous analysis and demonstrate some novel features of the dataset.

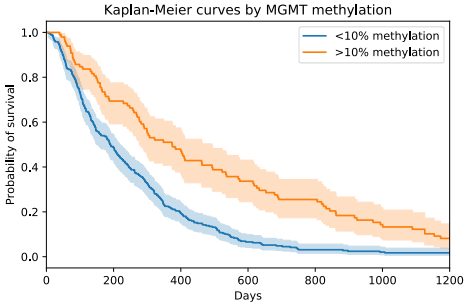
Figure 4.1 shows KM curves where the patients have been split by various features. Figure 4.1a confirms our analysis that age is a very significant factor in survival, and appears to demonstrate that it is largely time-invariant. Similarly Figure 4.1d reinforces the importance of Karnofsky score for survival – strikingly even a score of 90 vs 100 has a very noticeable impact. Figure 4.1c shows how much better survival outlook is for patients with high MGMT methylation (see Section 3.3.4 for discussion). Probably most interesting is Figure 4.1b which splits the patients by sex. Although males and females have overall a very similar survival outlook (mean of 352 days for males vs 349 for females), this KM curve demonstrates that this feature has a time-dependent effect. According to the fitted survival functions, males are marginally more likely to survive longer during the first ~ 600 days, whereas once patients have survived this long, females are marginally more likely to survive longer. This supports other research which suggests that men's survival advantage is only present during an initial period [56].



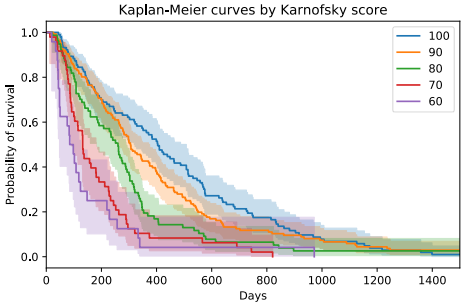
(a) Kaplan-Meier curves split by age



(b) Kaplan-Meier curves split by sex



(c) Kaplan-Meier curves split by MGMT methylation



(d) Kaplan-Meier curves split by Karnofsky score

Figure 4.1: Kaplan-Meier curves for patient groups stratified by various features. The utility of the curves beyond previous analysis is that they allow us to see how the effects of some features may be time-varying.

4.2 Cox proportional hazards model

We demonstrated in the previous section that the assumption of stationary effects of the Cox model is violated since sex has a time-dependent effect. The analysis in Section 3.3.1 and the success of the imputation demonstrates the violation of the assumption of independence. We will test this further in the next section. Nonetheless, we can fit the Cox to the data and analyse the model to learn what we can. Figure 4.2 shows the (directional) feature importances above a threshold value (the exact value is 0.02 but is not meaningful). A number of these results confirm our understanding: the total amount of tomozolomide is by far the strongest predictor, closely followed by high total dose of radiotherapy and KPS. We also see, as expected, that excision surgery is a positive predictor of survival; that tumours which cross into both sides of the brain or cause a larger midshift present a higher risk to survival; and that receiving no radiotherapy predicts poor survival. According to the collaborating neurosurgeon, these all agree with the clinical expectations.

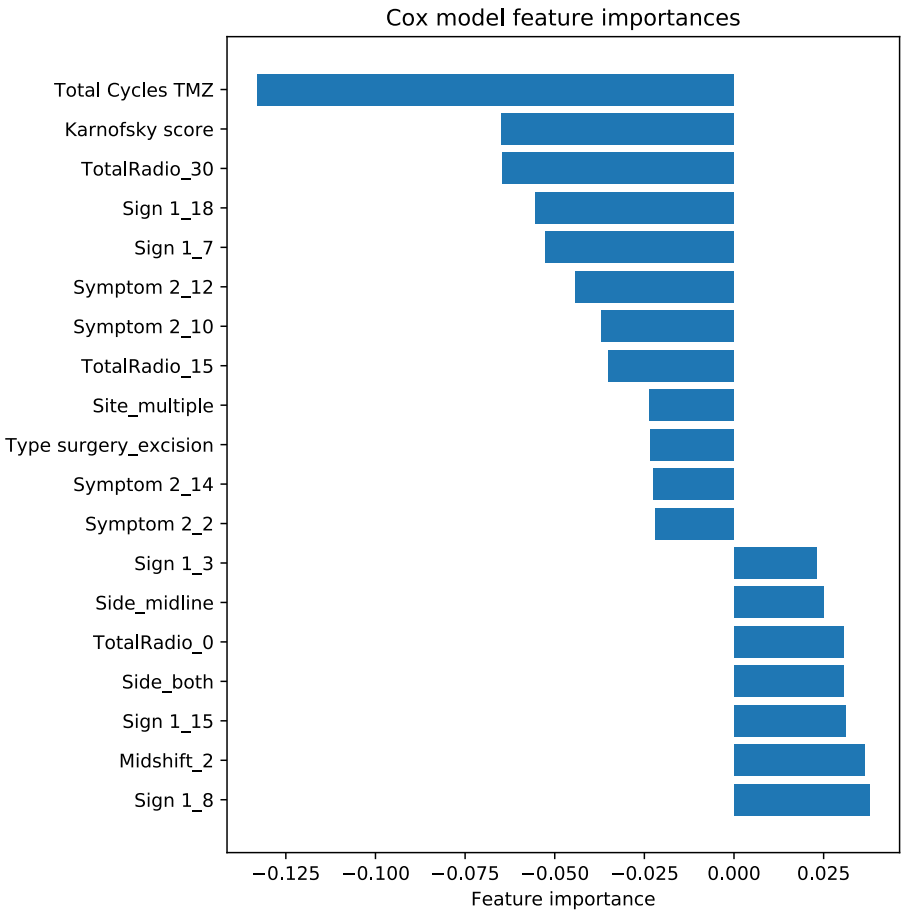


Figure 4.2: Bar chart showing the feature importances of the Cox model fitted to all the data. Generally this agrees with expectation; it is interesting to see several symptom classes of fairly high importance.

The most surprising result is that a tumour being present in multiple lobes appears to positively predict survival. One possible explanation for this is that this feature may be highly correlated with other features which are negative predictors of survival. The Cox model cannot model these dependencies, so it mistakenly thinks that the tumour being in multiple lobes is, in and of itself, a positive predictor, to 'balance out' the negative effects of the other highly correlated features. This is evidence of a shortcoming in the Cox model, and a violation of its assumptions. We also see a number of sign and symptom features. The strongest positive predictors of survival are drowsiness (sign 1 = 18), double vision (sign 1 = 7); the strongest negative predictors are unilateral numbness (sign 1 = 8) and problems walking (sign 1 = 15). However we suspect that these are really a case of the model 'overfitting'; for most of these symptoms there are very few patients who present with them, so the model can assign

a strong weight according to how long those few patients happen to have survived.

