Bayesian statistics and modeling for the prediction of radiotherapy outcomes : an application to glioblastoma treatment

Oscar Daniel Zambrano Ramirez

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité PHYSIQUE

Préparée au sein de l'Université de Caen Normandie

Bayesian statistics and modeling for the prediction of radiotherapy outcomes. An application to glioblastoma treatment

Présentée et soutenue par
Oscar Daniel ZAMBRANO RAMIREZ

Thèse soutenue publiquement le 18/12/2018 devant le jury composé de		
M. JACQUES BALOSSO	Professeur des universités, Université Grenoble Alpes	Rapporteur du jury
M. PHILIPPE MEYER	Physicien, Centre Paul Strauss	Rapporteur du jury
M. RENAUD DE CREVOISIER	Professeur des universités, Centre Eugène Marquis	Président du jury
Mme ISABELLE GARDIN	Physicien, Centre Henri Becquerel Rouen	Membre du jury
Mme JULIETTE THARIAT	Professeur des universités, Université Caen Normandie	Membre du jury
M. JEAN-MARC FONTBONNE	Ingénieur de recherche au CNRS, Université Caen Normandie	Directeur de thèse

Thèse dirigée par JEAN-MARC FONTBONNE, Laboratoire de physique corpusculaire (Caen)

UNIVERSITÉ CAEN NORMANDIE

ED 591 PSIME

Acknowledgment

I sincerely thank all the people whom directly or indirectly participated for the accomplishment of the work presented in this manuscript.

The thesis work represents the collective effort of numerous people and institutions. First, I would like to thank my main financial sponsor CONACYT (COnsejo NAcional de Ciencia Y Tecnología) for its generosity, and the Laboratoire de Physique Corpusculaire de Caen (LPC) for providing the necessary space and equipment that made this work possible.

Thank you to the directors of LPC, currently Gilles Ban and previously Dominique Durand for looking after the laboratory.

My deepest and sincere thanks to my thesis director Jean-Marc Fontbonne for all his effort including the rapid and precise suggestions and corrections as well as his personal support.

I thank the members of the medical physics and applications group for the discussions and guidance.

My sincere thanks to the jury members for taking the time to read the manuscript: Philippe Meyer, Jacques Balosso, Isabelle Gardin, Juliette Thariat, and Renaud De Crevoisier.

Lastly, I thank my colleagues, friends, other PhD students and post-docs who became family to me. I thank my immediate family for their patience and unconditional support.

Contents

List of Figures

List of Tables

Introduction

1.1 Introduction

1.1.1 Motivations

The constant improvement of oncology treatments has led to a significant increase of medical data in the form of electronic records. The computational power increase and the need to move towards a personalized approach to improve patient care is leading towards the development of predictive models

based on clinical evidence. Medicine is directing towards an era of personalized medicine [8]. Personalized treatment is of particular interest since the biological response of individual patients can vary greatly despite administering a treatment in a similar way. Hence, tailoring treatments could significantly improve outcomes. Predictive models are currently being developed to integrate the considerable amounts of oncology data [5]. The potential of using Machine Learning (ML) techniques on big data is a promising path towards reaching personalized patient care. Such techniques can be used to support clinicians to reach more knowledgeable treatment decisions based on previous clinical evidence of pass patients [51]. Bayesian approaches can be used for learning from clinical evidence in a continuous manner allowing the acceptance of the incorporation of new data leading to a continuous learning approach.

1.1.2 Objective

The main objective is to develop biophysical models to predict clinical effects by finding patterns from previously documented clinical records of patients including medical images. While developing these prediction models we can find parameters that play important roles in the clinical outcome. Bringing to light these hidden important parameters leads towards a personalized approach.

1.1.3 Personalized medicine approach

A personalized approach can surge from using the important parameters identified during the development of biophysical models. Examples of parameters may include, number of radiation sessions, Grays per session, type of machine used, technique used, response to treatment due to tumor size, age related response, etc. The idea is to be able use those identified parameters to personalize the approach of a patient being treated.

Personalized medicine refers to the tailoring of treatments taking into account the response to therapy or undesired effects of the treatment for subgroups of patients [3]. Precision medicine is a term often used in personalized medicine. Precision medicine is a strategy used for the amelioration of patient-specific therapies, diagnosis etc [10]. Patterns are searched from clinical data with the idea that these patterns are linked to the character-

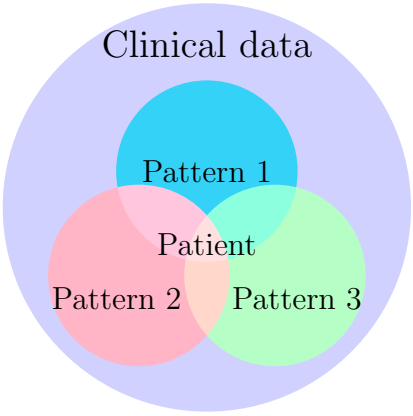


Figure 1.1: Personalized medicine diagram in which a hypothetical patient has three specific patterns represented by the overlapped section.

istics of an individual patient. As an illustration, the diagram in figure 1.1 shows the hypothetical identification of patterns from a clinical database with a patient having the characteristics corresponding to each of the patterns. Patterns found can overlap, for instance if a pattern involved requires age as a parameter and another pattern requires number of radiation sessions, and a last pattern involves blood type then the overlapping correspond to a patient which has the required age, radiation sessions, and blood type. Hence, according to the patient characteristics, a personalized approach could be developed for that particular patient.

1.2 Glioblastoma database

1.2.1 Glioblastoma

Glioblastoma is a very aggressive type of brain cancer typically resistant to treatments, including to chemotherapy and radiotherapy. Glioblastoma affect patients at different ages but predominantly affects older patients, and the patients often show EGFR overexpression, PTEN (MMAC1) mutations [28]. Glioblastomas are one of the most vascularized and highly invasive cancers and lamentably prognostics have not shown much improvement in decades[1]. Much work needs to be done to improve the prognostics from

understanding the genetic mutations associated to it as well as very early detection before it migrates anywhere else. The classification of brain tumors ranks from grade I to grade IV, the latter being the most aggressive [27]. A Glioblastoma multiform is a grade IV hence it is a very aggressive cancer. In "the 2016 World Health Organization classification of tumors of the central nervous system" report mentions that the classification now uses molecular parameters in addition to histology [41].

1.2.2 Clinical data

Clinical data increase
Doctors or medical practitioners document the medical history of the patient including some clinical effects caused by the therapy. The well being of the patient, hematologic grade, neurologic grade, blood pressure among many other fields can potentially be found in clinical records. These clinical records become utterly handy to estimate the efficacy of a treatment. Furthermore, those clinical records containing quantifiable physical units are even more relevant from a mathematical point of view. In addition, to the clinical records documented, the medical images are ever more present in medicine providing a great source of quantifiable descriptive information about the patient. Most Electronic Health Care Records (EHR) now include quantitative data [45]. The amount of data that is being produced is drastically increasing and the clinical records from clinics and hospitals often comes as EHR that are technologically easy to acquire.

Data privacy considerations
The gathering of oncology data is technologically feasible and relatively uncomplicated in many cases. However, the administrative and necessary steps for ensuring data privacy can considerably lengthen the process of working with the data for research purposes. In this work, meticulous steps were taken to ensure data privacy. For instance, the name of the patient is not written in the database used for research, and the facial features of the medical images concerning head scanning are unidentifiable to the researchers. Another privacy measure considered was the protection of the dates used. That is, measures were taken to avoid being able to identify a patient by using purely the database.

Clinical data acquisition
The clinical data used in this work was obtained from the oncology center "Centre François Baclesse" located in Normandy, France. The database

consists of clinical records recollected throughout the medical history of approximately 90 patients suffering from Glioblastoma. The brain cancer was chosen for several reasons such as the amount of medical imaging performed during the Glioblastoma treatments. For this study a fix amount of data was acquired but a long term goal would be to continuously keep adding data to the database as a continuous study approach instead of a retrospective study. Hence, at long term additional clinical records could be added.

Glioblastoma database
The data recollected is divided in two types of clinical records, data entries and medical images. The data entries refers to fields that are registered and entered in the computer such as the symptoms and gender of the patient, overall data entries refer to most other fields that are not medical images. Data entries can be quantitative or qualitative and come in different formats. The gender is male or female, age is a float number, complications can be written in grade scales. Some of the data entries can also be true or false, such as if surgery was performed or not. The type of surgical removal was recorded: either no surgical extraction, partial or complete extraction of the tumor. The database also contains several different Magnetic Resonance Images (MRIs) and Computer Tomography (CT) images, and these type of clinical records are referred simply as medical images. The database contains several different MRI sequences including the conventional sequence T1-Gd and T2-Flair, and the diffusion sequence DW-MRI. The MRI sequences refer to the specific type of MRI imaged performed and their specific meanings are explained later on in the third chapter. Lastly, a field in this case represents any clinical record; for instance age is a field and a single set of CT images is another field.

Example of field: Age is one field, and a set of CT images is another

Even though the medical history of patients in our database can vary depending on the individual case we can illustrate a typical medical history timeline of the patients. The medical history timeline representation is illustrated in figure 1.2. Typically the patient goes through a medical consultation where data entries are recorded, for instance the symptoms and age are written down. Medical exams are performed which can include medical imaging exams. A diagnostic is made with the help of the medical exams and the patient is treated in some cases starting with surgery then the radiotherapy planning is performed following by the radiation therapy and the chemotherapy. The database contains the CT scans, the 3D dosimetry, the contouring of the targeted areas the organs at risks etc. Complications and several side effects are also recorded. Due to the aggressive nature of Glioblastoma adjuvant treatment(s) may be performed. The database also contains the clinical

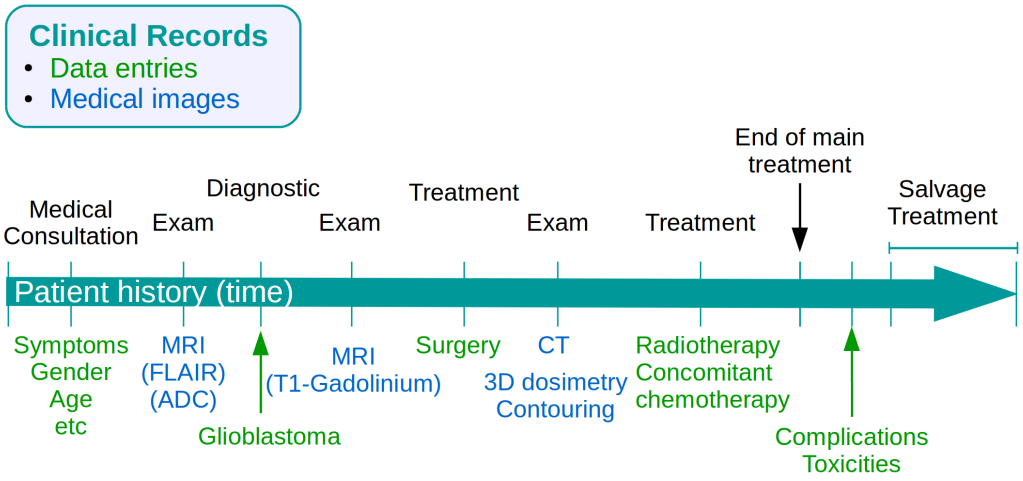


Figure 1.2: Medical history timeline representation of patients suffering from Glioblastoma brain cancer. The clinical records generated throughout the medical history are specified.

records of those treatments.

Overall the database contains a rich amount of clinical records which allow for the analysis and development of clinically based models. However, even though there are many fields the amount of clinical records vary from patient to patient since the medical history of the patient is specific to each patient. Hence, this affects the development of models since not all the data of patients can be used for each model. That is, it reduces the number of subjects used in the models. In general, interesting models can be developed using this rich field database.

1.3 Main parts of the document

1.3.1 Bayesian formalism

The first part is about the development of a Bayesian framework for the modeling of clinical data. The chapter starts with a comprehensive and

intuitive explanation of the Bayesian approach that allows us to develop clinically based prediction models. The three main inference objectives of Bayes' theorem are presented which are: estimating the parameter usually by finding its posterior distribution, making a prediction for a new data set using the posterior distribution of the parameter, comparison of models. One of the main advantages of the Bayesian approach is that the uncertainty of the parameters is predicted.

Several practical examples of personalized prediction models were developed using the Bayesian framework which includes neurologic grade prediction models. The numerical approach to solving Bayes' theorem is emphasized because it allows for the modeling of a wide range of situations. Lastly, the solving process involves machine learning techniques.

1.3.2 Tumor recurrence predictions

The second part applies a reduced case of the Bayesian framework in which Generalized Linear Models (GLM) were built to explore the correlations of pre-treatment medical images in the form of magnetic resonance (MRI) and computer tomography (CT) with the recurrence of the tumor. The MRI sequences used include DW-MRI, T2-Flair, and T1-Gd. Then in a similar manner more complex models using decision trees, from machine learning techniques, were performed to unveil possible hidden correlations to the recurrence which GLM models are incapable of finding.

The intensity values of the images corresponding to locations in the outer surface of the tumor and its mirror image on the opposite side of the brain were analyzed and compared. The mirror images were also used for normalization purposes. Layers (of about 2 mm) were created normally increasing from the surface of the tumor. A comparison of intensity values was done independently for each of the several layers. This was possible because the position of the recurrence and non-recurrence areas, within the layers, are known from the medical image data. The relevance, and limitation of the study are also discussed.

1.3.3 Spherical mapping and cartography representation tools

The third part involves the development of spherical mapping and cartography representation tools derived by the necessity to visually analyze the tumor recurrence link to the intensity values of the recurrence. The development of these tools are a continuation of the recurrence study and they were developed because the mathematical models in part two failed to strongly correlate the recurrence to a change in intensity values.

The first step to develop these tools is to construct a mesh covering the expanded tumor structure in which each of the vertices has known Cartesian coordinates. Then a couple of angles, latitude and longitude, are associated to each vertex by following a clever algorithm developed by Brechbühler et al., [6] which involves the concept of heat diffusion. With the latitude and longitude angles we were able to create a spherical map of the expanded tumor. Then 2D maps, of the surface of the spherical representation, in the form of Mercator and Mollweide projections were constructed. The 2D maps allow for a direct analysis of intensity values of the recurrence versus non-recurrence locations.

1.4 Conclusions

The framework developed in this work allows for making predictions based on evidence using the data rich Glioblastoma database. The potential uses of the framework were illustrated by several examples including the neurologic grade prediction models and Glioblastoma tumor recurrence predictions. Overall, prediction based on clinical evidence allows to move forward towards a personalized approach in health care. In this approach, unique traits of patients can be used for developing individual predictions.

Bayesian framework for modeling clinical data

2.1 Introduction

The purpose of this chapter is to illustrate the framework for the development of clinically based biophysical prediction models using Bayesian statistics.

Medical data is being increasingly generated in hospitals and clinics. Analyzing the stored data opens the door to learning from evidence. The data can be further exploited by searching for patterns and correlations. In the traditional style health practitioners rely on experiences with their own patients and in discussions with colleagues. Later on in retrospective studies, in which the medical outcome of a fixed number of patients are studied, supportive evidence of a claim is analyzed. However, a promising approach is to do continuous studies to achieve personalized medicine. Data containing the information of many patients can be stored and with relative ease more and more patients can be added which is statistically favorable. Furthermore, the rich oncology data has been known for its well suited candidacy to be applied to big data analytics in order to improve cancer treatment [15].

The Bayesian approach is a mathematical tool that ca be used to perform big data analytics. Such approach has been used in different areas such as in finance and oceanic research [26, 21]. Health care is yet to fully benefit from such approach. The main idea is to use a Bayesian approach to construct a general framework to create prediction models based on growing evidence using clinical data since the data of new patients could be added with ease. This approach allows to move forward towards a continuous study instead of relying purely in retrospective studies.

Intuitive Bayesian reasoning
Getting a feeling of the Bayesian reasoning is quite important to understand the mathematical formulation later described in the chapter. The human brain often does Bayesian reasoning by constantly searching for patterns and constantly doing intuitive guesses to predict an outcome which comes quite handy even for daily tasks. The brain looks for patterns quickly, we look at the clouds and the brain finds figures such as shapes of animals; it is mostly a matter of how quickly we look and then we see what we want to see. Directing our eyes towards the full moon and the smiling face could be found, or legends of human beings becoming mountains such as the Iztaccihuatl Aztec legend because a mountain resembled a woman laying down on the ground. The brain finds patterns nearly everywhere, and organize thoughts in order to cases. The observed pattern helps us create a belief and the strength

of the belief is modified in the next observation. The brain is constantly modifying its believes, according to observations. Let us look at a quick brain experiment.