One of the benefits of the Cox model is that we have a confidence interval for the feature importances. Figure 4.3 shows the feature importances with 95% confidence interval error bars only when the p-value (against the negative hypothesis that the importance is 0) is less than 0.05. Here we see only a portion of the features remain – most interestingly, excision surgery is not considered important at this confidence level.

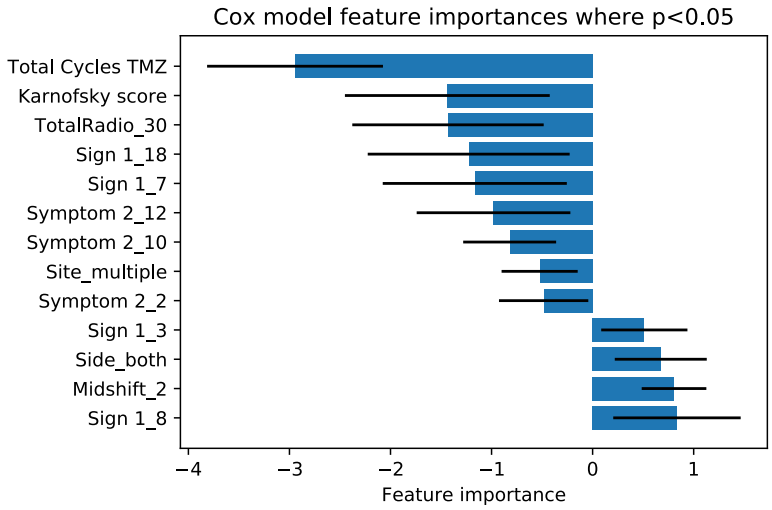


Figure 4.3: Bar chart showing the feature importances of the Cox model fitted to all the data. Generally this agrees with expectation; it is interesting to see several symptom classes of fairly high importance.

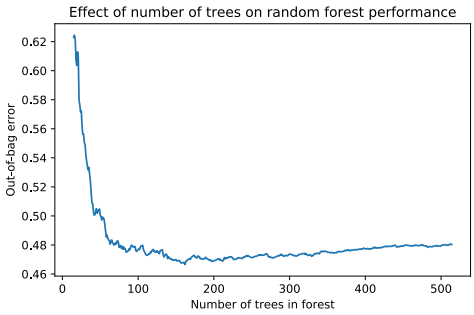
4.3 Methods and techniques

In this section we discuss the details of the models, as well as the experimental setup.

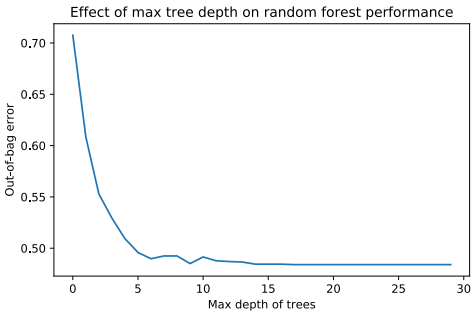
4.3.1 Random forest hyperparameter settings

Figure 4.4 shows the effects of different hyperparameter settings on fitting a random forest to our dataset. Note that in this case we are only varying one hyperparameter at at time rather than searching the multidimensional hyperparameter space. These make clear that for the number of trees, the maximum possible depth, and the number of features considered at each split, the performance initially increases as a function of these, before plateauing. Figure 4.4d shows how the minimum number of samples

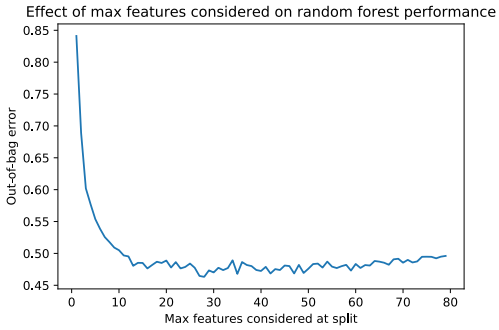
at a leaf node affects performance. It appears that requiring fewer samples is always better. During hyperparameter optimisation we find that the best hyperparameters for the random forest seem to be: (i) to use bootstrapping, (ii) to train at least 200 trees, (iii) to set the maximum depth to around 15, (iv) to consider around 10 features for each split, and (v) to have a very low (1-3) minimum number of samples at each leaf. These are similar for the random forest, Extra Trees, and AdaBoost.



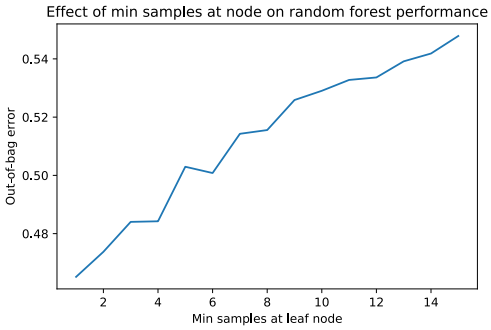
(a) Random forest performance as a function of number of trees.



(b) Random forest performance as a function of max possible tree depth.



(c) Random forest performance as a function of number of features considered at each split.



(d) Random forest performance as a function of the minimum of samples allowed at leaf nodes.

Figure 4.4: Exploring the effects of different hyperparameters on random forest performance

4.3.2 Experimental setup

We preprocess the data in much the same way as for the imputation, although here we normalise all features. This allows for better analysis as the coefficients in the Cox model are directly comparable to one another. One small point to note is that fitting the Cox model with the Lifelines package seems to produce infinite predictions when

there are highly collinear columns. This was particularly a problem with the symptom and sign features, where many classes had very few instances. As a result we need to be more aggressive in our removal of features than for the imputation.

Model comparison We aim to compare the tree-based models with the Cox model (as well as a KNN baseline), both in terms of their performance, and what they learn from the data. This will address the questions in the first line of enquiry as laid out in Section 1.3. The performance metric we use is root mean square error, since we are treating survival as continuous. Unlike in cases of binary classification, we do not create receiver operating characteristic curves.

For each iteration, we randomly select a portion (15%) of the data to be the test set. We then train each model on the rest and evaluate on the test set. For the tree-based models, we cross validate within the rest of the data to select hyperparameters as discussed. We average the errors of each model over the iterations. We also repeat the whole analysis with several runs of imputation, to ensure the reported results are not highly dependent on the specific run of the imputation. One useful feature of random forests is that they are able to internally calculate a generalisation error without explicitly putting aside a portion of the data set for testing, by evaluating each tree against the samples which it is not trained on (called out-of-bag samples). However, since we are comparing against the Cox for which we need to put aside a test set anyway, we choose not to utilise this here.

We are also able to perform an additional evaluation of the data imputation. Here we compare performance of all models on the unimputed dataset as well as the imputed dataset, and see what the effect of the imputation is on the errors of the models.

Once we have tested our models, we can refit them to the full dataset with the optimal hyperparameters and use this model to explore the questions in the third line of enquiry laid out in Section 1.3. We are able to alter feature values 'by hand' to investigate the effect this has. We perform several experiments, described in detail in Section 4.4.1.

4.4 Results and discussion

The central result of the first experiment is that tree-based models can reliably outperform the Cox model on this dataset: table 4.5 shows the results. Thus this answers the question of whether we can construct a model which outperforms the Cox model

in the affirmative. It also goes some way to demonstrating that the assumptions of the Cox are violated. This is the likely reason for the outperformance of the tree-based models, along with their greater flexibility. Since a tree recursively partitions the feature space, it can define a 'separate model' for different sections of the space. Thus, a tree might initially split the patients between young and old, and then the effects of any other feature on the two groups need not be at all similar. The Cox, on the other hand, has just one parameter for each feature which defines the strength of the effect on all patients, as well as through time. That is, the Cox model essentially defines a hyperplane through the feature space. We see that the various tree-based models have similar performance, but AdaBoost performs marginally better than the other two on average.

Model	Unimputed	Imputed 1	Imputed 2	Imputed 3	Imputed avg
Cox	353.1	289.3	209.2	307.1	289.7
Random forest	325.0	269.5	**195.6**	274.8	2662.
Extra Trees	330.2	274.3	**195.6**	281.0	270.2
AdaBoost	**319.6**	**263.2**	198.6	**261.4**	**260.7**
KNN	360.3	303.7	216.6	301.5	295.5

Figure 4.5: Table demonstrating the average mean square error for each model on each dataset. The tree-based models reliably outperform the Cox model.

We also find that imputing the missing data improves the performance of the random forest, which may be contrary to expectations. Imputing data adds noise (and likely bias) into the dataset; since the random forest is a more flexible model, we might expect it to learn spurious patterns in this noisy dataset. However, random forests are significantly more data-hungry than the Cox (since they have many more parameters to learn), and it seems that a larger dataset outweighs the cost of this noise. Given that all models perform better with the imputed data it was clearly worthwhile.

We also check the residuals of the fitted random forest (see 6.6 in the Appendix for the plot). Ideally we hope that the residual is independent of the prediction value. Unfortunately we can see a relationship, and indeed the Pearson correlation coefficient between the prediction and the residual is -0.46, $p < 0.01$. This suggests that, although the random forest is a fairly good model, it is not able to perfectly model the underlying processes.

Random forest feature importance Breimen states that a random forest is "impenetrable as far as simple interpretations of its mechanism go" [12]. To some extent this is the case: it does not have a small set of parameters one can easily inspect to understand its functioning. That being said, we now discuss what we can learn from the model. This analysis will provide a possible answer to the question of which factors predict survival as laid out in Section 1.3.

One approach to understanding how features are contributing to survival in the random forest is feature importance. This is usually measured with mean decrease in impurity (MDI): how much, on average, does the splitting based on this feature reduce the variance at the child nodes. However, there is some evidence that this form of variable importance is not necessarily the best measure for how 'impactful' a feature is on the model, particularly when the features are of varying types (e.g. continuous and categorical), have different scales, or are correlated [57, 58]. In his original paper Breiman suggests defining the importance of a feature as the extent to which the performance of the model is affected by permuting the feature[1] [12] . We thus report the permutation importance for the random forest in Figure 4.6. Notably, this has no valence: it does not tell us whether features positively or negatively affect survival.

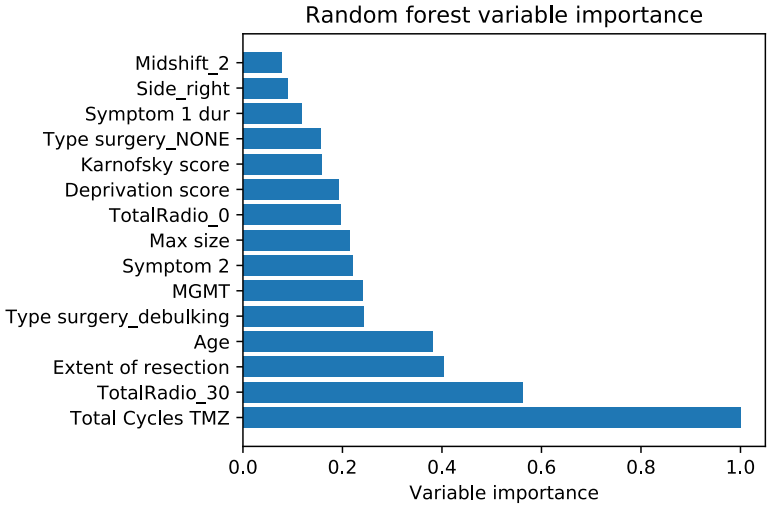


Figure 4.6: Bar chart showing the feature importances of the random forest. Note that these do not have a valency like the Cox model feature importances.

Overall, this ranking of feature importance seems more reasonable than the Cox ranking. Once again we see that the amount of temozolomide is by the most predictive

[1] An indepth discussion can be found at https://explained.ai/rf-importance/

factor, followed by the total dose of radiotherapy. The next two features in the random forest variables, EOR and age, were both fairly insignificant under the Cox model. As we noted previously, since the Cox models the effect of each variable as time-invariant and independent from other features, it is poor at taking into account features which do not fulfil these requirements. There is general consensus that EOR is only significantly predictive after a threshold value [30, 33] (although the evidence may not be strong [59]), so in its raw form the Cox would likely find it not to be significant. The random forest, however, can model this easily, and indeed we see EOR near the top of the importances. Figure 4.7 compares the feature importances in the two models.

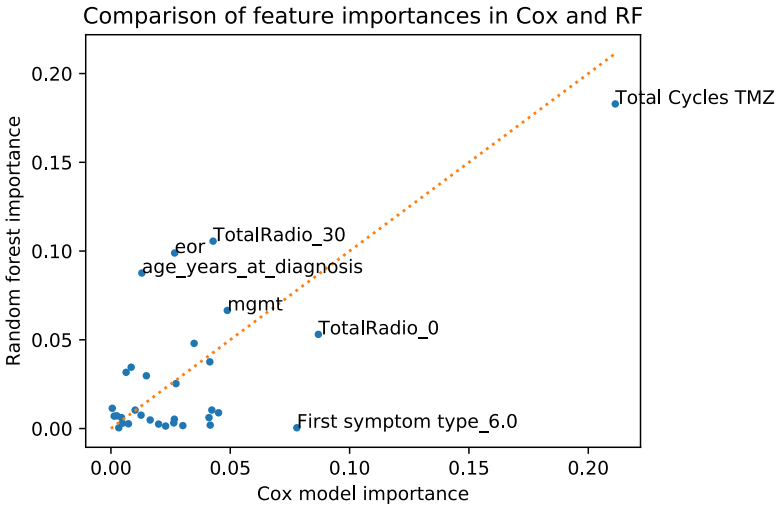


Figure 4.7: Scatter plot comparing feature importances of the two models. Note that we plot the absolute value of the feature importance for the Cox model.