Quick brain experiment

What is the probability of throwing a coin and landing tails? And heads?

If the answer is 50%, then it was assumed that the coin is fair. The collision of the two great statistic branches, the frequentists and the Bayesians, can be distinguished from this extremely simple brain experiment. Frequentists require to know a parameter linked to whether the coin is fair or not while a Bayesian approach can guess if the coin is fair or not by using the previous belief that most coins are fair. Bayesian reasoning can be summarized from this innocent example in which there was a previous belief of the coin fairness. In the Bayesian approach, previous believes are constantly being modified due to experiences. For instance, if a person keeps encountering that coins are fair, then the belief that coins are fair keeps getting stronger.

Description of the chapter

The chapter begins by a quick review of essential statistical concepts to understand the Bayesian formalism, then Bayes' theorem is presented. Following, the full practical framework to develop clinically based prediction models is presented then exemplified. The data mining concept is discussed which refers to searching and looking for specific patterns in a dataset, in this case medical oncology data. Machine Learning modeling is then introduced and its potentiality is emphasized. Machine learning refers to developing algorithms to teach the machine, computer, to learn automatically. Following, practical examples are presented. In certain cases a problem involving Bayes' theorem can be solved analytically specially taking advantage of conjugate priors. Nonetheless, in practice it can seldom be solved analytically and a numerical approach is an excellent alternative. The numerical approach used (recommended) in this work is the Metropolis-Hastings algorithm M-H which is a type of Monte Carlo Markov Chain method. In general terms the current trend of the revival of the Bayesian approach is thanks to the increase of the computer power that allows for the accomplishments of large data storage, manipulation and for numerical approximations to be carried on. In other words, in practice, numerical approximations are generally the norm.

The model building framework is exemplified by developing neurologic

grade prediction models using the Glioblastoma database. First, a simpler model was developed in which the Neurologic grade is predicted after the treatment based on the initial neurologic grade before the treatment. A second, enhanced, model is developed which also predicts the neurologic grade before the treament but it uses as an input an additional parameter, Gross Tumor Volume (GTV), which is related to the size of the tumor.

2.1.1 Three useful inference objectives

The three useful inference cases that the Bayesian formulation helps us achieve are discussed. The first is parameter predictions, the second is the estimation of data values, and third is the model comparison. The parameter prediction is useful to predict the value and its uncertainty. The second inference case is useful to make predictions for other datasets and the third inference objective allows us to compare several models. A review of the Bayesian methodology with emphasis on the three inference cases can be found in John Kruscke's book which it is worth taking a look at for additional reading in the subject [34].

Estimating parameters
The first objective is to estimate all the possible values of a parameter and the weight or likelihood of each of these values. This answer is responded by using the posterior function which is the solution of Bayes' theorem!

To better understand, let us illustrate it with a real example from our Glioblastoma database. We are often interested in finding correlations relating toxicities or complication parameters. For instance, lets investigate a parameter related to the skin toxicity after a radiation treatment. Lets call this parameter "Toxicity frequency" θ ranging from zero meaning no change of developing skin toxicity and 100 being completely certain (100%) of developing the toxicity. We are going to relate this parameter to the type of radiotherapy technique used during the treatment, either 3D, ArcTherapy, or TomoTherapy.

The number of patients who developed or not a skin toxicity are recorded in table 2.1 and from this data, we can construct our likelihood function and ultimately the posterior function. In a later section we will show specifically how we solved for the posterior function analytically but for now let us assume we know the posterior functions answered by Bayes' theorem. The posterior functions for each of the three possible cases are the density functions plotted

	3D	ArcTherapy	TomoTherapy	Total
Toxicity	6	3	7	16
No toxicity	53	14	7	74
Total	59	17	14	90

Table 2.1: Technology related skin toxicity data. The number of patients who developed skin toxicity or not corresponding to each type of technology used are recorded.

in figure 2.1. For instance, the posterior or density function corresponding to 3D is equals to $P(\theta_{3D}|D = Toxicity, No\ toxicity) = P(\theta|D = 6, 53)$ and it tells us the range of parameter θ_{3D} and the weight of the believe of this parameter according to the value. Our strongest belief of the value of this parameter is a value of around 10% seen from the toxity frequency corresponding to the peak; the width of the graph represents the uncertainty of the parameters. With the posterior distributions we have covered the first inference objective.

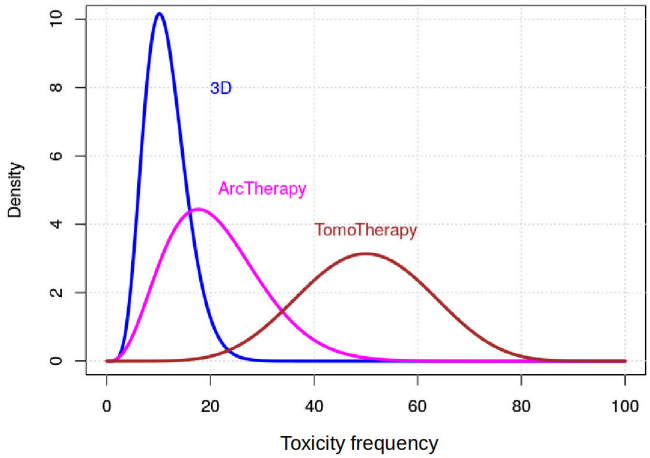


Figure 2.1: Technology related skin toxicity induction pdf (probability density function) for the three possible cases: 3D, ArcTherapy, and TomoTherapy. The peak is related to the most likely value and the width of the graph represents the uncertainty of the prediction.

Finding data values
The posterior or density functions previously calculated can be used to predict the next data value. For instance, we are interested in calculating the probability of developing skin toxicity following the treatment for some other new patient, using one of the three radiotherapy techniques previously discussed. In other words, we are interested in making a prediction for new data. More specifically a patient would have around 10% chances of developing skin toxicity if using the 3D technique during the treatment and a little bit less than 20% if using the ArcTherapy, and finally around 50% chances if using the TomoTherapy technique. These three values correspond to the peaks of the Toxicity frequency. We can predict in multiple ways; we can use the peak as the a reference as we just did but we can also use the mean, the median etc but the objective of making a prediction for new data values remains the same. The great advantage of having the posterior function of the parameter is that the uncertainty is estimated as well.

Comparing models
The last inference objective is easier to understand but not so straight forward to be carried on. Several different priors can be proposed therefore different results for the posterior can be obtained. Each of different predictions correspond to a different model. So the objective is to determine which model is predicting the most accurately. Overall, choosing the most appropriate prediction is not always an easy task and a special attention should be emphasized when comparing the models.

Dilemma

Which prediction is the most appropriate? Not always an easy task.

2.2 Bayesian formalism

2.2.1 Review of basic statistical concepts

Three basic concepts are reviewed in this section to better understand Bayes' theorem; joint probability, conditional probability, and marginal probability. We will illustrate these concepts using a simple example. We have 100 coins which are either shiny or not. Some of the coins are made of gold and some are not; we are interested in testing the gold content. After the measurements,

we obtain table 2.2 that summarizes the results.

Joint probability
The joint probability is written as the probability of A and B,

$$p(A \cap B)$$

A visual representation of joint probability is shown in figure 2.2 (using the data in table 2.2), where the probability of a coin being shiny is represented in the first circle, and the second circle represents the probability of a coin being made out of gold. The overlapping section corresponds to the joint probability of $p(Shiny \cap Gold)$. The probability of a coin being shiny and made of gold is equals to the probability of a coin being made of gold and being shiny.

We can say that the probability of A and B is the same as the probability of B and A

$$p(A \cap B) = p(B \cap A) \tag{2.1}$$

This concept may seem basic but it is essential for the derivation of Bayes' theorem later on.

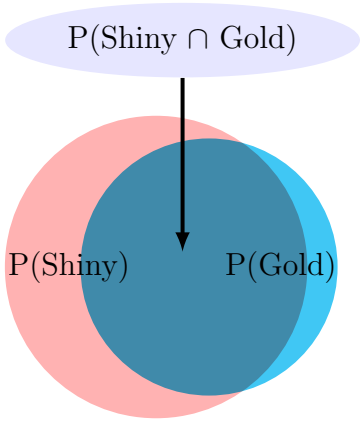


Figure 2.2: Joint probability illustration represented by the interception of the circles. The area of the circles and their interception are proportional to their corresponding probabilities. The probability of a coin being shiny is p(shiny) = 0.28, the probability of a coin being made out of gold is p(Gold)=0.20, and the intercept has a probability of 0.18.

Conditional probability
The mathematical notation for conditional probability is read as probability

	Gold Coin	Regular Coin	Coins
Shiny	18	10	28
Not Shiny	2	70	72
Coins	20	80	100

Table 2.2: Coin data for illustration

of A given B,

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

For instance, given the coin data in table 2.2 let us find out certain probabilities. First, the probability of being a gold coin given the coin is shiny is,

$$P(Gold|Shiny) = \frac{p(Gold \cap Shiny)}{p(Shiny)} = \frac{.18}{.28} = 0.64$$

which is not the same as the probability of the coin being shiny given it is a gold coin

$$P(Shiny|Gold) = \frac{p(Shiny \cap Gold)}{p(Gold)} = \frac{.18}{.20} = 0.9$$

Notice that,

$$P(Gold|Shiny) \neq P(Shiny|Gold)$$

It is worth highlighting that this formulation only works if A and B are dependent variables. The probabilities calculated say that gold coins are very likely to be shiny but not all shiny coins are made of gold.

Marginal probability

The last concept to understand before proceeding to Bayes' theorem is the concept of marginal probability. The marginal probability of the gold or regular coin illustration is the following. The marginal probability of a coin being shiny is,

$$p(Shiny) = p(Shiny\ and\ Gold) + p(Shiny\ and\ Regular) = 0.28$$

The marginal probability of a Gold coin is,

$$p(Gold) = p(Gold\ and\ Shiny) + p(Gold\ and\ Not\ Shiny) = 0.20$$

which in this case summarizes as the total number of Shiny (S) coins. In a more mathematical way the marginal probability of variable Gold (G) would be,

$$p(G) = p(G \cap S) + p(G \cap N) + ...$$

which is equal to,

$$p(G) = p(S \cap G) + p(N \cap G) + ...^{1}$$

According to the rules of joint probability $p(S \cap G) = p(S)p(G|S)$ Now, let θ be the brightness parameter in which it includes if the coin is shiny or Not Shiny. $\theta = [S, N]$ Then,

$$p(G) = \sum_{i=1}^{n} p(\theta_i)p(G|\theta_i) \tag{2.2}$$

The data was normalized on purpose to 100 coins, therefore the marginal distribution is the probability of a coin being made out of Gold. We could even say it is simply the probability of Gold. The marginal distribution can be used as a normalization constant with respect to a variable in this case G. This concept will be used as a normalization factor of Bayes' theorem.

2.2.2 Bayes' theorem

Reviewing joint probability, conditional probability, and marginal probability help us to better understand Bayes' theorem. Bayes' theorem provides a mathematical formula to determine the probability of A given B and the probability of B given A. Applied to the gold coin example, it provides a mathematical formulation to determine the probability of a coin being made of gold given it is shiny and the probability of a shiny coin being made of gold. I describe it as a bridge between $p(G|S)$ and $p(S|G)$. This invaluable bridge provides the foundation to study and develop models from clinical data.

Derivation
To derive Bayes' theorem the relation between the probability of A and B, $P(A \cap B)$, and probability of B and A, $p(B \cap A)$, are essential:

$$p(A \cap B) = P(A)P(B|A)$$

and

$$p(B \cap A) = P(B)P(A|B)$$

Thanks to equation 2.1 we know that,

$$P(A \cap B) = P(B \cap A)$$

[1] Variable inversion for later convenience when deriving Bayes' theorem.

Hence,

$$P(A \cap B) = P(B)P(A|B)$$

Rearranging items,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

thanks to our first relation

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

As a personal choice I prefer to derive it using A and B parameters but for practicality in this work I would substitute θ for A and D for B. So that it reads, probability of parameter[2] θ given D (in which D refers to data). Hence, Bayes' theorem is,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{2.3}$$

Where $p(D)$ is marginal probability

$$p(D) = \sum_{i=1}^{n} p(\theta_i)p(D|\theta_i) \tag{2.4}$$

$P(D)$ serves as a normalizing factor and it is often called the evidence when referencing to the Bayes' theorem.

Formulation
Bayes' theorem creates a wonderful bridge between the probability of a parameter or parameters θ given data D is $p(\theta|D)$ and the probability of data D given a parameter θ which is $P(D|\theta)$. For the moment it is sufficient to keep in mind this connection but it would be further emphasized progressively throughout the chapter.

It is customary to separate and name the Bayes' theorem in four parts: The Posterior, Likelihood, Prior, and the Evidence as shown in figure 2.3.

Previously we have seen that the Evidence acts as a normalizing factor, which is the marginal probability of parameter D over all possibilities. The Prior $P(\theta)$ corresponds to the confidence in the belief of parameter θ. For

[2]The notation is not restricted to a single parameter therefore θ can represent multiple parameters.

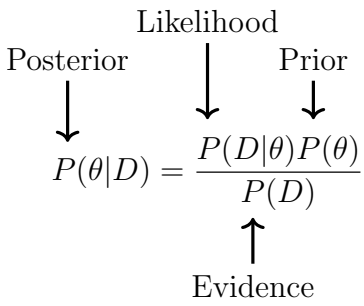


Figure 2.3: Bayes' theorem component names commonly used.

instance, in clinical data modeling we might have a complication parameter (e.g., Alopecia parameter) of which we have a previous belief that a certain oncology treatment causes hair loss. Prior is how confidence one is that hair would fall after that treatment. The likelihood is related to the observation, data or experiment and specifically means how likely or probable is an observation to happen. For instance, if out of three patients only one suffer from hair loss then how likely is this configuration to occur give the Alopecia parameter previously discussed. In mathematical terms the likelihood is,

$$P(D|\theta) = P(D_1|\theta) \times P(D_2|\theta) \times ... P(D_i|\theta) \tag{2.5}$$

In which D_1 refers to event one in data $D = [D_1, D_2, ...D_i]$ and i refers to the total number of patients. In product notations the likelihood is written as,

$$P(D|\theta) = \prod_{i=1}^{N} P(D_i|\theta) \tag{2.6}$$

However, when solving for the likelihood it is often much more convenient to use the log likelihood. A further description of the log likelihood can be found in the following reference [16]. The log likelihood is,

$$LL(D|\theta) = \sum_{i=1}^{N} ln\big(P(D_i|\theta)\big) \tag{2.7}$$

The Posterior is the new belief modified by the observations. That is, if one believed a treatment causes hair loss 60% of the time (that is, Alopecia parameter $\theta = 0.6$) and after observing that only one patient of three loss hair, then the new Posterior belief might be that perhaps a treatment does not causes as much hair loss as previously believed. That is, the new belief (Posterior) that θ is less than 60% seems reasonable.

Simple analogy to understand Bayes' theorem!

In simple words Bayes theorem can be read, a current belief (Prior) is altered after making observations or experiments (Likelihood) resulting in a new (Posterior) belief.

Choosing the prior

Deciding what the prior belief corresponding to a certain parameter θ is not always an easy task. It can be challenging to state mathematically such belief; the aim is to create a prior distribution function. The priors can be divided in two main branches, one branch corresponds to the informative priors and the other to uninformative priors. A branch diagram illustrating the two main branches is shown in figure 2.4. The informative prior correspond to

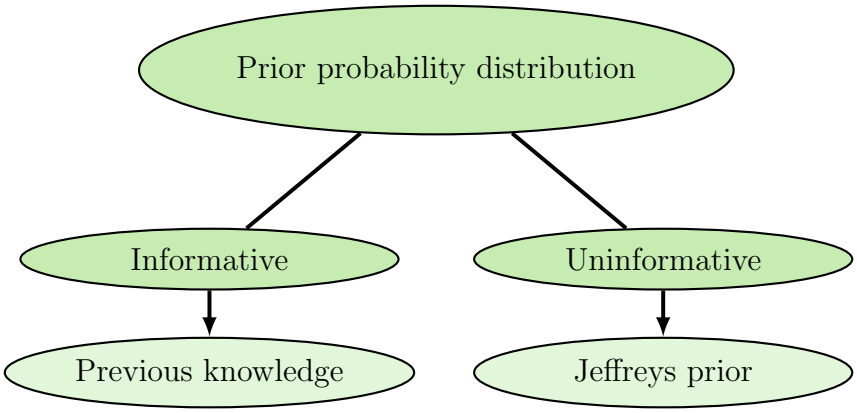


Figure 2.4: Two main prior branches, the informative where previous knowledge is known and the uninformative priors where no previous knowledge is known in that case Jeffreys prior is used.

those priors which can be constructed from previous information about the parameter of interest. A previous knowledge of the parameter can be found in a previous study in the literature or from previous data. For instance, in clinical model building we are often interested in complication or side effect parameters. If a study already has certain knowledge about a complication that study could potentially be used to construct the prior.

The uninformative priors correspond to the cases where no experimental or observed information about the parameter is known. In that case a prior probability distribution needs to be proposed in an educated manner. The prior distribution needs to be a function that can be normalized otherwise it would be an incorrect prior.

In order to come up with prior distributions, Jeffreys proposed an indirect

way for proposing them [49][3]. Jeffreys prior is the most widely used for the non-informative (uninformative) priors [17]. Notice that multiple names are used to refer to the uninformative prior such as non-informative, objective, vague, among other names. However in this work, we will use only the name uninformative for consistency.

Jeffreys prior is composed of the square root of the fisher information for the parameter of interest θ [23],

$$P(\theta) \propto \sqrt{det(I(\theta))} \tag{2.8}$$

$I(\theta)$ is the Fisher information which is related to the information that an observable variable carries about some unknown parameter θ. The fisher information equation is,

$$I_{i,j}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta_i\theta_j} log f(x|\theta) \right] \tag{2.9}$$

Where $\mathbb{E}$ means the expectation over x given θ

$f(x|\theta)$ is the likelihood and $\frac{\partial}{\partial\theta} log f(x|\theta)$ is the gradient of the log likelihood known as the score function. For instance, if the likelihood function is a binomial distribution and we solve equation 2.8 we would end up with

$$P(\theta) = \theta^{-1/2}(1-\theta)^{-1/2} = \frac{1}{\sqrt{\theta(1-\theta)}}$$

Despite the possible limitations of Jeffreys prior that can occur in certain cases, such as where the prior is not possible to be normalized, it is still a good departure point for uninformative priors in which no previous knowledge of variable θ is known.

2.2.3 Practical examples

Up to now we have worked towards the derivation of Bayes' theorem including a review of probability concepts to better understand the derivation. In this section the usage of Bayes' theorem is emphasized. First, a simple example is presented then a practical one closer to real life applications in order to better understand the application of Bayes' theorem in simple and practical cases.

[3] Section 2.3 in page 5 of the cited document.

Simple case
Let us create a simple hypothetical scenario in which a person is interested in learning about hair loss after a medical treatment "M". The person believes there is a 60% chances of loosing hair, but she is not 100% sure (she is actually only 50%) because she also believes it could be 45% or even 70% but she only believes half strongly of those two last percentages in comparison with the 60%.
Therefore the three priors would be (letting hair loss parameter, Alopecia, be equals to θ):

$$p(\theta_1 = 0.45) = 0.25$$
$$p(\theta_2 = 0.60) = 0.5$$
$$p(\theta_3 = 0.70) = 0.25$$

She feels the need to talk to patients who underwent the treatment, that is, she is gathering data. She talks to 5 patients and records that the first two patients lost their hair and the following three did not have any hair loss after the treatment. After that encounter (observations) what does she believes are the chances of loosing hair? Let us apply Bayes' theorem to respond the question.

The likelihood is how probable the configuration, of 2 patients loosing hair and 3 not loosing hair, is to occur. According to equation 2.5 the likelihood would be,

$$P(Patients|\theta) = P(Patient_1|\theta) \times P(Patient_2|\theta) \times ...P(Patient_5|\theta)$$

in this case it is just the multiplication of patients. Where the probability of not loosing hair is

$$P(Patient_i = No\ hair\ loss|\theta_i) = 1 - \theta_i$$

and the probability of loosing hair is,

$$P(Patient_i = hair\ loss|\theta_i) = \theta_i$$

Therefore, the likelihood of that configuration is simply the probability of not loosing hair times the probability of hair loss to the power of a (number of patients with no hair loss) and b (number of patients with hair loss) correspondingly. That is,

$$P(Patients|\theta) = (1 - \theta)^a \theta^b = 0.02304 \tag{2.10}$$

For θ_2 would be,

$$P(Patients|\theta_2 = 0.60) = (1 - 0.60)^3 0.60^2 = 0.02304$$

The evidence $P(Patients)$ is simply the marginal probability of patients who lost hair,

$$P(Patients) = \sum_{i=1}^{n} P(\theta_i) P(Patients|\theta_i)$$

She is interested in knowing the probability of hair loss after the treatment given her hair loss prior believes and the new observations she did by talking to the 5 patients. We can use Bayes' theorem 2.3 to answer her question,

$$P(\theta|Patients) = \frac{P(Patients|\theta)P(\theta)}{P(Patients)}$$

Therefore,

$$P(\theta_2 = 0.60|Patients) = \frac{P(Patients|\theta_2)P(\theta_2)}{\sum_{i=1}^{n} P(\theta_2)P(Patients|\theta_2)}$$

$$P(\theta_2 = 0.60|Patients) = \frac{0.02304 \times 0.5}{0.02325} = 0.4954$$

She had a prior belief for the hair loss parameter to be that 60% (chances of hair falling after the treatment) and the strength of that specific belief was 50% confident. After doing Bayes' theorem she believes almost with the same strength, 49.54% instead of 50%. Her belief did not change much for that particular Alopecia parameter of $\theta = 0.60$. At the beginning she believed half strongly for $\theta_1 = 0.45$ and $\theta_3 = 0.70$. Now she believes more strongly then before on θ_1 and less strongly for θ_3. As can be seen after doing the calculations which are display on the left side of figure 2.5 when comparing the plotted three points in the prior versus the posterior three points.

Practical case
It would be more reasonable to say that the person interested in knowing the alopecia parameter has a wider range of beliefs not just θ_1, θ_2, and θ_3. For instance a more legitimate prior would be composed of a myriad of believes ranging from 0 to 1.0. I chose 20 theta values and plot prior believes $P(\theta)$

using a normal distribution function since her prior beliefs are relatively symmetric center at $\theta = 0.60$ which can be observed in the upper right panel in figure 2.5. Notice how the prior starts to look more like a probability distribution than before.

In the simple case there were 5 observations, 2 people suffered from hair loss after the treatment while 3 remain with the same hair. In practical cases the more observations the better, in this practical case the observations were increased to 25, 10 lost hair, while 15 remain the same.

Notice that for the practical case, displayed on the right panel of figure 2.5, how the likelihood pushes the posterior curve to the left in comparison to the prior. It is important to keep in mind that in practical cases:

- Priors are often density functions
- It can be challenging to decide which prior to use
- The larger the number of observations the better
- The log likelihood is rather used
- Numerical approximations are often implemented

Priors are often density functions since as previously mentioned, it is of vital interest to describe parameter θ over all of its possibilities. Notice that θ does not have to be a probability ranging from 0 to 1; in this case it is but it is not restricted, it can even range from a negative number to a positive number. In real cases, choosing the right prior is not always an easy task because it is hard to state exactly what we believe on. Therefore it seems appropriate to analyze more in-depth the construction of the prior.

The next bullet point is more evident, statistically speaking the more data one has the better. However, in medical data acquisition can be a burden primary due to privacy issues. Authors in the literature start to discuss the issues concerning privacy [42]. As the number of patients increase the likelihood function values become extremely small since it is much less probable to have a large sample configuration than a small configuration because in the bigger configuration there are many more possible configurations. Hence it is convenient to use the natural log of the likelihood. This can be seen in comparing the likelihood of the simple case vs the practical one.

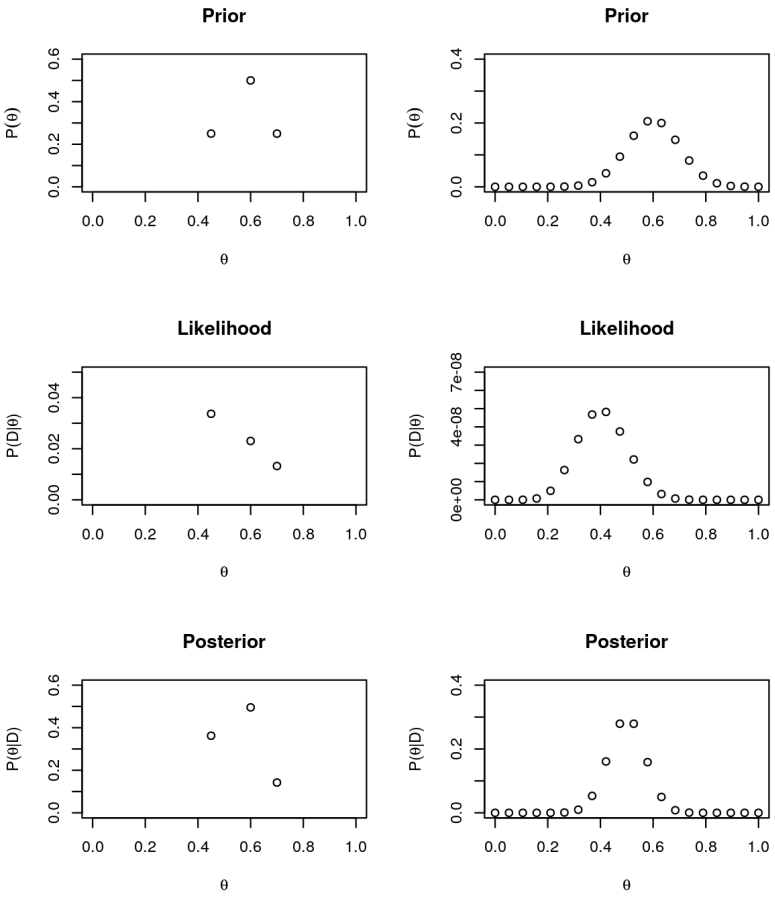


Figure 2.5: Prior, Likelihood and Posterior calculations for the simple case (3 calculations on the left panel) and for the practical case (20 calculations on the right panel). The practical case starts to resemble probability density functions.

2.2.4 Analytical and numerical solving techniques

In the ideal case we would like to solve Bayes' equation analytically. We could solve it analytically in the traditional way by multiplying the likelihood function times the prior and dividing by the evidence to get an exact answer. Unfortunately this option is restricted to simple cases and can not be applied to practical (i.e. often complex) cases in real life. Another option is to take advantage of the mathematical inverse relations known as the conjugate priors which will be explained in the analytical approach. When the analytical approach becomes such a burden or it cannot even calculated we rely on the alternative of numerical approaches which refers to finding an approximate but close solution to the problem. Relying on numerical approaches is very handy since we can solve more complex problems otherwise nearly impossible to solve or too tedious to be solved analytically. In this work we proposed the use of a Monte Carlo Markov Chain algorithm know as the Metropolis-Hastings algorithm (M-H) because it provides an effective and relatively fast approach to solve the problem.

Analytical approach
In the section for the simple and practical case we solved Bayes' theorem using numbers not mathematical expressions. We had 3 believes for the simple case and 20 believes for the practical one. From the practical example we became aware that the prior is often a probability distribution. We are going to retake the skin toxicity example shown at the beginning of the chapter but we are going to solve it by relying on a conjugate prior.

Recalling the problem: We investigated the skin toxicity frequency θ given a certain number of patients developed skin toxicity depending on the technique used (3D, ArcTherapy, TomoTherapy). The data corresponding to this is shown in table 2.1. Let us choose one of the three techniques, for instance the 3D case. Bayes theorem would look like this,

$$P(\theta_{3D}|D) = \frac{P(D|\theta_{3D})P(\theta_{3D})}{P(D)} \tag{2.11}$$

The first step is to understand the type of data that we have to fill out for each component of Bayes' equation in figure (2.3). We know the likelihood is related to how probable the configuration of the data is,

$$D = D_{3D} = [Patient_1, Patient_2, ...Patient_5]$$

with a total of 59 patients treated with the 3D technique. Where $Patient_i$ can be equals to either Toxicity or No Toxicity. To express out data in a mathematical way let Toxicity be 1 and no toxicity be 0. Hence the data would

be expressed as, $D = [0, 1, 1..1]$ extracted from the Glioblastoma database [4]. Where θ represents the probability developing toxicity and $1 - \theta$ represents

Patient ID	1	2	3	...	59
Toxicity Data	0	1	1	...	1
Probability	(1-θ)	θ	θ	...	θ

the probability of not developing a toxicity. By recalling that the likelihood is the product of individual probabilities we can obtain the general likelihood function,

$$P(D|\theta_{3D}) = P(Patient_1|\theta_{3D}) \times P(Patient_2|\theta_{3D}) \times ... P(Patient_{59}|\theta_{3D})$$

Notice that each individual probability can be represented by Bernoulli's distribution (shown in appendix A.2). Hence the product of Bernoulli's equations is the likelihood for this example which is,

$$\begin{aligned} P(D|\theta_{3D}) &= P(Patient_1|\theta_{3D}) \times P(Patient_2|\theta_{3D}) \times ... P(Patient_{59}|\theta_{3D}) \\ &= \left\{\theta_{3D}^{Patient_1}(1-\theta_{3D})^{1-Patient_1}\right\} \times \left\{\theta_{3D}^{Patient_2}(1-\theta_{3D})^{1-Patient_2}\right\} \times ... \\ &\quad \left\{\theta_{3D}^{Patient_{59}}(1-\theta_{3D})^{1-Patient_{59}}\right\} \\ &= (1-\theta_{3D}) \times (\theta_{3D}) \times ... \times (\theta_{3D}) \end{aligned}$$

Step one
Find likelihood

Lastly, the product of Bernoulli's distribution is proportional the Binomial distribution! Therefore we can further simplify the likelihood function to,

$$P(D|\theta_{3D}) \propto \binom{N}{n} \theta_{3D}^{n}(1-\theta_{3D})^{N-n} \rightarrow Binomial(N, n)$$

Where, N represents the total number of patients treated with the 3D technique (59 patients) and n represents the number of patients that develop a toxicity (6 patients) with this technique that is, D=[N=59, n=6]; these numbers can be found in table 2.1.

Step two
Find the prior

The next step is to define our prior function which needs to predict a probability of the belief ranging from 0 to 1 likely in a non-uniform manner. We choose a non-informative prior since we have no idea of the toxicity.

We are going to propose the beta distribution with parameters $\alpha = 1/2$ and $\beta = 1/2$ as our prior since this function, with those specific values, is

[4] The exact order of zero and one in this precise case was used to illustrate since multiplication order does not matter in the end result of the likelihood for this example.

the solution after solving for Jeffreys prior for a binomial likelihood function. The proposed prior is,

$$P(\theta_{3D}) = \frac{(1-\theta_{3D})^{\beta-1}\theta_{3D}^{\alpha-1}}{\beta(\alpha,\beta)} \rightarrow Beta(\alpha,\beta)$$

Step three
Find evidence

Where $Beta(\alpha,\beta)$ represents the Beta distribution. The next step is to determine the evidence, Using the data in table 2.1 we can calculate the evidence to be,

$$P(D) = \sum_{i=1}^{n} P(\theta_i)P(D|\theta_i) = \left(\frac{59}{90}\times\frac{6}{59}\right) + \left(\frac{17}{90}\times\frac{3}{17}\right) + \left(\frac{14}{90}\times\frac{7}{14}\right) = 0.177$$

Now we can multiply the likelihood times the prior normalized by the evidence and get an answer, which can be tedious. Instead we are going to solve it using a mathematical relation called the conjugate priors.

Advice

Use conjugate prior relations whenever possible to solve Bayes' theorem since it is much easier to directly obtain the posterior!

In this case, Bayes' theorem can be solved using the Beta-Binomial conjugate prior relation since the likelihood is a binomial distribution and we used a Beta distribution for the prior. If the likelihood function is a binomial distribution and the prior (the conjugate prior of the binomial likelihood) is a beta distribution the posterior is also a Beta distribution; a mathematical proof can be found in the A.7 from the appendix section. In this case Bayes' equation is,

$$\begin{aligned} P(\theta_{3D}|D) &= \frac{P(D|\theta_{3D})P(\theta_{3D})}{P(D)} \\ &\downarrow \\ Beta(\alpha_1,\beta_1) &= Binomial(N,n)\times Beta(\alpha,\beta) \end{aligned} \tag{2.12}$$

Where $\alpha_1 = (n+\alpha)$ and $\beta_1 = (N-n+\beta)$. The α_1 and β_1 parameters are derived in equation A.6 from the appendix section. Hence, $\alpha_1 = (6+1/2) = 6.5$ and $\beta_1 = (59-6+1/2) = 53.5$. Therefore the posterior function for the 3D case is simply,

$$Beta(6.5, 53.5)$$

We have solved indirectly using the conjugate prior relation, and it is much easier! Extra work was done in this section with the intention of

Likelihood $P(D\|\theta)$	Prior $P(\theta)$	Posterior $P(\theta\|D)$
Normal	Normal	Normal
Binomial	Beta	Beta
Poisson	Gamma	Gamma
Multinomial	Dirichlet	Dirichlet

Table 2.3: Some of the most used conjugate prior relationships.

demonstrating the process. In practical work, we only need to use the corresponding conjugate prior relationships according to the prior and likelihood cases; in table 2.3 some very useful conjugate priors are listed.