Another striking difference between the two is that the random forest does not rank many symptom or sign features as high importance. This is likely because, although each symptom can probably be used to strongly predict those patients who present with it (possibly due to it being a genuine predictor of survival but also just due to the random variance in that group of patients' survival times), it is does not affect the predictions for the vast majority of patients. Thus the overall performance of the random forest model is not significantly affected by leaving it out.

4.4.1 Treatment decisions

One of the benefits of these models is that they allow us to tweak the input data and rerun it through the model to see what effect the changes have. We will use this now

to address the third line of enquiry as described in Section 1.3: whether the treatment decisions made by the clinic are optimal.

Radiotherapy and chemotherapy for old patients A significant portion of patients do not receive the full treatment recommended by the Stupp protocol [4] (e.g. 254 of 451 did not receive chemotherapy). These tend to be older, worse-performing patients, with the rationale being that they would be less able to resist the negative effects of the treatment. However the collaborating neurosurgeon expressed interest in interrogating whether this is optimal in terms of survival, particularly given that both models consider treatment with temozolomide as the strongest predictor of survival. Previous randomised trials appear to show that treatment with temozolomide seems to be optimal for older patients, and particularly for those with MGMT methylation [60, 61].

We look at those patients who are older than 70, and did not receive chemotherapy or radiotherapy ($n = 68$). The mean survival for these patients was 105.0 days; the mean survival predicted by the random forest model is 109.4 days. The root mean squared error of the model on these patients is 59.5. We set total radiotherapy dosage to 15 and total cycles TMZ to 6, and run these synthesised 'patients' through the model. The result is that mean survival rises to 385.2 days – a significant improvement.

A major caveat for this result (and the following results) is that the number targets who have similar characteristics to the 'target' (in this case, older patients who received full treatment) is often low. In this case there are only 10 patients who received radiotherapy with total cycles TMZ 15 or 30. This means that the model is training its predictions on very few cases, so may not be accurate. Additionally, it is likely that there is some selection bias: the patients who were given full chemotherapy and radiotherapy were judged to be better positioned for treatment. Thus using the same model to make predictions for those patients who were judged *not* to be suitable for the treatment will induce another layer of inaccuracy. That being said, we would hope that most of the differences are reflected in other features (e.g. age, KPS, MGMT), which the model can take into account.

Is biopsy sufficient? Discussions with the consultant neurosurgeon also brought up the question of comparing biopsy to debulking. Although generally if a patient is deemed reasonably fit the surgeons will opt for debulking surgery, he wondered whether there might be a cohort of patients for whom a biopsy, followed by radiotherapy and chemotherapy, would have sufficed.

We look at young patients (younger than 60) who have 100 KPS, MGMT methylation above 0.3, and received debulking surgery ($n = 7$). The mean survival for these patients was 787.1 days, and the mean RF prediction is 727.3. The root mean square error for these patients is 181.2 days. If we set the surgery type to biopsy, the mean survival drops slightly to 673.7. This is well within the error, which suggests that perhaps indeed a biopsy not have been significantly worse than debulking in terms of survival.

A similar caveat applies here to the previous example: there are only 10 patients in the 'training' for this case. There is also an additional consideration here; given that we have demonstrated the many dependencies among features already, it may not be realistic to change just one without modelling how such a change would affect other features. In particular, it is likely that a different surgery would affect the patients response to radiotherapy and chemotherapy, so just changing one feature without the other may not be realistic.

Does debulking surgery with low EOR confer a survival advantage? Previous research has suggested there may be 'cutoff' point in EOR around $80-90\%$, below which debulking surgery does not confer a survival advantage [30, 33]. We investigate this here by looking at patients who received debulking surgery with EOR $< 0.78\%$ ($n = 76$). The mean survival of these patients is 387.8, and the mean of the RF model predictions is 499.0 (with root mean square error 149.4). Changing EOR to 0 leads to a mean predicted survival of 378.8 - a fairly small drop and within the error. This suggests that perhaps indeed if EOR is below 0.78 the patients would fare similarly with no debulking surgery. Again similar considerations apply here, where we cannot be sure than a lower EOR might not lead to, for example, a different response to chemotherapy and radiotherapy.

Should we treat MGMT-unmethylated patients with temozolomide? Previous research has demonstrated that the survival benefit of temozolomide may be dependent on MGMT methylation [62]. We investigate by looking at unmethylated patients (MGMT < 5) who received temozolomide treatment ($n = 6$). These patients had a mean survival time of 263.7 days, and the mean survival estimate for the RF model is 244.8 (RMSE is 187.0). Setting their total cycles TMZ to 0 pushes survival down to 183.4 days. This is a fairly significant drop but still within the error: we consider the results of this experiment inconclusive.

Chapter 5

Future work and conclusion

This research lays the ground for much further work; it is most readily seen as the first, exploratory analysis of this dataset, the size and richness of which mean it is very suitable and deserving of further research. There are a number of directions such work could take.

We have here compared only one group of machine learning models (tree-based) to our baseline, but exploring other types of model would be a natural next step. One interesting option would be to treat cancer as categorical (i.e. 6-months, 12-months survival), and experiment with classification algorithms. We could also investigate incorporating other features into our model, including possibly raw MRI images as discussed in Section 2.3. Another avenue for exploration would be to investigate models with greater explainability (perhaps neural networks of kernel random forests [63]), which is a significant downside to random forests.

Another downside of the random forest model is that, unlike the Cox model, it does not produce confidence intervals for its estimates. A possible extension to this work could be to create such estimates. This could be done in a number of ways: adding a noise element to the inputs to test the sensitivity of the estimate to the inputs; retraining the models several times to test the variance in the ensemble of models' estimates; or using multiple imputation [64] for the data imputation stage.

It would also be interesting to further interrogate the random forest model's capacity for generalisation, either to data from another institution, or to the first, larger dataset. For the latter, the heterogeneity would likely present a significant challenge: one option would be to perform an initial clustering of the data and then train a model for each cluster; another would be to train separate models for the various different types of cancer.

We expect further discussion with the collaborating neurosurgeon to raise more questions about treatment decisions which could be interrogated through these models as in Section 4.4.1.

5.1 Conclusion

This thesis has made significant progress towards the objectives laid out in Section 1.3:

1. In Chapter 4 we were able to demonstrate which factors predict survival using two different models, and evaluate this in a medical context. By comparing the random forest model which was able to model interactions, and the Cox model which was not, we were able to get some sense of how some of the features may be interacting to affect survival. We provided evidence that the Cox model's assumptions were not valid, and we demonstrated that the tree-based models can reliably outperform the Cox model.