Numerical approach
Many posterior solutions cannot be obtained analytically so we have used conjugate priors in the previous section. However, we are limited by the type of prior we can use since the type of prior proposed depends on the likelihood. For instance, if the likelihood used is a Binomial distribution then we can propose a Beta distribution for the prior in order to also obtain a Beta distribution as can be seen in table 2.3. However, if we propose a different prior then the conjugate prior relations do not remain valid. Hence the solution is to approach the problem in a numerical manner.

In this work we propose the use of a Monte Carlo Markov Chain (MCMC) method called the Metropolis Hastings (M-H) to find the posterior function $P(\theta|D)$. The M-H algorithm is a versatile MCMC algorithm which was developed in the 1950s and later generalized by Hastings in the 1970s [12].

In general terms Monte Carlo methods refer to the use of random numbers, and MCMC methods refer to the use of random numbers for a calculation that immediately depend on a previous iteration. One example of this is the famous random walk simulation. Overall, the random walk consist of actually walking randomly (for instance, blindfolded) from a starting position in which each random step taken directly depends on the previous step since the previous step is the point of departure of the new step.

In this work we are going to use the M-H because it allows us to walk or cover the whole function over an entire parameter's range while still finding the most probable sections which is the peak or peaks. If we use an algorithm that only finds the maximum peak then that would be an opti-

mization algorithm but we are interested in finding all the possibilities no just the peak.

To illustrate the numerical approach using the M-H we are going to use the technology related skin toxicity data in table 2.1. In the analytical section we have solved the posterior distribution, using conjugate prior relations, for the 3D technique, which was equals to,

$$P(\theta_{3D}|D) = Beta(6.5, 53.5)$$

Similarly we can obtain the posterior distribution corresponding to the ArcTherapy technique,

$$\begin{aligned} P(\theta_{ArcTherapy}|D) &= \frac{P(D|\theta_{ArcTherapy})P(\theta_{ArcTherapy})}{P(D)} \\ &\downarrow \\ &\propto \prod_{i=1}^{N} P(\theta_{ArcTherapy}|D_i) \times Beta(\alpha, \beta) \\ &\propto Binomial(N, n) \times Beta(\alpha, \beta) \\ &\propto Binomial(17, 3) \times Beta(1/2, 1/2) \end{aligned} \tag{2.13}$$

We will call equation 2.13 the general posterior distribution and we are going to use it to illustrate the M-H.

We know the analytical solution, thanks to the conjugate prior relations, which is equal to,

$$P(\theta_{ArchTherapy}|D) = Beta(3.5, 14.5) \tag{2.14}$$

in which $\alpha_1 = (n+\alpha) = (3+1/2) = 3.5$ and $\beta_1 = (N-n+\beta) = (17-4+1/2) = 13.5$ and total of 17 patients ($N = 17$) and 3 ($n = 3$) of them develop some sort of skin toxicity after the radiation therapy treatment.

It makes no sense, other then for illustration purposes, to calculate it numerically since we know the analytical solution which is equation 2.14. Also, in this case we are very lucky that the likelihood reduces to the Binomial distribution, but what happens when we do not know the equation ruling each individual probability $P(\theta_{ArcTherapy}|D_i)$ or when Bayes' theorem requires multiple θ parameters? In such situation, the numerical method is a very useful approach.

We are solving the ArcTherapy problem numerically with the purpose of showing how the M-H works well and that even with a low number of

Monte Carlo iterations the numerical solution starts to approach the analytical posterior distribution. The solid line in figure 2.6 represents equation 2.14 and the dots represent calculations done using the M-H. Concerning the dots in the figure: a random number $\theta_{ArcTherapy}$ is generated and used to solve Bayes' theorem in equation 2.13 (posterior solution) represented by the solid dot (first graph); the $\theta_{ArcTherapy}$ value is accepted since it is the starting position. A new number is randomly generated with a delta value chosen by the user and is used to calculate again Bayes' theorem, this new posterior solution (solid dot in second graph) is compared to the previous posterior solution (empty circle in second graph) and if the probability is higher we keep the new $\theta_{ArcTherapy}$, which in this case it is clearly higher as it is visually observed in the second graph from left to right. The third graph shows the new posterior solution is lower (solid dot) then the previous posterior, hence the acceptance of new random value ($\theta_{ArcTherapy}$) generated is subjected to the condition shown in the box below

Condition

Let u = random number generated in the interval [0,1]

$$\text{if } \frac{posterior_{new}}{posterior_{previous}} > u$$

then we accept the random $\theta_{ArcTherapy}$ generated,
otherwise we reject the new value and keep the previous $\theta_{ArcTherapy}$ as the new value.

Similarly the fourth graph is subject to condition where as in the fifth graph it is quite clear the the posterior solution of the new $\theta_{ArcTherapy}$ is higher then the previous so we keep the new value.

- Keep doing the process for about a thousand times
- Get rid of the first couple hundred accepted values commonly known as the burn-in
- Store the accepted values of $\theta_{ArcTherapy}$
- Make a histogram of the accepted values and then we now obtain the shape of a posterior pdf for the ArcTherapy technique which starts to resemble the analytical solution given by equation 2.14 shown in solid line in figure 2.7.

In figure 2.7 we have shown that the analytical M-H can be used to estimate the posterior distribution and it is very helpful when we have multiple θ values which can become difficult to calculate. Hence with the numerical approximation the problem is much simpler in terms of computational time. Lastly, the M-H algorithm has demonstrated the ability to converge to the true posterior pdf, which makes it a powerful tool to discover the pdf.

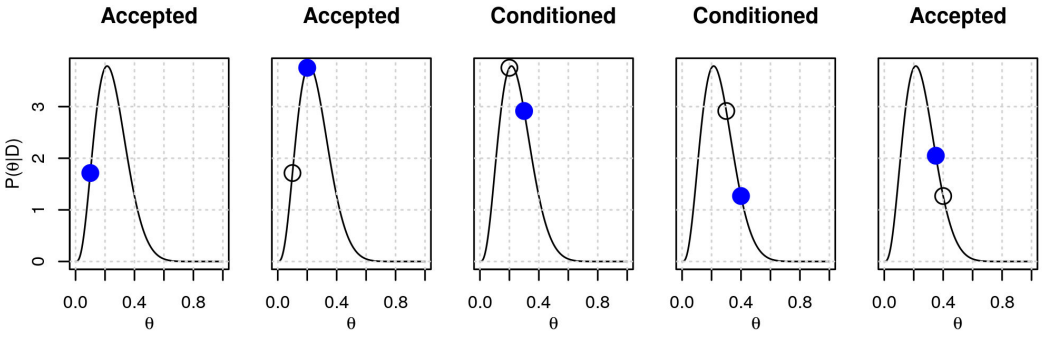


Figure 2.6: Posterior pdf generation using the Metropolis Hastings (M-H) algorithm for ArcTherapy technique in which new posterior points are created by using random $\theta_{ArcTherapy}$ numbers and accepting the posterior point only if it meets a specific condition depicted in the "Condition" box located in the text.

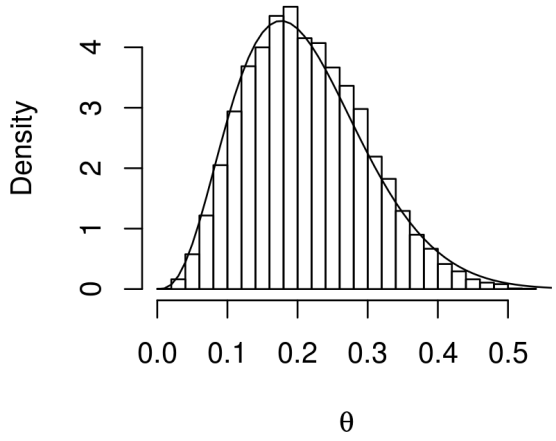


Figure 2.7: Posterior pdf illustration for ArcTherapy technique. The solid line represents the analytical solution and the histogram represents the numerical approximation using only a low number of iterations. It shows that even at low iteration numbers the numerical approximation starts to closely resemble the analytical solution.

2.2.5 Machine Learning

In this section a description of the framework for developing biophysical prediction models based on clinical data is presented in order to better understand the model building methodology. The Bayesian framework presented so far allows for a continuous learning approach in which the posterior function becomes the new prior for some other calculation. Hence, we can add new data to our database and run the code in order to get a new posterior distribution meaning that new data can keep improving the predictions.

The need for continuous learning from large and complex datasets led us into the Machine Learning modeling. A wide variety of Machine Learning tools for oncology are being used which includes, Artificial Neural Networks, Decision Trees, Support Vector Machines, and Bayesian Networks [31].

In this chapter we are going to focus on a Bayesian framework Machine Learning modeling approach which is illustrated in figure 2.8. The typical modeling approach we used to develop clinically based prediction models using machine learning and Bayesian statistics is as follows:

- A set of clinically based data is gathered

- Clinical observations are identified
- Certain known parameters specific to the patient are used as inputs to initialized the computer model
- The user makes an educated guess of the parameters that we are interested in finding out; this is the initialization of the model parameters.
- The patient parameters and the model parameters complete the computer model and the simulation is ran
- A prediction is made usually by a decision making algorithm such as the M-H numerical approximation presented in the previous section
- The prediction is compared with the clinical observations
- Prediction errors are determined, for instance by data fitting methods
- The model is adjusted and new model parameters are passed to the computer model to make a new simulation
- The decision making algorithm is ran thousands of times for the purpose of exploring the posterior joint probability density function (pdf) of the model parameters.

The workflow displayed in figure 2.8 describes the model building methodology which can be used as a guide to develop a wider range of clinically based models. It is worth insisting on the importance of related methodologies. In the literature, the importance of machine learning methods to help us understand cancer progression and treatment is highlighted [31, 13].

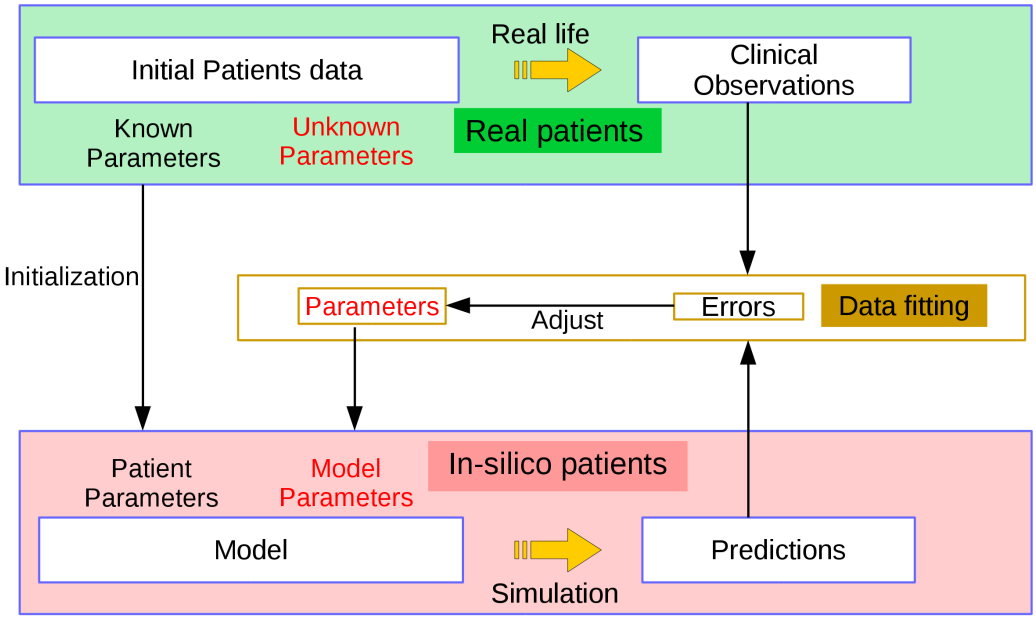


Figure 2.8: General Machine Learning diagram used as a guide to develop clinically based models. The objective is to find the unknown parameters by a continuous data fitting process to be able to make a prediction as close as possible to the clinical observations.

2.3 Applications of the Bayesian formalism in neurologic grade prediction models

We know how to solve a wider range of posterior functions using the numerical approximations so now we can build more complex models guided by the Machine Learning model building workflow presented in figure 2.8. A simpler model and an enhanced model were developed which explore the probability of developing neurologic issues after the first treatment. Identifying the early signs of neurologic complications are important in managing patients suffering from cancer [37]. Hence, neurologic complication models is of relevant interest. The same methodology can be applied to any kind of toxicity (skin, alopecia, ...), if the toxicity is correctly recorded and given that we have a sufficiently large relevant data.

The mathematical point of view of modeling is emphasized in these neurologic prediction models. The simpler model predicts the neurologic grade

after the first treatment based on studying the neurologic grade before and after the first treatment[5]. The enhanced model additionally requires the Clinical Target Volume CTV with the purpose of directing towards a more realistic prediction. We chose the CTV over the GTV because there was a slightly better correlation between CTV and neurologic complications. The tumor size is often relied on for assessing response to therapy, however not a great number of studies have performed clinical correlation of tumor size with patient outcome [14] .

2.3.1 Neurologic grade data

Let us follow the Machine Learning diagram in figure 2.8 to develop the neurologic models. The first step is to gather, from the Glioblastoma database, the clinical observations that we are interested in modeling which in this case are the neurologic grade before and after the treatment.

The number of patients according to their neurologic grade before and after the treatment are presented in figure 2.9. For instance, 26 patients started with neurologic grade of one before the treatment and remain with the same neurologic grade of one after the first treatment. From this figure we can see that the larger the circle the higher the number of patients corresponding to each case. There are 25 possible cases, five initial grade cases starting from no neurologic complications (grade 0) all the way up to grade 4 which is the grade with the highest neurologic complications. Even though we have the data of about 90 patients we can clearly observe that only the neurologic grades of 0 and 1 are more relevant.

[5]Due to the aggressive nature of Glioblastoma multi-treatment approach might have been performed.

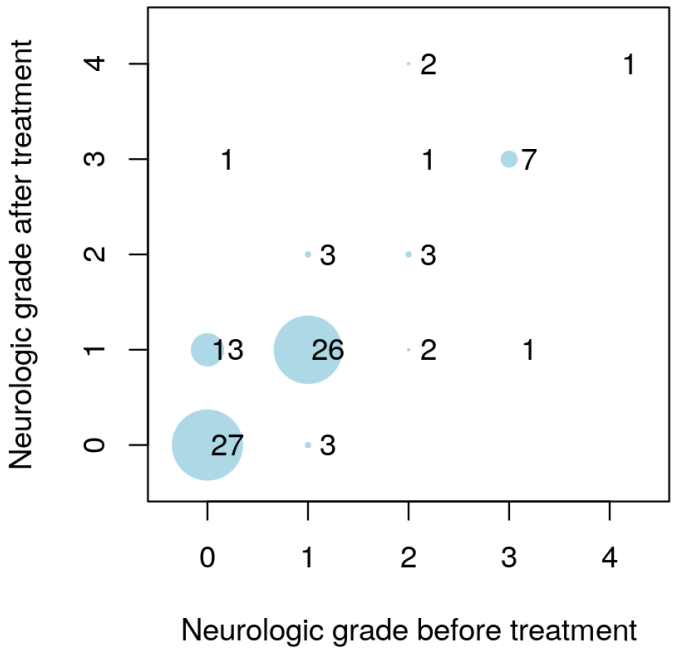


Figure 2.9: Neurologic grade change before and after the treatment. The number of patients are recorded according to each case; the higher the number of patients the larger the corresponding circle.

2.3.2 Simpler model setup

We are going to use the neurologic grade data in figure 2.9 for those patients who started with neurologic grade 0 in order to illustrate the modeling process. The number of patients corresponding to each neurologic grade after the treatment were used as input parameters for the model. To construct the model we need to construct Bayes' theorem, thus we need to create the likelihood function and propose the prior distribution.

First we are going to calculate the likelihood. For the likelihood we are interested in finding the probability of the configuration of 27 patients remaining with neurologic grade 0, 13 with grade 1, 1 with grade 3 and no patients with grade 2 nor grade 4 after the first treatment (Data=$D[x_1$=27, x_2=13, x_3=0, x_4=1, x_5=0]). The following multinomial distribution is a good candidate to predict such configuration therefore the likelihood is,

Construct the likelihood

$$P(x_1, ..., x_k|\theta_1, ..., \theta_k) = \frac{\Gamma(\Sigma_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} \theta_i^{x_i} \tag{2.15}$$

Propose a prior

Where parameter θ_1 represents the probability of developing grade 0, θ_2 the probability of developing grade 1 etc. A reasonable prior to propose[6] could be a prior of the form of a dirichlet distribution since the likelihood is a multinomial distribution. Hence the prior chosen was,

$$P(\theta_1, ..., \theta_k) = \frac{\Gamma(\Sigma_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{2.16}$$

where $\alpha[\alpha_1{=}1, \alpha_2{=}1, \alpha_3{=}1, \alpha_4{=}1, \alpha_5{=}1]$ would be for a uniform prior or we could set a value of $1/2$ for each θ component to make it a Jeffreys prior.

Lastly the evidence is given by,

$$p(D) = \sum_{i=1}^{n} p(\theta_i) p(D|\theta_i) \tag{2.17}$$

Solve for the posterior

Bayes' theorem can be solved analytically using conjugate prior relations in table 2.3 therefore for this model the posterior is the Dirichlet distribution,

$$P(\theta_1, ..., \theta_k|a_1, ..., a_k) = \frac{\Gamma(\Sigma_{i=1}^{k} a_i)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} \theta_i^{a_i - 1} \tag{2.18}$$

Marginalize the posterior pdf

Where $a_i = (\alpha_i + x_i)$ The issue with the posterior in equation 2.18 is that it is a five dimensional distribution therefore a method to represent it in a one dimensional manner is needed. Hence, we need to marginalize the dirichlet distribution. The analytical solution of the marginalized dirichlet distribution is the following beta distribution,

$$P(\theta|a_i, b_i) = \frac{(1 - \theta)^{b_i - 1} \theta^{a_i - 1}}{B(a_i, b_i)} \tag{2.19}$$

where $B(a_i, b_i)$ is the beta function described in the appendix A.5 and $b_i = (a_0 - a_i)$.

The model was solved in both ways, by the analytical approach using conjugate prior relations and numerically using the M-H approach. For the numerical solving technique educated random θ values are proposed. The

[6] For more information in choosing a prior please take a look at Bayes' theorem section.

decision making algorithm (M-H) keeps track of the accepted θ values. Then histograms are created for each of the final neurologic grade cases. From the histogram shapes we can determine the marginalized pdf. A diagram of the workflow highlights (figure 2.10) the comparison between the analytical and numerical approach in order to obtain a one dimensional pdf of the parameter of interest. The results are presented in the following section.

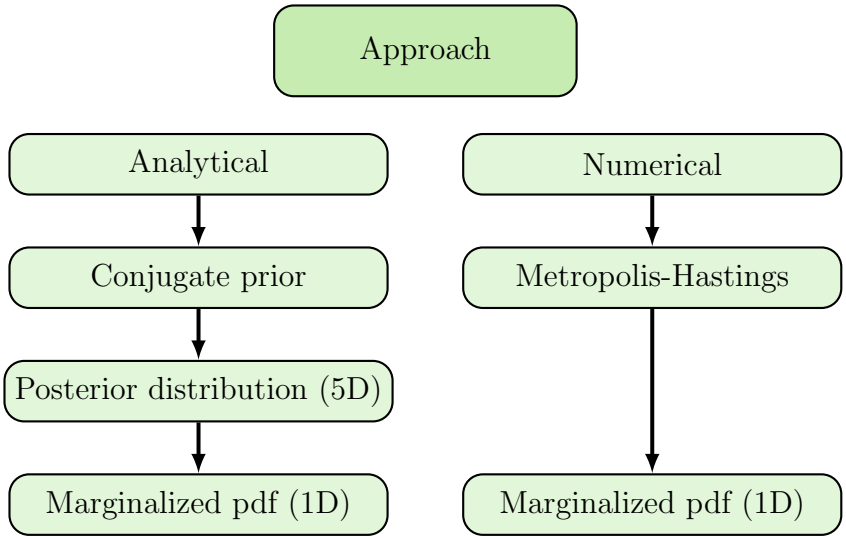


Figure 2.10: Workflow, for the simple model, comparison between the analytical and numerical approach. The analytical approach is possible especially when using conjugate priors otherwise the numerical approach is recommended.

2.3.3 Simpler model predictions

The normalized marginal probability density functions (pdfs) for each of the θ_i parameters corresponding to the five possible neurologic grades after the treatment are plotted in figure 2.11. We rely on equation 2.19 to obtain normalized analytical pdfs which are represented in solid lines. Recall that the θ_i parameters represent the probability of developing a neurologic grade given the patient started with neurologic grade 0. Finding the posterior distributions of the parameters θ_i accomplishes the first inference objective. The second inference objective was to predict a data value for another set of data. This objective is accomplished by using the pdfs in figure 2.11 for the prediction of the probability of developing a neurologic grade for an arbitrary

new patient. For instance, according to the pdf corresponding to the final grade 0, a patient has about 60% chances ($\theta_1 \approx 0.60$) of remaining with no neurologic complications, this is depicted by the peak of the pdf. Similarly a patient has about 30% chances ($\theta_2 \approx 0.30$) of developing a grade 1 after the first treatment. The higher the vertical value the more probable θ is. The width of the pdfs correspond to the uncertainty of the prediction.

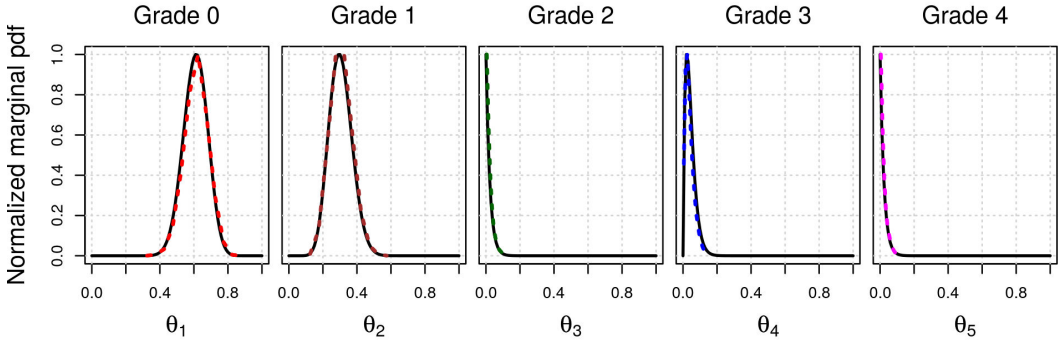


Figure 2.11: Normalized marginal pdfs representing the probability of developing a neurologic grade given patients started with grade 0. The solid lines represent the analytical solutions and the color dotted lines represent the numerical approximations.

The numerical predictions are overprinted in figure 2.11 using dotted lines. The importance of plotting together the analytical and numerical pdfs is to show how closely the numerical approximation is to the analytical prediction. This resemblance can be used to validate the numerical algorithm. Markov chains and autocorrelation functions (ACF) were plotted in order to verify the stability of the numerical approximation. In figure 2.12 we can diagnose and analyze the Markov Chains using the Auto Correlations (AC) (please see the appendix for more information); these simulations are stable. In addition the first two histograms have a more defined form which seems logical since θ_1 and θ_2, associated with final grade 0 and grade 1 respectively, include more data (27 and 13 patients correspondingly as stated in figure 2.9) than for the rest of the neurologic grades after the treatment corresponding to initial grade 0. The numerical approximations depicted by the dotted lines in figure 2.11 are created using the histogram shapes in figure 2.12. The middle (red) vertical solid line over the histogram correspond to the mean of the corresponding θ_i value, the two outer lines immediately after the mean correspond to the first quartile, and the outer most vertical lines correspond

to the second quartile.

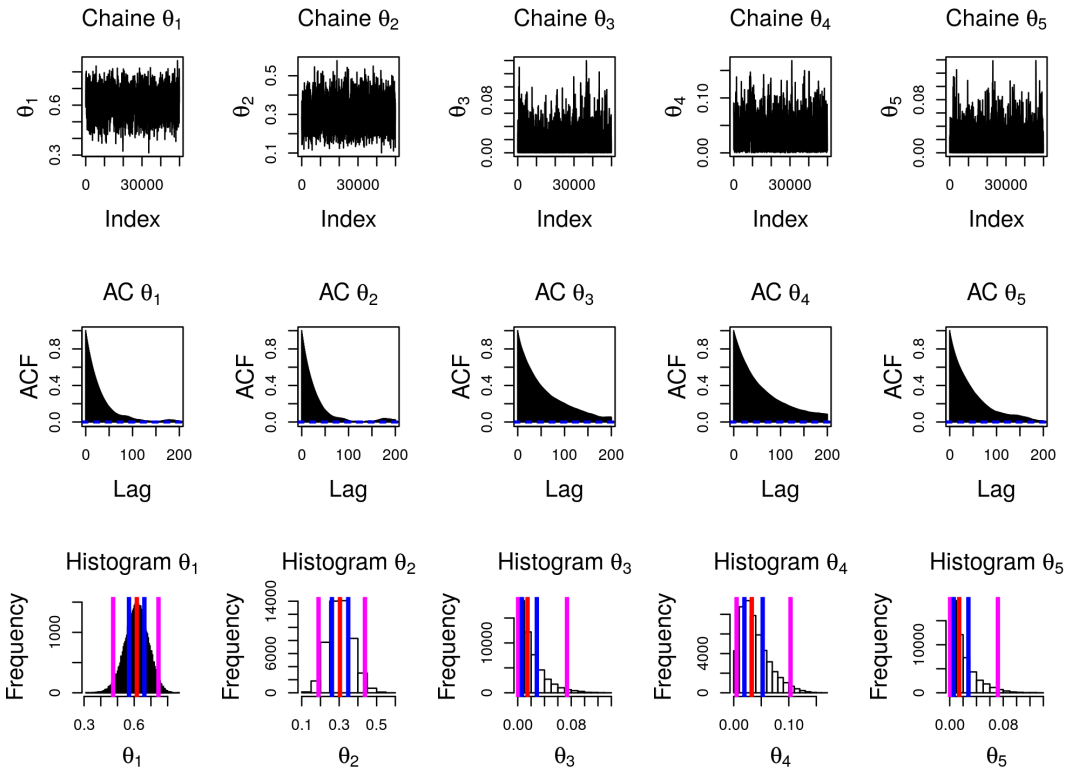


Figure 2.12: Markov chains and Auto Correlations (AC) show the more stable numerical simulations correspond to the θ_1 and θ_2 parameters. The simulation results in a histogram for each θ_i is direcly linked to the pdfs depicted in dotted lines in figure 2.11.

2.3.4 Enhanced model setup

The objective is to work towards a more realistic model in which other parameters, in addition to the neurologic grade, are taken into account. The enhanced model requires additionally the CTV. First we plot the neurologic grade after the treatment versus the CTV only for grades 0 and 1 as shown in figure 2.13 because significantly more data is available for these two grades.

The objective is to find a function that can predict the neurologic grade with respect to the CTV. For this work the most relevant issue is not nec-

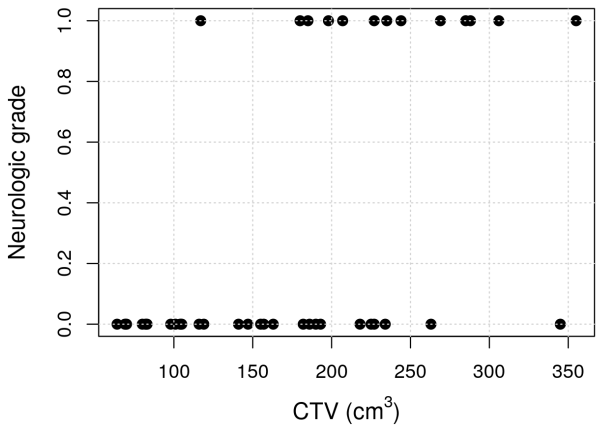


Figure 2.13: Neurologic grade after the treatment vs CTV in which each point represents a patient. Patients with smaller CTV sizes did not develop neurologic grade 1 immediately after the treatment. This data is for patients who started with no neurologic issues before the treatment.

essarily the proposed function but the methodology to numerically solve the unknown parameters of the proposed function. The logistic function is a type of sigmoid function that is widely used when dealing with binary data. Hence we proposed the logistic function, to fit the data in figure 2.13,

$$f(CTV, a, b) = \frac{1}{1 + e^{-(CTV-a)b}} \tag{2.20}$$

with unknown parameters a and b. $f(CTV, a, b)$ is the probability of developing a neurologic grade 1.

The aim is to solve for the posterior distribution of parameter a and b. Since it is not possible to solve the problem analytically we will rely on the numerical approach. First we start by constructing the likelihood which is,

$$P(GrF|CTV, GrI, a, b) = \prod_{i=1}^{N} P(GrF_i|CTV_i, GrI_i, a, b) \tag{2.21}$$

The likelihood is the product of individual probability, and each individual probability can be calculated by the proposed logistic function,

$$P(GrF_i|CTV_i, GrI_i, a, b) = \begin{cases} f(CTV_i, a, b) & GrI_i = 1 \\ 1 - f(CTV_i, a, b) & GrI_i = 0 \end{cases} \quad (2.22)$$

Where GrF refers to the grade after the treatment, GrI to the grade before the treatment and the index i refers to the patient number.

Following, we proposed a uniform prior distribution for a and b to simplify the problem. Then the posterior distribution of a and b was determined using the logistic function completed with these two parameters. A wide range of parameters are available from the numerical simulation, hence there are many possible solutions to the logistic function. The solutions are presented in the following section.

2.3.5 Enhanced model prediction

The normalized marginal pdf for parameter a and b are plotted in figure 2.14. The a parameter ranges from about 160 to about 320 and has units of cubic centimeters and b ranges from greater than zero to about 0.06. The peak represents the most likely value and the width can be interpreted as the uncertainty of the parameter. The most likely value for a is around 225 cm^3 and about 0.025 cm^{-3} for b. The plot of the values of parameter b versus a displayed in figure 2.15 allow us to see the most likely parameters in which the color of each hexagon is associated to a certain number of simulations. It is worth noting that the joined pdf (a vs b in this case) contains all the information of our dataset for the model. It can be used as the prior for a and b when new data becomes available.

The stability of the numerical simulation can be observed from the Markov chains in figure 2.16 in which the a and b values vary randomly around a certain value covering the full range of parameter a and b meaning the simulation is stable. The logistic function with the most likely a and b values is plotted in figure 2.17 represented as a solid black line. About a hundred other a and b values were also used and the gray hallow circles next to the main solid line represent those 100 plots. Each black solid circle, horizontal to the zero or one grade value, represents a patient with its respective CTV size.

The enhanced model by using the completed logistic function can predict the probability of developing grade 1 given the patient started with no neurologic toxicity. It is not sufficient to determine the probability but also the

uncertainty of the prediction which can be visually observed with the hallow gray circles in image 2.17.

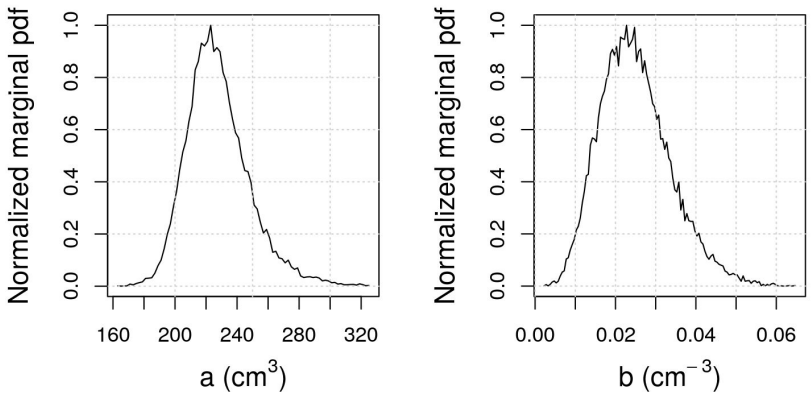


Figure 2.14: Normalized marginal pdf for parameter a and b. The most likely value of a is about 225 and about 0.025 for b.