2. In Section 3.3.1 we investigated the dependencies among the features, and addressed how these fit into previous research as well as how these impact our understanding of treatment. In Section 3.3.4 we gave further evidence of the dependencies between the features.

3. In Section 4.4.1 we investigated whether the treatment decisions made by the clinic were optimal, and compared our results to the existing medical literature.

We were also able to identify a number of weak points in the existing literature – most notably a closer analysis of the random forest training and performance, as well as the incorporation of MGMT and IDH and symptom features into the analysis – and made progress towards addressing these.

One of the significant outputs of this paper will be its investigation into data imputation, and the creation of a larger, more filled-out dataset. This can be used in all future research, and will be especially valuable for any machine learning done on the dataset. This is particularly worthwhile as, given that brain cancer – and glioblastoma in particular – are rare diseases, it is not possible to collect large datasets for analysis. This is an intrinsic difficulty in the investigation of such diseases, so improving the existing dataset is a significant contribution.

Difficulties This project had a number of risks which were evaluated before beginning. Most significantly, this was a piece of research performed on a dataset which had not been used previously and which we had not seen, and which turned out to be of relatively poorer quality than expected. This was compounded by the fact that the collaborating neurosurgeon was at times unavailable or unable to go back and check/correct the data. By and large this was mitigated by the exploratory and open-ended approach we took, with the direction of the research shaped by the ongoing development of our understanding of the dataset.

Aspects for improvement One of the approaches to this research which could have been improved was that we should have been clearer earlier on what was or was not going to be available over the course of the project. Discussions were being had up until the final weeks with the collaborating neurosurgeon about possibly providing more complete records, and this back-and-forth delayed the research. This was likely worsened by the short duration of the project. It would have been better to establish more clearly at the start exactly what was available.

The initial goals of the research were in fairly drastic contrast to most 'traditional' medical research: we did not conduct randomized trials, we did not focus on p-values, and not all of our results are simply and unequivocally interpretable. What this work offers is an alternative approach and perspective, which ideally should challenge established views and stimulate further research and discussion. Given the moderate size of the dataset, further work is needed to evaluate and interrogate our conclusions. The final pieces of analysis in particular are deserving of further exploration, to establish some confidence level for these results and explore possible shortfalls of the approaches used.

That being said, one of the major benefits of this machine learning approach is that we are able to include all patients in the dataset, and even synthesise 'new patients' by hand-altering inputs. Traditional medical research involving randomised trials is plagued with problems of eligibility, which can induce significant bias in any results, as well as leading to small sample sizes and underpowered results. It is also hard to run trials when testing for likely poor options like less treatment, whereas we are able to experiment with this freely. Taking this data-driven approach avoids these problems and allows for a different perspective. This dataset thus provides an excellent opportunity for further research.

Bibliography

[1] Dong Nie, Junfeng Lu, Han Zhang, Ehsan Adeli, Jun Wang, Zhengda Yu, LuYan Liu, Qian Wang, Jinsong Wu, and Dinggang Shen. Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific reports*, 9(1):1103, 2019.

[2] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):10353, 2017.

[3] BWKP Stewart, Christopher P Wild, et al. World cancer report 2014. 2014.

[4] Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

[5] Derek R Johnson and Brian Patrick ONeill. Glioblastoma survival in the united states before and during the temozolomide era. *Journal of neuro-oncology*, 107(2):359–364, 2012.

[6] David Roxbee Cox. *Analysis of survival data*. Chapman and Hall/CRC, 2018.

[7] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[8] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[9] Leo Breiman, Jerome H Friedman, RA Olshen, and CJ Stone. Classification and regression trees (belmont, ca: Wadsworth international group). *Biometrics*, 40(3):17–23, 1984.

[10] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[11] Walter J Curran Jr, Charles B Scott, John Horton, James S Nelson, Alan S Weinstein, A Jennifer Fischbach, Chu H Chang, Marvin Rotman, Sucha O Asbell, Robert E Krisch, et al. Recursive partitioning analysis of prognostic factors in three radiation therapy oncology group malignant glioma trials. *JNCI: Journal of the National Cancer Institute*, 85(9):704–710, 1993.

[12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] Hemant Ishwaran and Min Lu. Random survival forests. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2007.

[14] Elena A Manilich, Ravi P Kiran, Tomas Radivoyevitch, Ian Lavery, Victor W Fazio, and Feza H Remzi. A novel data-driven prognostic model for staging of colorectal cancer. *Journal of the American College of Surgeons*, 213(5):579–588, 2011.

[15] Rajan Jain, Laila M Poisson, David Gutman, Lisa Scarpace, Scott N Hwang, Chad A Holder, Max Wintermark, Arvind Rao, Rivka R Colen, Justin Kirby, et al. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology*, 272(2):484–493, 2014.

[16] Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):11707, 2017.

[17] M.R. Segal. Machine learning benchmarks and random forest regression. *eScholarship Repository*, 2004.

[18] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[19] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[20] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[21] V Panca and Zuherman Rustam. Application of machine learning on brain cancer multiclass classification. In *AIP Conference Proceedings*, volume 1862, page 030133. AIP Publishing, 2017.

[22] Sunil Gupta, Truyen Tran, Wei Luo, Dinh Phung, Richard Lee Kennedy, Adam Broad, David Campbell, David Kipp, Madhu Singh, Mustafa Khasraw, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ open*, 4(3):e004007, 2014.

[23] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, 36(2):3240–3247, 2009.

[24] Woojae Kim, Ku Sang Kim, Jeong Eon Lee, Dong-Young Noh, Sung-Won Kim, Yong Sik Jung, Man Young Park, and Rae Woong Park. Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of breast cancer*, 15(2):230–238, 2012.

[25] Kanghee Park, Amna Ali, Dokyoon Kim, Yeolwoo An, Minkoo Kim, and Hyunjung Shin. Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, 26(9):2194–2205, 2013.

[26] Siow-Wee Chang, Sameem Abdul-Kareem, Amir Feisal Merican, and Rosnah Binti Zain. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*, 14(1):170, 2013.

[27] Hai Yan, D Williams Parsons, Genglin Jin, Roger McLendon, B Ahmed Rasheed, Weishi Yuan, Ivan Kos, Ines Batinic-Haberle, Siân Jones, Gregory J Riggins, et al. Idh1 and idh2 mutations in gliomas. *New England Journal of Medicine*, 360(8):765–773, 2009.

[28] Haley Gittleman, Daniel Lim, Michael W Kattan, Arnab Chakravarti, Mark R Gilbert, Andrew B Lassman, Simon S Lo, Mitchell Machtay, Andrew E Sloan,

Erik P Sulman, et al. An independently validated nomogram for individualized estimation of survival among patients with newly diagnosed glioblastoma: Nrg oncology rtog 0525 and 0825. *Neuro-oncology*, 19(5):669–677, 2017.