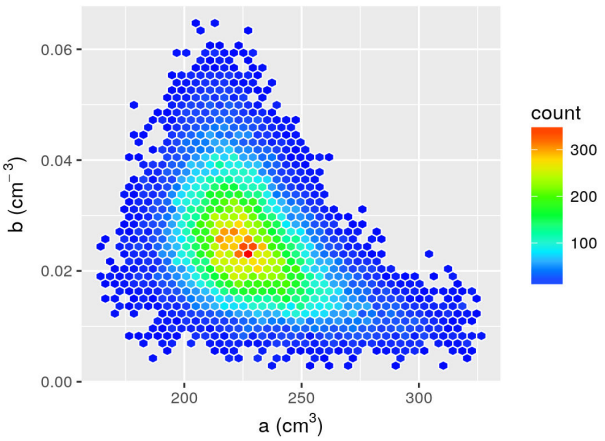


Figure 2.15: Value of parameter b versus a for each simulation in which the color of the hexagons represent a certain number of counts of the simulation. The most common value of a is about 225 and about 0.025 for b.

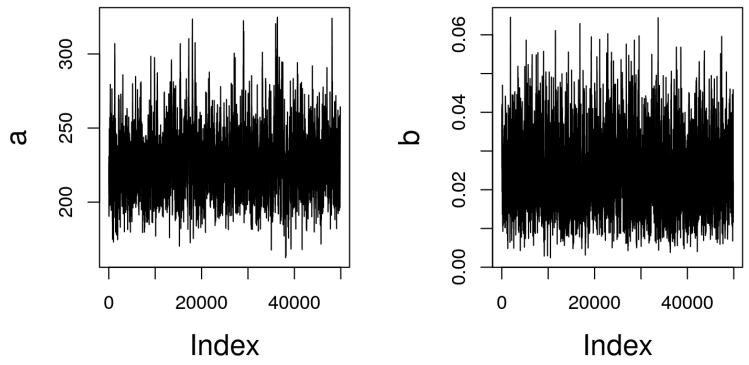


Figure 2.16: Markov chains for parameter a and b. The stability of the numerical simulation can be observed from the Markov chains in which the a and b values vary randomly about the most likely value while covering the full range of the parameters a and b.

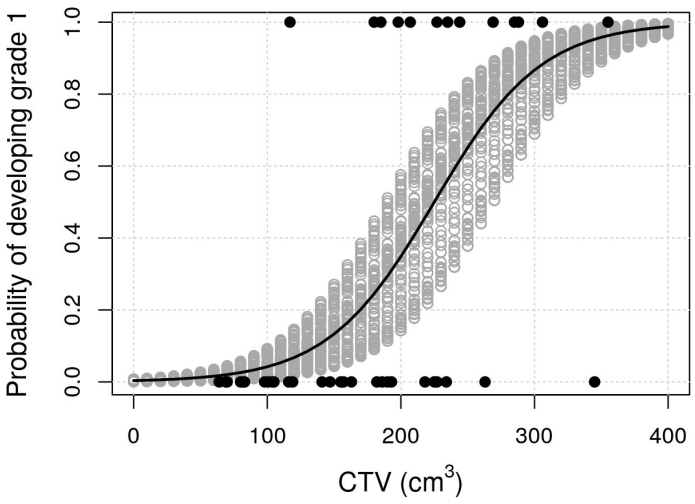


Figure 2.17: Probability of developing neurologic grade one based on CTV size. The solid black line is the logistic function with the most likely a and b parameters. The hallow circles surrounding the solid line represents about 100 logistic functions with other a and b parameters chosen randomly. The solid black circles at the very bottom and top represent the original data to be fitted.

2.3.6 Simple versus enhanced model comparison

The simple model versus the enhanced model are compared in order to notice the high dependence on the CTV size. From figure 2.17 we can observe the uncertainty of the probability of developing grade one (parameter θ_2) depending on the CTV. As an illustration let us observe three different cases for the enhanced model; for CTV=100, 200, and 300 cm^3. The uncertainty represented in figure 2.17 is bigger for the 200 case then for the 100 case. The uncertainty depicted by the vertical width created by the gray circles correspond to parameter θ_2, and those probabilities are plotted for the three cases in figure 2.18. The simple model predicts a θ_2 about 30% chances developing a grade 1 where as the enhanced model greatly depends on the CTV size. The enhanced model provides a more realistic prediction than the simple model which does not take into account the CTV. Hence the enhance model provides advantages in terms of personalized medicine in which specific patient parameters are taken into consideration for making a prediction.

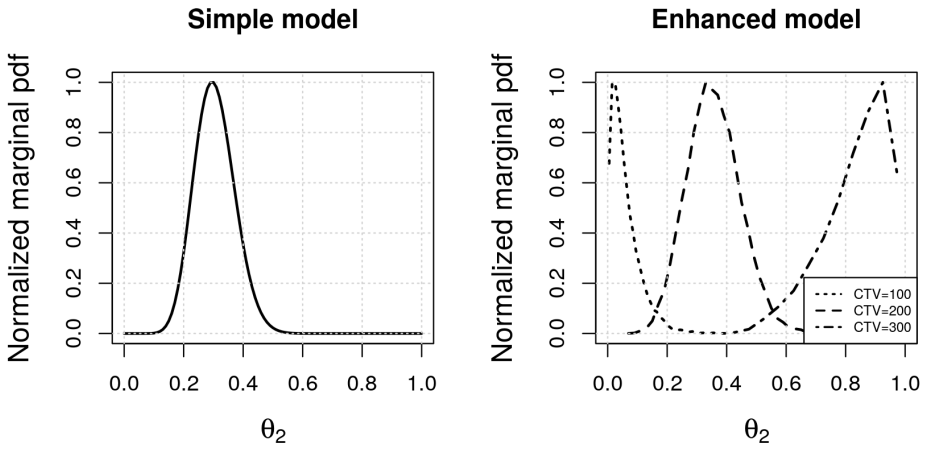


Figure 2.18: Normalized marginal pdfs for parameter θ_2 for the simple model and the enhanced model. Three pdfs are plotted for the enhanced model using three different CTV sizes. These plots show that parameter θ_2 is highly dependent on the CTV size.

2.4 Conclusions

The framework to develop clinically based models with the use of Bayesian statistics and Machine Learning was presented in this chapter. A comprehensive and intuitive construction of Bayes' theorem was described starting from basic statistical concepts. Thereafter, we proceeded to construct the likelihood, proposing a Prior, and determining the evidence. Proposing a prior can be challenging since it is hard to clearly state what it is previously believed in about a parameter of interest. It was shown that determining the posterior distribution of the parameter of interest is essential for practical cases. The three main inference objectives when using Bayes' framework were emphasized. Real life examples of clinical data of patients suffering from Glioblatoma were used to develop the model building methodology. The neurologic enhanced prediction model demonstrates the high dependency on the CTV size. The potential of the numerical approximations especially for large datasets was emphasized.

The comprehensive intuitive build-up of Bayes' theorem described using real life examples and quantities intends to reach a wider range of audience. For instance, a wider audience can have a grasp of the meaning of skin toxicity, and neurologic complications following a radiotherapy treatment. Quantities related to the size of the tumor seem intuitive to have an impact on the prediction. However there can be other quantities that are not so intuitive to guess that are playing a significant role. As an example, the numerically calculated a and b parameters of the logistic function, start to have a more difficult real life meaning. Hence, we can see that very easily we can be confronted with many parameters that mostly have a mathematical meaning. One way to try to identify those unknown parameters is to keep developing models similar to the neurologic enhanced model presented in this work. In order to develop more sophisticated models we need the database to keep increasing. As a general rule, the larger the database the better it is for developing more complex models. One of the toughest part of the Bayesian methodology is to determine a reasonable prior, though the influence of the prior is reduced as the volume of data increase. Lastly, a main significance of the posterior pdf is that it contains all the information of the dataset.

The methodology in this chapter serves as foundation work for the construction of models in the PMRT (Plateforme de Modélisation pour la Radiothérapie) currently being built at LPC-Caen in collaboration with other institutions. It is worth noting that this work does not intend to provide a medical interpretation of the results but it is focused on the mathematical

approach on modeling using real clinical data. For instance, the use of CTV in the enhanced model instead of GTV was motivated by a better correlation with the observed outcome. This observation suggests that the outcome is related to the treatment rather than to the patient, but it is only a proposal, not a demonstration. Such statement is made with the premise that the GTV is inherent to the patient while the CTV is part of the treatment since by definition the CTV is the GTV plus a certain margin belonging to a protocol and not inherent to the patient. The methodology demostrates the potential of the predictive capability of the Bayesian statistics together with the Machine Learning approach. This approach can be especially useful for large databases which directs towards a personalized predictive approach.

Tumor recurrence analysis

3.1 Introduction

The purpose of this chapter is to study the medical images in the form of MRI images and CT scans, from the Glioblastoma database, to explore tumor evolution with special attention to the prediction of the location of the

recurrence using pre-treatment imaging.

The medical images, including the Diffusion Weighted (DW), T2-Flair, and T1-Gd MRI (Magnetic Resonance Imaging) sequences as well as the CT (Computer Tomography)-scans, have become essential in the diagnostics and treatment of brain cancers. The DW-MRI has gain particular recognition in the study of brain cancer. Kono et al., hypothesis is that DW imaging by means of the Apparent Diffusion Coefficient (ADC) values could potentially provide further valuable information in diagnosis of brain tumors [29]. A particular intriguing relation of a change in the intensity values of the Glioblastoma recurrence area (from the pre-treatment images) was found in a thesis work done by Emmanuel Meyer [44]. We decided to further analyze such relation. His previous work consisted of manually selecting three regions from the pre-treatment DW-MR images: (1) region corresponding to the inside of the recurrence, (2) region corresponding to a location far from the tumor (healthy region), (3) region corresponding to a non-recurrence location near, but outside, the tumor contour. The ADC[1] values were lower in the recurrence (region 1) compared to the non-recurrence peritumoral (region 3).

The observations suggest that the ADC pixel values of the MRI images could provide helpful information for predicting the recurrence location. Along a similar line, it is mentioned in the literature that tumor ADC values could potentially contribute useful information in the diagnosis of brain tumors [29]. The potential useful information goes beyond diagnostics. For example, DW imaging is useful in diagnostics but also in: grading tumors and amount of tumor infiltration, looking for early responses or progression, and in evaluating the residual or recurrence of the disease [54].

Guided by the supportive information that the medical images play a role in brain cancer analysis, we decided to further investigate the influence of DW, T2-Flair, and T1-Gd MRI sequences with special attention to the ADC values of the DW-MR. Recently Chang et al., have observed a small but statistically significant changes in the intensities of the ADC and Flair values of the Glioblastoma recurrence regions [9]. Additionally, our work allows us to investigate not only the ADC and Flair values but also the intensity values of the T1-Gd and CT in relation to the recurrence.

[1]The DW-MR intensities in each voxel are measured in ADC values.

3.2 Imaging

3.2.1 Medical images used in this work

MRI
The magnetic resonance imaging represents and incredible way to study matter by means of the nuclear magnetic resonance phenomenon [50]. In our case this phenomenon is used to study and "see" the human tissues with the objective to identify an abnormality in the tissue. The introduction of MR imaging in the clinic represented a very important advancement in the care of patients with brain cancer [20]. MRIs play a crucial role in many areas such as the detection of the tumor. The MRIs were used for several purposes; for detecting the tumor area, identifying healthy tissues (necessary for the radiotherapy planning), and monitoring the behavior of the tumoral mass. Some of the basic principles of MR imaging are briefly presented in this section to better understand the work of this chapter.

Magnetic resonance imaging is a type of medical imaging which rely on the nuclear magnetic resonance properties to create an image of the internal structures of the body. Powerful magnets surround gradient coils in such a way such that a strong magnetic field is created. The field aligns protons, from the Hydrogen atoms of the water (H_2O) molecule, parallel or anti-parallel to the field as illustrated in figure (3.1). The p inside a circle in the figure represents a proton, the top one refers to the proton aligned parallel to the magnetic field and the bottom proton represents the anti-parallel alignment. The protons wobble in a cyclic motion (w_0) which is often referred as precession movement. Before the magnetic field is applied, the protons wobble in non-preferential directions. After the magnetic field is applied, the net alignment is directed parallel to the field. Following, radiofrequency is applied perpendicular to the magnetic field causing both the parallel and anti-parallel protons to align parallel to the radiofrequency (represented in step 2 of figure 3.2). However, this is not a stable state and the protons have a tendency to go back to the parallel or anti-parallel positions illustrated in step 3. The time it takes for the protons to relax and go back to the parallel or anti-parallel positions is often called T1 relaxation time. At the same time some protons are being miss aligned from their parallel anti-parallel positions, a quantity related to this time is called the T2 relaxation time.

The MRI machine is able to detect or measure the T1 and T2 values and the detection location. Since different tissues and parts of the body have

different T1 and T2 values, an image can be created by using these values and the detected locations. Therefore different tissues of the body can be identified in the image.

The basic physics of MRI was briefly described above mainly to have an idea of how the MRI images are created. However, the MRI is a vast field in which many different types, commonly known as "MRI sequences", of MRI images are generated. Not all MRI requires the T1 and T2 times, but they are important values for describing many types of MRI sequences. There is a zoo of MRI images used in the medical field, some techniques additionally require a contrast agent to better highlight certain tissues of interest. The MRI sequences used in this work (for each patient) include the DW or ADC, T2-Flair, and T1-Gd, sequences. Typical resolution of the MRI images are slices of 2 mm with x and y resolutions of about 1-2 mm (often 1.15 mm). The x and y resolution is of 1.15 mm and the slides were taken each 2 mm. The patients included 5 women and 12 men. The average age of the patients was 60.5 years old with standard deviation of 13.9.

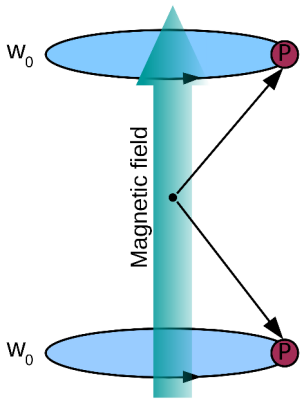


Figure 3.1: Precession movement of protons under a magnetic field. The top proton (p) is aligned parallel with the magnetic field where as the proton on the bottom part is aligned anti-parallel. The precession movement is characterized by a wobbling motion.

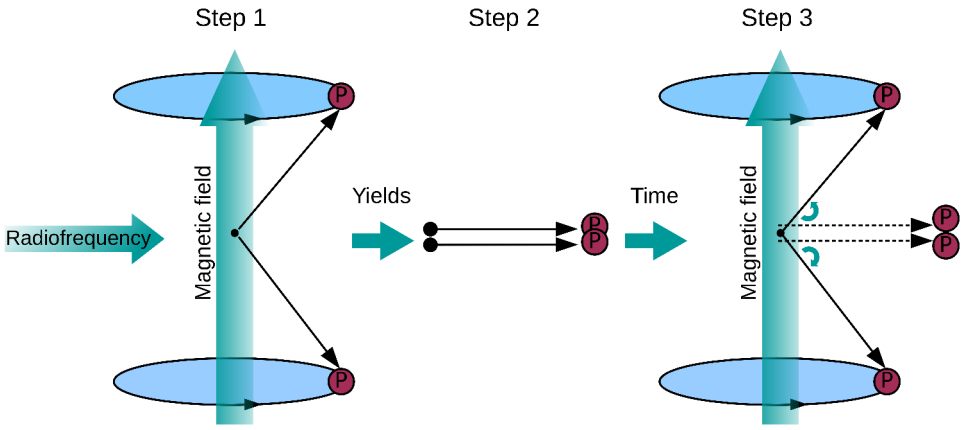


Figure 3.2: Proton alignment process in MRI. Radiofrequency is applied which forces the wobbling protons to momentarily align parallel to the direction of the radio-frequency. Then after sometime the aligned protons return to the wobbling motion. The time it takes for protons to return to align parallel or anti-parallel to the magnetic field is called T1 relaxation time. Simultaneously those newly aligned protons return to be aligned with the radio frequency; this is called T2 relaxation time. These two time values and their location of detection are used to created MRI images since different tissues have different T1 and T2 values.

ADC or DW-MRI sequence

The contrast of the diffusion weighted (DW) MRI is determined by the tiny microscopic movements of the protons of the water molecules [2]. In other words, this MRI sequence is related the diffusion of water molecules. These microscopic movements are important because they can be used for the study of the human tissues. The diffusion MRI can be used for the study of intracranial tumours [32]. The DW MRI are not only used for cancer purposes but rather a wider variety of reasons such as the study by Nakahara et al., in which they studied the severity of brain injuries [46]. Nonetheless, one main use of the DW-MRI is for detection of brain cancer since the water in different tissues have different diffusion coefficients. For instance, the diffusion coefficient of water in tissues is about two to three times less than free water [35]. The ADC value for free water at 37 degrees Celsius is about $3.0 \times 10^{-9} m^2/s$ [36]. An explanation of the effect is that physical items such as cell membranes, fibers etc impede the diffusion effect. The idea is that tumor areas have different compositions compared to normal tissues hence their diffusion coefficients differ, resulting in different contrast in the DW-MRI image.

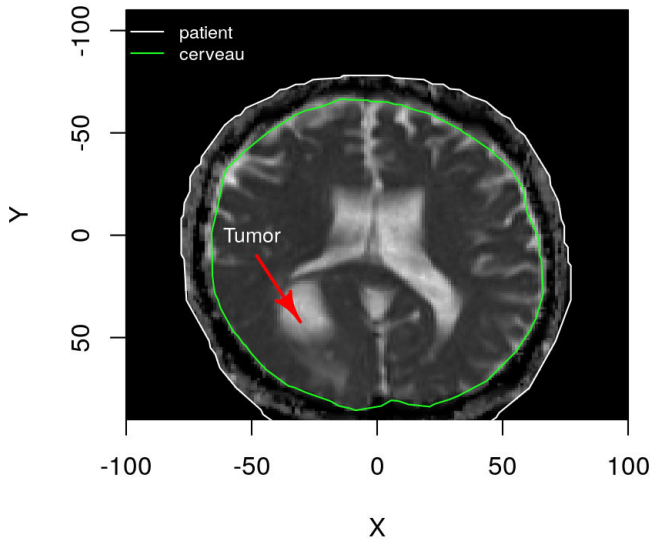


Figure 3.3: Diffusion Weighted Magnetic Resonance Imaging (DW-MRI) sequence of the brain of a patient suffering from Glioblastoma. The red arrow indicates the tumor location which displays higher ADC pixel values.

The DW-MRI helps medical doctors diagnose Glioblastoma by identifying the abnormal tissue. For instance, in figure 3.3 a DW-MRI sequence of the head of a patient suffering from Glioblastoma is shown. The image shows the delineation of the contours of the patient in white and the brain (french word *cerveau*) in green. The tumor location is distinguishable by the brighter area in the image which represents higher diffusion coefficient intensity values.

T2-Flair MRI sequence

The etymology of the word *tumor* comes from the latin word *tumere* which means "to swell". The T2-Flair MRI is a type of sequence that allows for the identification of the swelling section. Identifying the edema is important especially for diagnostics since it depicts anatomical details. This MRI sequence is very useful since it assists in the identification of the swelling versus the cerebrospinal fluid [4]. The acronym FLAIR stands for FLuid Attenuated Inversion Recovery [19]. Identifying the swelling is important because there is a link between cancer and inflammation [24].

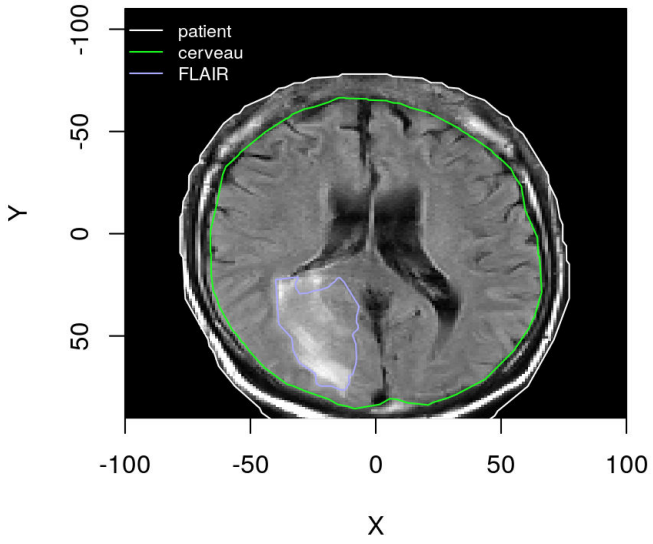


Figure 3.4: T2-Flair MRI sequence of the brain of a patient suffering from Glioblastoma. This type of MRI sequence is used to identify the swelling which can be observed inside the "FLAIR" contour.

In medical terminology the inflammation is referred as an edema. A considerable edema is perceived in the T2-Flair MRI sequence displayed in figure 3.4. The medical practitioners were responsible for all the contouring of the anatomical details such as the edema shown in the image in figure 3.4. From the mathematical modeling perspective, the edema represents different pixel intensity values compared to the healthy tissue. The difference in values is useful because it allows for the discrimination of the lesion or damaged area.

T1-Gd MRI sequence

After excising as much tumor mass as possible by a surgical procedure, the radiotherapy planning can begin. However, also a T1-Gd MRI sequence was performed since this sequence helps in the identification of organs at risk and the tumor area which are necessary for the contouring in the radiotherapy planing. Typically the T2-Flair sequence is used for identifying the edema where as the T1-Gd is used for identifying healthy tissues as wells as the border of the tumor region [22]. Hence, the T1-Gd sequence is used for tumor contouring.

The post-surgical T1-Gd MRI image is shown in figure 3.5 in which we can see that there is still tumor mass left after the excision. The patient head, the brain, and the Gross Tumor Volume (GTV) are contoured in the

image. The borders of the GTV, the tumor mass, are quite bright making it easier to identify it. It is worth clarifying that the T1-Gd MRI sequence refers to the T1-weighted sequence with contrast agent Gadolinium (Gd). The contrast agent is typically injected intravenously to the patient with the objective of enhancing the contrast of the image. The borders of the tumor are highlighted since the contrast agent accumulates in that region [4].

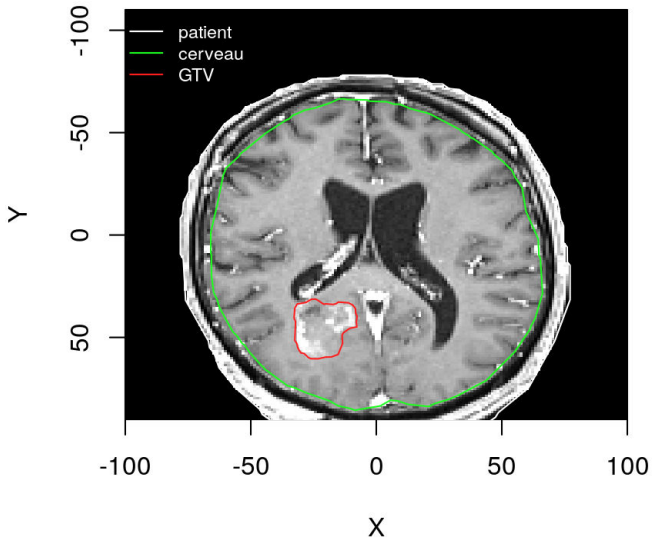


Figure 3.5: T1-Gd MRI sequence, pre-radiotherapy, of the brain of a patient suffering from Glioblastoma. The GTV, the brain (french word "cerveau"), and the head of the patient are contoured. The borders of the GTV are brighter because the contrast agent Gadolinium accumulates in that region.

The water, the fat, and blood flow in the form of blood vessels can be identified from the image. The water is dark and it is shown in the middle of the image in the form of an x shape. As a useful rule of thumb for the T1-Gd MRI sequence the water is darker, the blood vessel is bright and the fat is bright too.

In many cases, when possible, another T1-Gd MRI was performed but after the treatment for following up the care of the patient. A post main treatment T1-Gd MRI sequence is shown in figure 3.6 in which unfortunately a tumor recurrence was observed. The recurrence area is delineated in dark red (crimson) color.

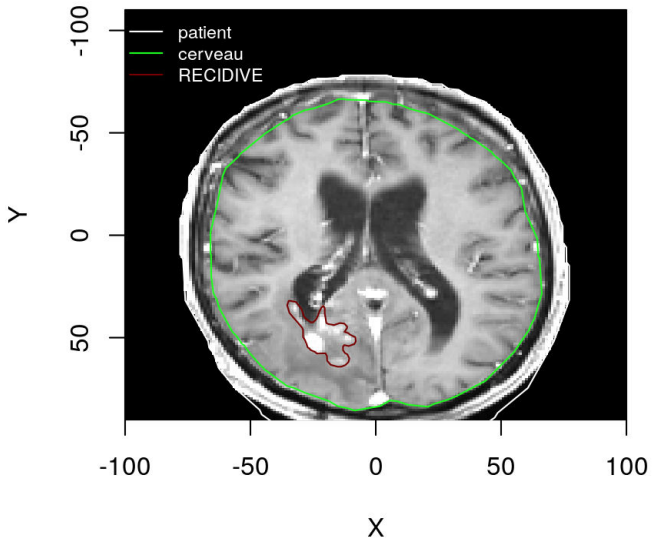


Figure 3.6: T1-Gd MRI sequence, after the main treatment, of the brain of a patient suffering from Glioblastoma. The image was used for follow-up purposes. Unfortunately, the image reveals a tumor recurrence (french word "RECIDIVE") which is contoured.

CT imaging

The computer tomography imaging is a technique used to reconstruct an organ or object in a 3 dimensional manner by means of directing x-ray beams around the patient (e.g., head) and measuring the attenuation of the beams. The attenuation of the x-rays are used to reconstruct an image and many slices are developed for the 3D reconstruction. An algorithm is used by the computer to correlate the attenuation of the beams, depending on the incoming direction, with the density of the tissue. The denser the tissue the more the beam is attenuated. Hence the beam that passes through bone is attenuated much more than a beam passing through soft tissue.

The Hounsfield units are used to refer to the pixel value in computer tomography. The CT scanner needs to be calibrated for the relation of Hounsfield units to tissue density [43]. For the calibration, an arbitrary value of -1000 is set for a beam passing through air and zero is set when the beam passes through water. One main important feature of the CT scanner is that it allows for low contrast discrimination meaning certain tissues can be better differentiated compared to a 2D x-ray image.

The CT started to answer the necessity for tumor and normal tissue

delineation and localization in such a way that can be used for the planning and delivery of therapy [18]. Perhaps the most important part of the CT imaging is its relation to the energy deposited; the radiation absorbed dose planning can be done using the material density inferred by the use of the CT value expressed in Hounsfield Units.

Example of a CT image

A CT image acquired from our database is shown in figure 3.7. Three contours are overprinted on the image: the tumor (GTV), the head of the patient, and the brain. The tumor area is darker than other areas in the CT image. However, the outer contours of the tumor are not as well defined as in the T1-Gd MRI depicted in figure 3.5.

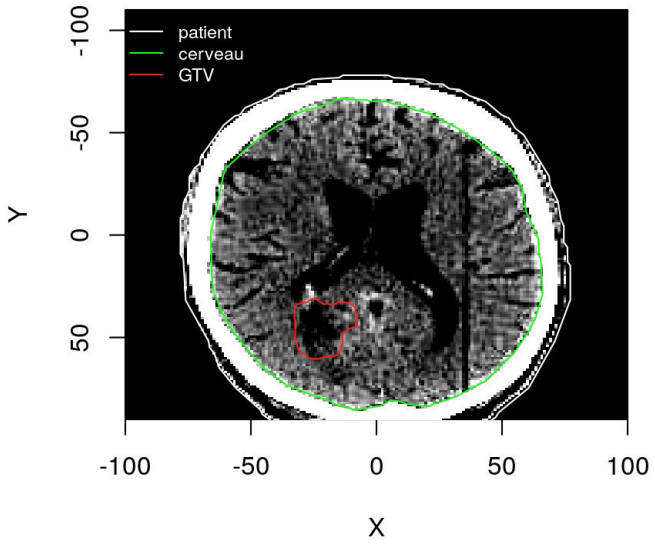


Figure 3.7: Computer Tomography (CT) image of the brain of a patient suffering from Glioblastoma. The GTV, the brain, and the head of the patient are contoured. Bone structures such as the skull, area between the head of the patient and the brain, are very clearly identified in CT images.

The attenuation is related to the electron density of the material as well as the beam energy. The reason certain soft tissue structures are not well differentiated can be explained by the close electronic density of the tissues. The higher the electron density the more the x-rays will interact with the material hence more x-rays are attenuated. For instance, the skull is well defined by the bright area, between the patient and the brain, because it has a higher electron density than soft tissue. Lastly, the CT image was used

primarily for the planning of the radiotherapy treatment.

3.2.2 Successive layers outwardly from the GTV surface

In order to study the recurrence, which happens in a progressive manner, we decided to investigate the contents of the medical images by creating successive layers outwardly from the GTV surface. Three of these outer layers are illustrated in this section [2]. First a tumor contour was selected then it was used to do a three dimensional reconstruction of the tumor as shown in figure 3.8. The units for all the three axes are in millimeters. In many cases the tumor is somewhat spherical just like the one shown here.

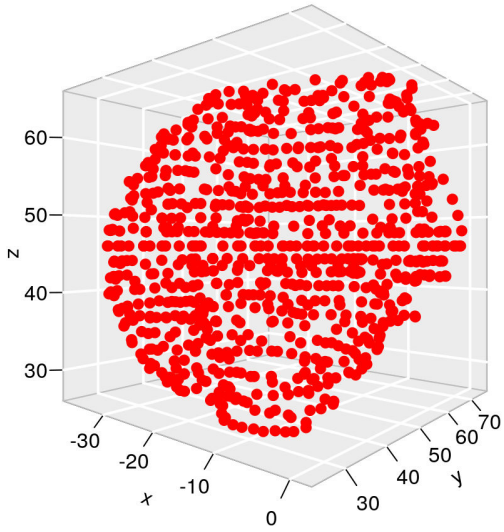


Figure 3.8: Tumor 3D representation (in mm) created by using the contour information. The tumor shows a somewhat spherical shape.

The contour of slice z=30 of the 3D image is plotted in a red curve in figure 3.9, the axes are in millimeter units as in the 3D image. Additionally three layers are represented as well. The first GTV contour expansion or layer

[2]Five layers were created each of 2 mm in thickness: 0-2, 2-4, 4-6, 6-8, 8-1.0 mm. After performing multiple modeling trials we discovered that the first couple of layers were more relevant to the models because the closer the layers are to the GTV the more recurrence area is involved. This is because the recurrence often grows immediately around the GTV.

is represented by golden brown circles ◯, the second layer is represented by dark salmon squares □ and lastly, the third layer is represented by dark cyan diamond shapes ◊. The outward expansions were done in a three dimensional manner which explains why certain widths of the layers are much wider than other sections as observed in figure 3.9.

Voxel intensity values of the CT-scan, DW-MRI, T1-Gd MRI, and T2-Flair MRI were associated to the corresponding positions represented by the small shapes in the figure 3.9. With this information we can calculate how the voxel intensity values change outwardly. Since the recurrence area is also known then we can associate the location of the recurrence and determine which voxels of the outer layers belong to the recurrence location. With this set-up we could analyze if there is a difference or not in voxel intensity values in the corresponding recurrence area in the layers versus the healthy tissue in the same layer.

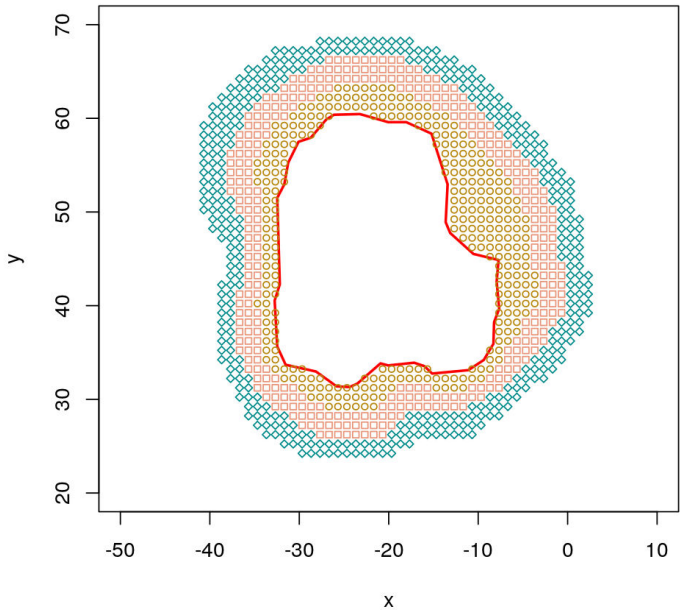


Figure 3.9: Tumor contour (red) with three outer layers. The first layer is represented by golden brown circles, the second by dark salmon squares, and the third by cyan diamond shapes.