[29] Jerome H Friedman. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, (4):404–408, 1977.

[30] Michel Lacroix, Dima Abi-Said, Daryl R Fourney, Ziya L Gokaslan, Weiming Shi, Franco DeMonte, Frederick F Lang, Ian E McCutcheon, Samuel J Hassenbusch, Eric Holland, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *Journal of neurosurgery*, 95(2):190–198, 2001.

[31] Ken Chang, Biqi Zhang, Xiaotao Guo, Min Zong, Rifaquat Rahman, David Sanchez, Nicolette Winder, David A Reardon, Binsheng Zhao, Patrick Y Wen, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. *Neuro-oncology*, 18(12):1680–1687, 2016.

[32] Dalu Yang, Ganesh Rao, Juan Martinez, Ashok Veeraraghavan, and Arvind Rao. Evaluation of tumor-derived mri-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Medical physics*, 42(11):6725–6735, 2015.

[33] Nader Sanai, Mei-Yin Polley, Michael W McDermott, Andrew T Parsa, and Mitchel S Berger. An extent of resection threshold for newly diagnosed glioblastomas. *Journal of neurosurgery*, 115(1):3–8, 2011.

[34] Zhenye Li, Xiangheng Zhang, Xiaobing Jiang, Chengcheng Guo, Ke Sai, Qunying Yang, Zhenqiang He, Yang Wang, Zhongping Chen, Wei Li, et al. Outcome of surgical resection for brain metastases and radical treatment of the primary tumor in chinese non–small-cell lung cancer patients. *OncoTargets and therapy*, 8:855, 2015.

[35] Laurie Gaspar, Charles Scott, Marvin Rotman, Sucha Asbell, Theodore Phillips, Todd Wasserman, W Gillies McKenna, and Roger Byhardt. Recursive partitioning analysis (rpa) of prognostic factors in three radiation therapy oncology group (rtog) brain metastases trials. *International journal of radiation oncology, biology, physics*, 37(4):745–751, 1997.

[36] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[37] Philipp Kickingereder, Sina Burth, Antje Wick, Michael Götz, Oliver Eidel, Heinz-Peter Schlemmer, Klaus H Maier-Hein, Wolfgang Wick, Martin Bendszus, Alexander Radbruch, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, 280(3):880–889, 2016.

[38] Guido Van Rossum and Fred L Drake. Python language reference manual. 2003.

[39] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90, 2016.

[40] Eric Jones, Travis Oliphant, Pearu Peterson, et al. Scipy: Open source scientific tools for python. 2001.

[41] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90, 2007.

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[44] Cameron Davidson-Pilon, Jonas Kalderstam, Paul Zivich, Ben Kuhn, Andrew Fiore-Gartland, Luis Moneda, Gabriel, Daniel WIlson, Alex Parij, Kyle Stark, and et al. Camdavidsonpilon/lifelines: v0.22.3 (late). Aug 2019.

[45] Joseph R Dettori. Loss to follow-up. *Evidence-based spine-care journal*, 2(01):7–10, 2011.

[46] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. In *Advances in Neural Information Processing Systems*, pages 8011–8023, 2018.

[47] David A Karnofsky, Walter H Abelmann, Lloyd F Craver, and Joseph H Burchenal. The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma. *Cancer*, 1(4):634–656, 1948.

[48] Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

[49] Ági Oszvald, Erdem Güresir, Matthias Setzer, Hartmut Vatter, Christian Senft, Volker Seifert, and Kea Franz. Glioblastoma therapy in the elderly and the importance of the extent of resection regardless of age. *Journal of neurosurgery*, 116(2):357–364, 2012.

[50] JR Simpson, J Horton, C Scott, WJ Curran, P Rubin, J Fischbach, S Isaacson, M Rotman, SO Asbell, JS Nelson, et al. Influence of location and extent of surgical resection on survival of patients with glioblastoma multiforme: results of three consecutive radiation therapy oncology group (rtog) clinical trials. *International Journal of Radiation Oncology* Biology* Physics*, 26(2):239–244, 1993.

[51] Paul D Allison. Imputation of categorical variables with proc mi. *SUGI 30 proceedings*, 113(30):1–14, 2005.

[52] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[53] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.

[54] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, 2003.

[55] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Nashville, USA, 1997.

[56] Jigisha P Thakkar, Therese A Dolecek, Craig Horbinski, Quinn T Ostrom, Donita D Lightner, Jill S Barnholtz-Sloan, and John L Villano. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiology and Prevention Biomarkers*, 23(10):1985–1996, 2014.

[57] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[58] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

[59] Timothy J Brown, Matthew C Brennan, Michael Li, Ephraim W Church, Nicholas J Brandmeir, Kevin L Rakszawski, Akshal S Patel, Elias B Rizk, Dima Suki, Raymond Sawaya, et al. Association of the extent of resection with survival in glioblastoma: a systematic review and meta-analysis. *JAMA oncology*, 2(11):1460–1469, 2016.

[60] Annika Malmström, Bjørn Henning Grønberg, Christine Marosi, Roger Stupp, Didier Frappaz, Henrik Schultz, Ufuk Abacioglu, Björn Tavelin, Benoit Lhermitte, Monika E Hegi, et al. Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma: the nordic randomised, phase 3 trial. *The lancet oncology*, 13(9):916–926, 2012.

[61] Wolfgang Wick, Michael Platten, Christoph Meisner, Jörg Felsberg, Ghazaleh Tabatabai, Matthias Simon, Guido Nikkhah, Kirsten Papsdorf, Joachim P Steinbach, Michael Sabel, et al. Temozolomide chemotherapy alone versus radiotherapy alone for malignant astrocytoma in the elderly: the noa-08 randomised, phase 3 trial. *The lancet oncology*, 13(7):707–715, 2012.

[62] Monika E Hegi, Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas De Tribolet, Michael Weller, Johan M Kros, Johannes A Hainfellner, Warren Mason, Luigi Mariani, et al. Mgmt gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, 352(10):997–1003, 2005.

[63] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

[64] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393, 2009.

[65] Dan Rohle, Janeta Popovici-Muller, Nicolaos Palaskas, Sevin Turcan, Christian Grommes, Carl Campos, Jennifer Tsoi, Owen Clark, Barbara Oldrini, Evangelia Komisopoulou, et al. An inhibitor of mutant idh1 delays growth and promotes differentiation of glioma cells. *Science*, 340(6132):626–630, 2013.

[66] Songtao Qi, Lei Yu, Hezhen Li, Yanghui Ou, Xiaoyu Qiu, Yanqing Ding, Huixia Han, and Xuelin Zhang. Isocitrate dehydrogenase mutation is associated with tumor location and magnetic resonance imaging characteristics in astrocytic neoplasms. *Oncology letters*, 7(6):1895–1902, 2014.

Appendix

Variable importance in imputation

An interesting further question is, when imputing some feature, which other features are important? Earlier we discussed correlations between features, but this does not necessarily draw out all dependencies between features (consider $X \sim Uniform(-1,1)$ and $Y = X^2$. Clearly the two features are related but their correlation will be ≈ 0.) Additionally this does not let us investigate categorical features.

One way to investigate this is to repeatedly impute a feature, each time choosing to drop one other feature. In this way we can observe which feature, when left out, leads to the greatest drop is accuracy. The results of this analysis are seen in Figure 6.1. Note that this is normalised by row but not overall, so that the importances in different rows are not directly comparable. Normalising the whole matrix just shows that surgery type used for extent of resection is far more predictive than any other variable pair (so we end up with an almost-white graph with a single coloured square). This graph thus needs some contextualising – the missing piece of information for each row is whether *any* feature being dropped results in a significant loss in imputation performance. For example, we might be surprised to see that MGMT seems important for imputing deprivation score, or that many features seem important for imputing IDH. However really this just shows that, for these two features, there is no single feature which, when dropped, makes imputation significantly harder.

What we *can* see is quite interesting however. By and large it confirms our earlier analysis of feature correlations: Karnofsky score is linked to age and size, size is linked to midline shift etc. One interesting observation is that some strongly correlated features – for example total cycles TMZ and EOR – do not depend on each other for imputation. The reason for this is that there are other features which are strongly correlated with both and can be used for the imputation – in this example, total cycles TMZ can be predicted from total radiotherapy just as well as from EOR. Also important is that for some features, the number of imputed values is quite small, so effects may not be hugely significant. [EDIT: CHECK THIS]

A number of somewhat surprising dependencies are drawn out through this analysis. IDH seems to be somewhat important for predicting Karnofsky score. It does seem possible that IDH-mutated patients are less likely to have lower Karnofsky score (2 out of 7 have score below 90, compared to 34 out of 100 for wildtype), but the numbers are too small to be confident. As mentioned previously, patients with a mutated

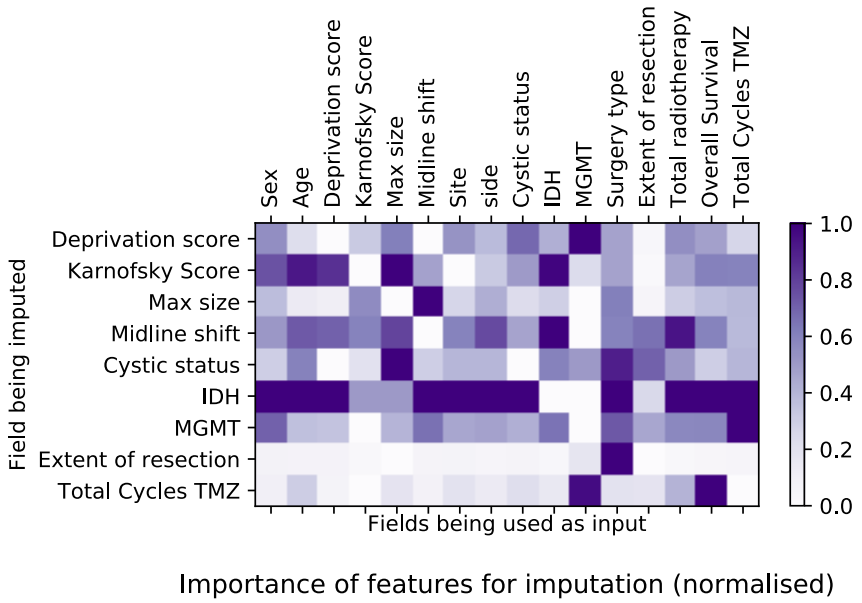


Figure 6.1: Feature importances for imputation. Note that it is normalised by row, so the importances are not comparable from one row to another.

IDH1 gene have been shown to have a significantly better survival outlook [27]. The current clinical view seems to be that tumours in patients with a mutated IDH gene tend to grow more slowly [65]. This could lead to them tending to present with higher Karnofsky score, since slower growth may lead to fewer symptoms. It also seems that IDH-mutated patients are less likely to have tumours in high-risk areas [66]. However, the number of patients with mutated IDH is only 7, so we should take care before drawing strong conclusions. Given that the test for IDH is relatively cheap and easy, further investigation may be possible in the future with larger cohorts of patients who have been tested for the mutation.

Another interesting and significant relationship is that total cycles TMZ (the total amount of chemotherapy given) seems important for predicting MGMT methylation. Figure 6.2 illustrates this further. There are two plausible explanations for this: firstly, there is growing consensus that patients with higher MGMT methylation are likely to see a greater benefit from treatment with temozolomide [62]. Given this, we might well expect clinicians to assign higher levels of temozolomide treatment to patients who have higher methylation. Secondly, given that patients with higher methylation have a better outlook (both with and without temozolomide), it is possible that they make it further through their planned treatment protocols. This would also lead to a

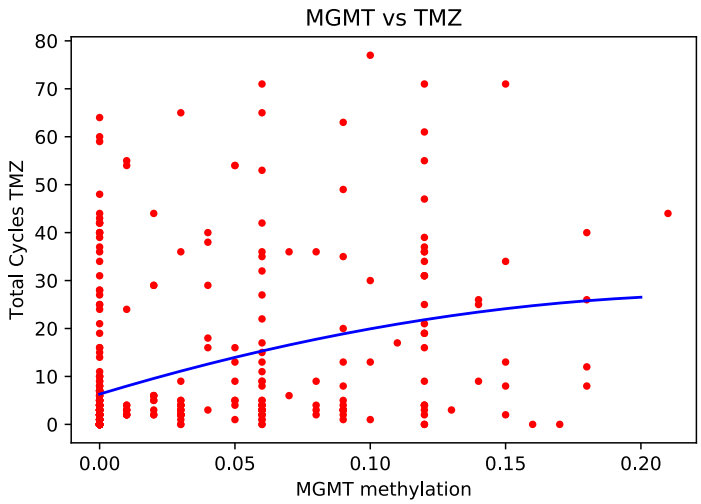


Figure 6.2: Illustration of relationship between MGMT methylation and total dosage of temozolomide (chemotherapy drug). A moderate positive relationship is clear.

positive relationship between the two variables.

Although this analysis affords some further insight into the relationships between the features, we note at this point that since all the models are essentially 'linear' or implicitly make assumptions of independence between features, they cannot provide us with insight into the multivariate interactions between features. However, in the next chapter we will investigate the use of models which can for survival modelling and hopefully provide some motivation for their use.

The collinear features also weaken this analysis somewhat. Clearly if you have two highly collinear features, dropping out one of the features is unlikely to significantly affect any model, since what can be predicted with both can be predicted with just 1. An extension of this analysis could be to drop out *groups* of features which are highly collinear and see the effect.

Glossary of features

Sex – the sex of the patient

Age – the age of the patient

Deprivation score – a measure of the relative wealth of the area which the patient lives in

Karnofsky performance score (KPS) – a measure of the patient's level of mental functioning

Max size - a measure of the maximum size of the tumour (across three planes)

Midline shift – a measure of how much the brain has moved over its center line

Site – the lobe in which the tumour is found

Side – whether the tumour is in the left or right side of the brain (or both)

MGMT methylation – a measure of how methylated the MGMT gene's promoter is

Type of surgery – either biopsy (just to get a piece of the tumour for testing); debulking (to try to cut out as much of the tumour as possible); excision (to try to cut out all of the tumour); or none.

Extent of resection (EOR) – a measure of what percentage of the tumour is cut out during surgery

Total radiotherapy dose – a measure of the total dosage of radiotherapy administered to the patient over the course of the treatment protocol

Total cycles TMZ – a measure of how many cycles of temozolomide, the chemotherapy drug, are administered to the patient over the course of the treatment protocol

Symptoms – the physical or mental symptoms which the patient presented with to their doctor (usually GP)

Signs – the physical or mental signs observed by the clinicians

Survival – the number of days between the patient's initial diagnosis and death

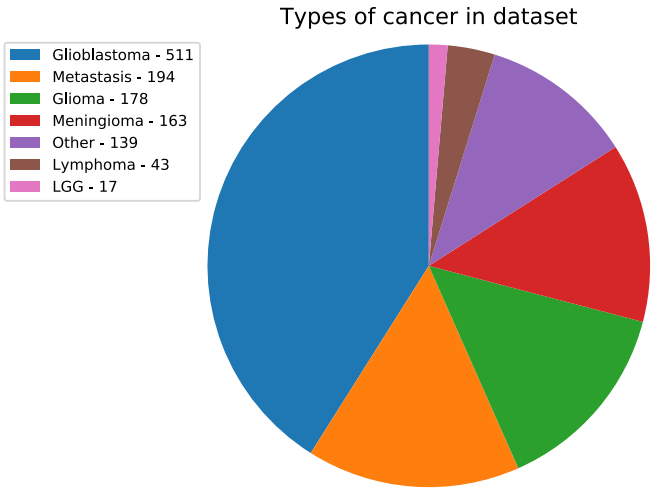


Figure 6.3: Pie chart showing the proportion of patients with each type of cancer. Note the heterogeneity present: of the 1,334 patients, only 511 have glioblastoma. The most common other types are Metastasis, Glioma, and Meningioma.

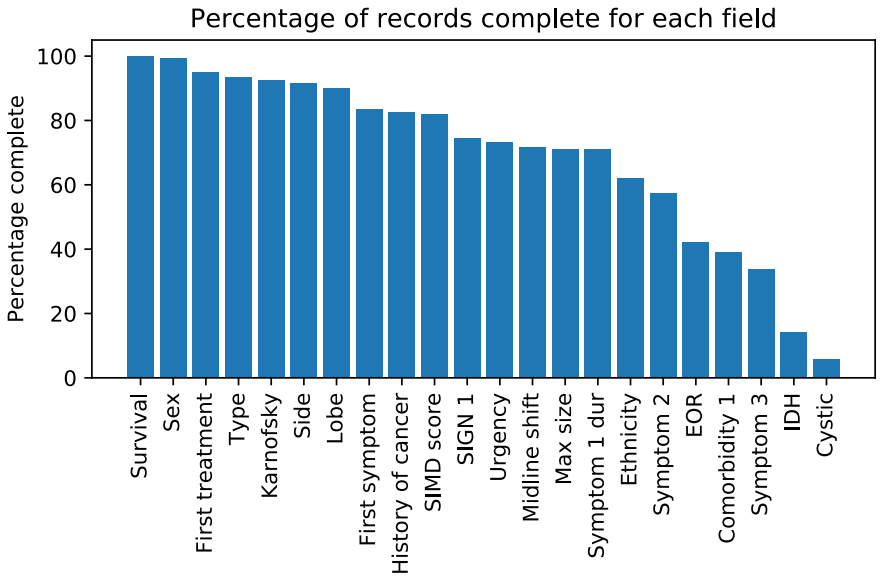


Figure 6.4: Bar chart showing the percentage of records complete for each field. Apart from ID, survival and first treatment, all fields have some missing data. The compounding effect of this means that very few patients have all fields complete.

TABLE I

PERFORMANCE STATUS

Definition	%	*Criteria*
Able to carry on normal activity and to work. No special care is needed.	100	Normal; no complaints; no evidence of disease.
	90	Able to carry on normal activity; minor signs or symptoms of disease.
	80	Normal activity with effort; some signs or symptoms of disease.
Unable to work. Able to live at home, care for most personal needs. A varying amount of assistance is needed.	70	Cares for self. Unable to carry on normal activity or to do active work.
	60	Requires occasional assistance, but is able to care for most of his needs.
	50	Requires considerable assistance and frequent medical care.
Unable to care for self. Requires equivalent of institutional or hospital care. Disease may be progressing rapidly.	40	Disabled; requires special care and assistance.
	30	Severely disabled; hospitalization is indicated although death not imminent.
	20	Very sick; hospitalization necessary; active supportive treatment necessary.
	10	Moribund; fatal processes progressing rapidly.
	0	Dead.

Figure 6.5: Original definition of Karnofsky Performance Index [47]. Although the score is introduced as a percentage score (i.e. a score of 100 meaning the patient is 100% able to 'carry on normal activity'), the authors did specifically define the meanings of the multiples of 10.

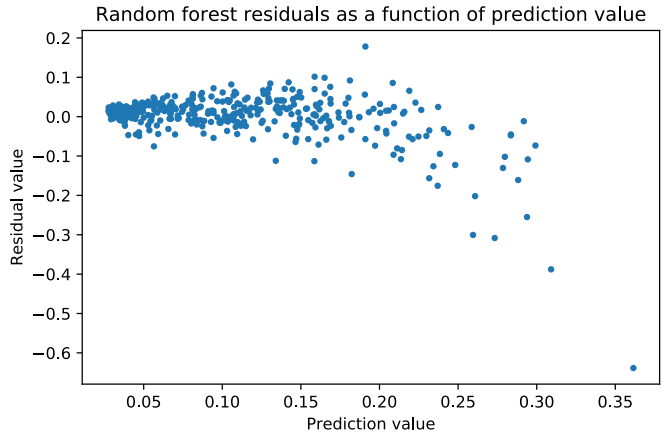


Figure 6.6: Plot of residuals (errors) of the random forest model as a function of the value of the prediction. There does appear to be somewhat of a pattern meaning that the random forest model is unable to properly model the underlying processes.