3.2.3 Normalization of images

We developed several computer models for the purpose of predicting the recurrence location which require the normalization of images; the normalization methodology is presented below . The approach taken was to start by comparing the healthy pixel values on the mirror image of the tumor then to expand the GTV contour outwardly in layers of 2 millimeters each. By creating the layers we can discriminate the voxels corresponding to a location where there was a recurrence or not. In figure 3.10 the contours of the mirror (green), the GTV or tumor (red), and the recurrence (navy blue) section are overprinted onto a T1-Gd MRI (this sequence was used for the contouring process) pre-treatment image. If we expand the tumor contour (red) we expect that there may be sections where the recurrence overlaps the expanded layers. The mirror image of the tumor was acquired by an algorithm, that we developed, using the tumor contour and the line that divides the brain in two sections. The division line was defined manually using the CT-MRI images.

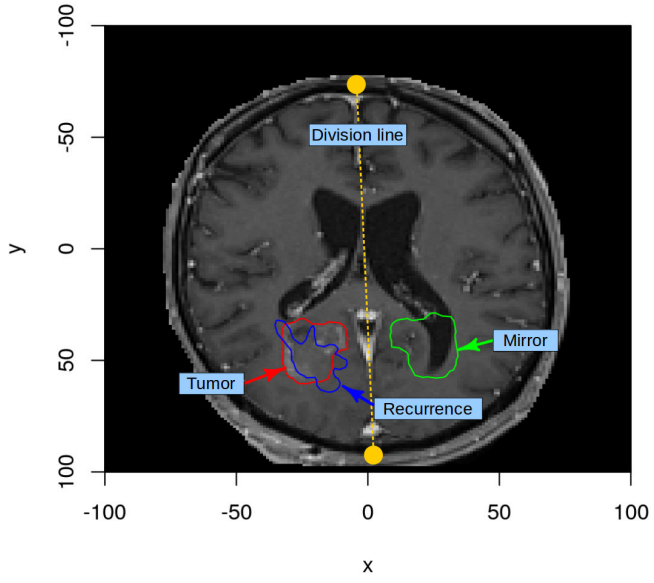


Figure 3.10: T1-Gd Magnetic Resonance Image with overprinted contours of the tumor, recurrence, and mirror. The division line was used to separate the brain in two parts in order for an algorithm to calculate the mirror contour.

The mirror image serves as a reference for healthy tissue analysis versus

the GTV and also for scaling or normalizing the pixel values of the GTV related contours. Additionally, it is worth noting that the recurrence contours were used to identify if a voxel position correspond to a location where the recurrence appears. This information is crucial for developing the recurrence prediction models presented in this chapter.

Pdfs of tumor and mirror sections

The region corresponding to the GTV was identified and the pdfs of intensity values were created for the CT-scan, the DW-MRI, the T2-Flair MRI and for the T1-Gd MRI; they are plotted in figure 3.11. About 20 pdfs are plotted in each of the four graphs and each pdf represents the data of a single patient. For instance, in the CT plot there are 20 pdfs in which we can observe that the form of the pdf seems to be conserved, sort of a Gaussian shape, but in some cases displaced. For the remaining three plots, the DW, FLAIR and T1Gd, the intensity values seem to be scattered and do not show a predominant pdf shape.

Similar pdf graphs concerning the mirror section were created and they are displayed in figure 3.12. The mirror section refers to a mirror image of the tumor which theoretically is a healthy region of the brain and somewhat symetric to the tumor. The idea is first to compare the tumor and the mirror region to observe a difference in intensities. One main difference between them is that the pdf of the DW-MRI has a much more defined shape for the mirror region than for the tumor region. The DW pdf shows higher ADC values in units of $mm^2/seconds$. That is, the ADC value called the Apparent Diffusion Coefficient (ADC) is higher for the GTV. Additionally the pdf for the T1-Gd corresponding to the tumor shows a more spread out shape ranging from 100 to 600 where as the mirror region shows a slightly less spread out pdf ranging primarily from 200 to 500.

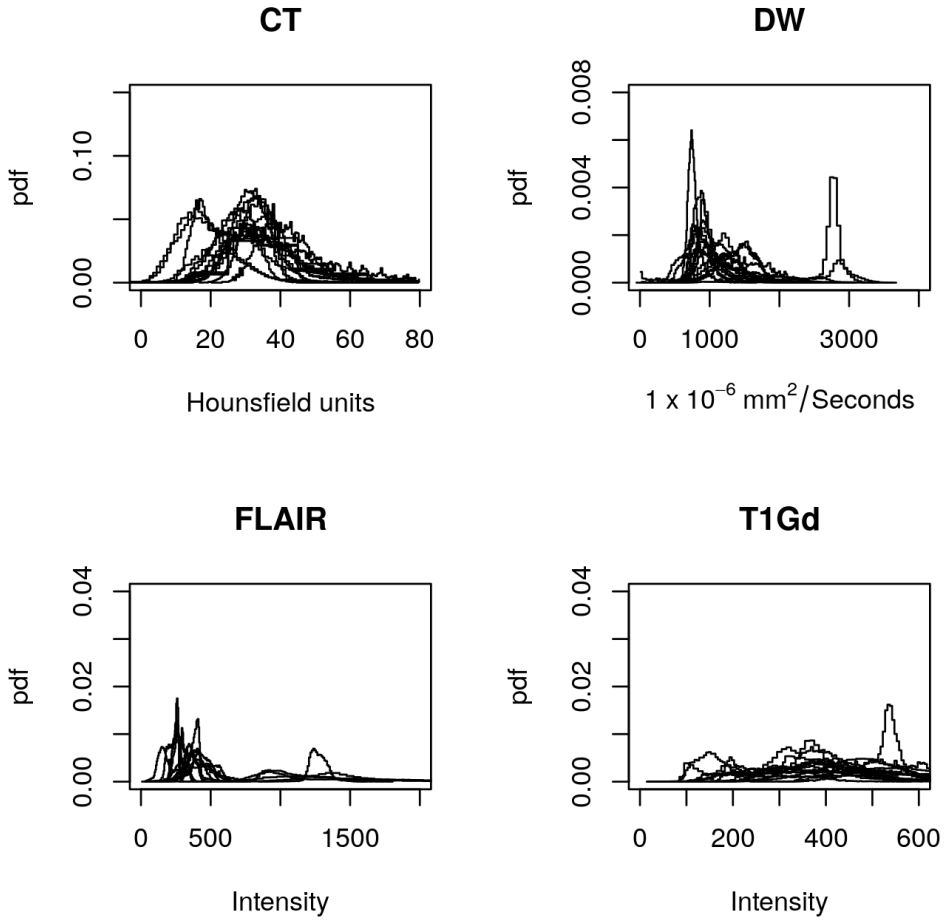


Figure 3.11: Pdfs of the intensity values of the GTV concerning about 20 patients; each individual pdf corresponds to a patient. The pdfs do not show very specific patterns.

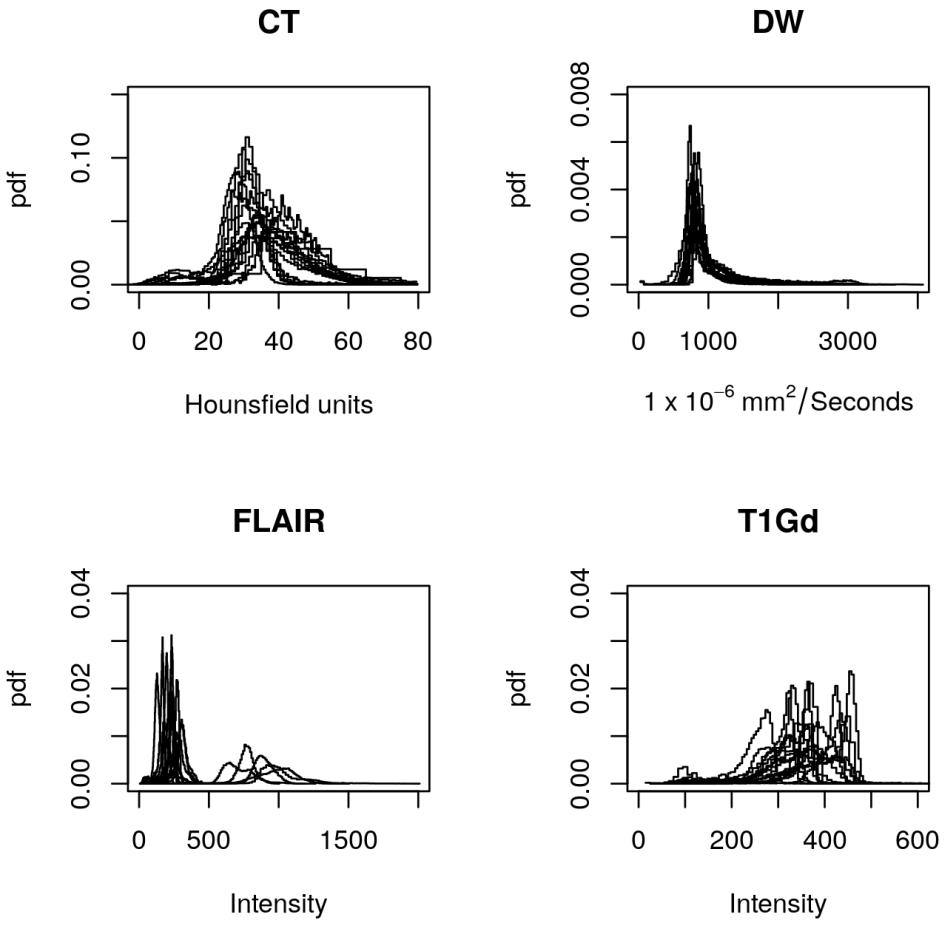


Figure 3.12: Pdfs of the intensity values of the mirror area. The pdfs display more organized values compared to the tumor pdfs displayed in figure 3.11.

Normalized pdfs of tumor and mirror sections

The mirror image was not only used as a reference to compare equivalent tissue (healthy tissue) to the GTV but was also used for normalization or calibration purposes. Due to a variety of reasons the pixel values of the voxels vary from patient to patient which perils the tumor and mirror pdf comparison. One reason of the pixel intensity variations is due to the fact that patients acquired their medical images at different imaging centers and those machines are not necessarily calibrated the same way. Second, biological differences such as age or other specific characteristics may influence the intensity discrepancies in healthy tissue since the pdfs are not normally

distributed and sometimes present important extensions (especially for the DW and T1Gd). The approach we took was to normalize all the mirror images by setting the median to zero value and dividing the values by the difference between the third and first quartile. The shape of the pdfs after the normalization procedure is better defined as can be seen in figure 3.13. These pdfs in figure 3.13 are the same as in figure 3.12 but normalized.

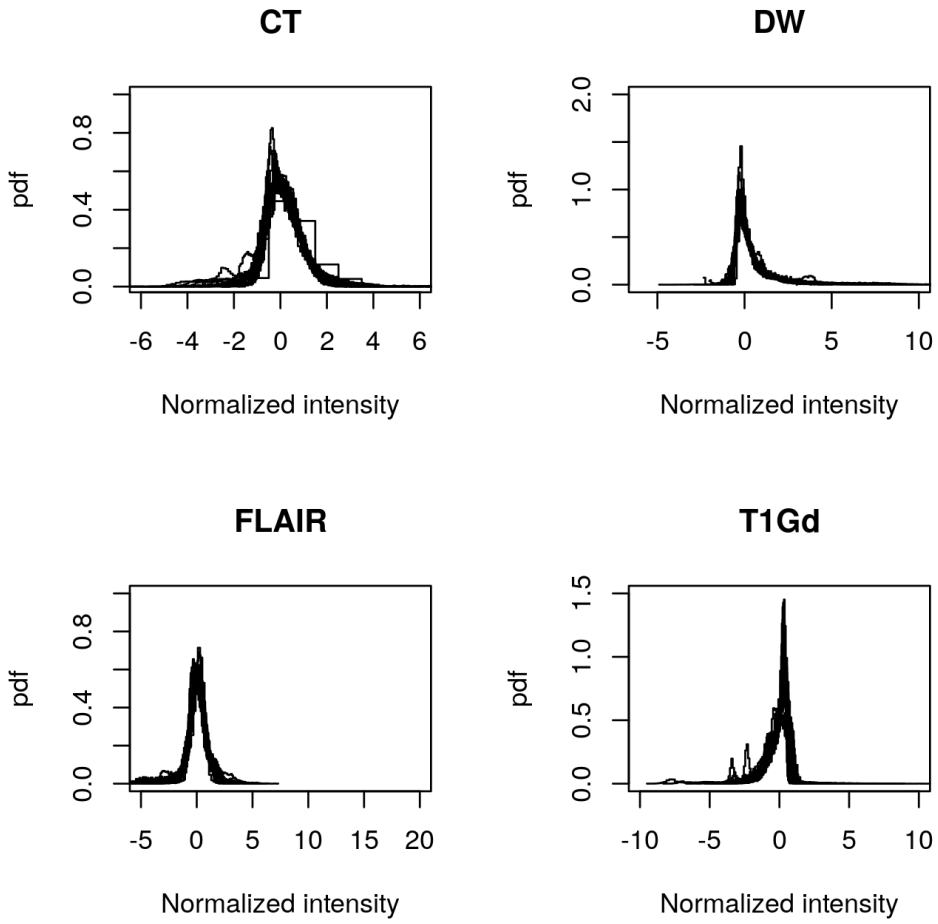


Figure 3.13: Pdfs of the normalized intensity values of the mirror area. The pdfs exhibit well organized distribution of values with consistent shapes.

Simultaneously, the pdfs corresponding to the tumor region were normalized as well under a similar procedure. The mirror image of patient one was used to calibrate the tumor image for patient one, and mirror image of patient

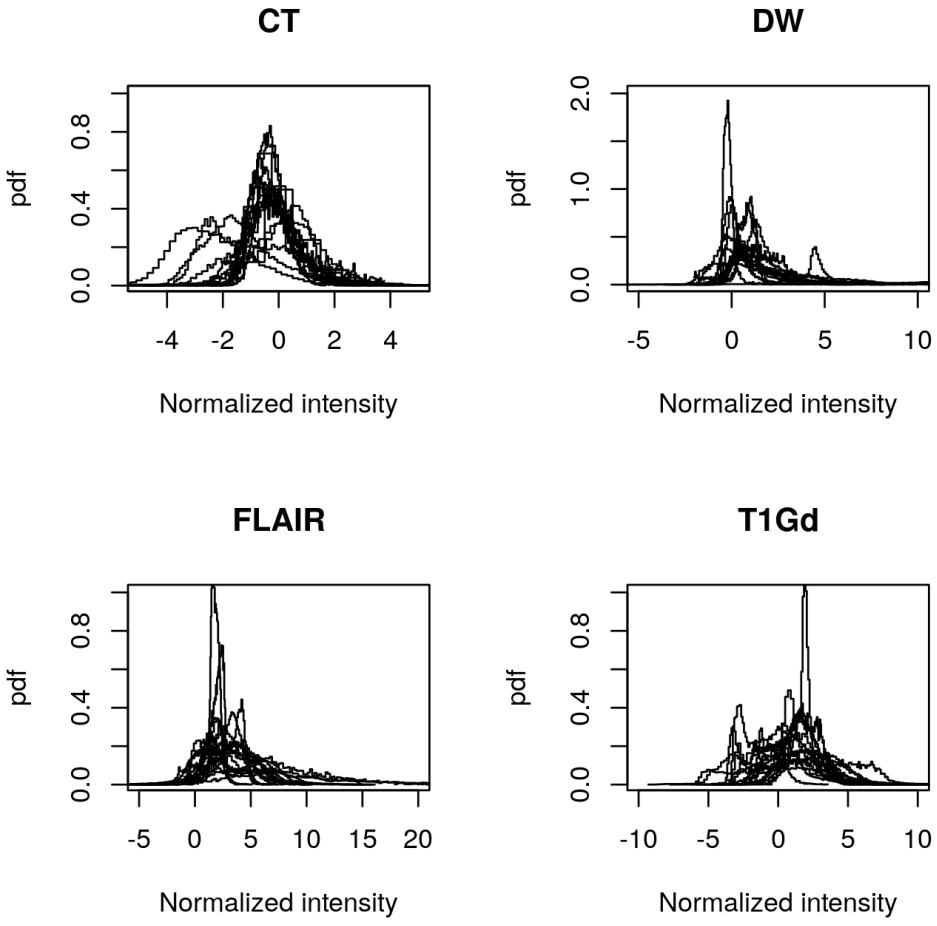


Figure 3.14: Pdfs of the normalized intensity values of the GTV present considerable less consistent shapes compared to the mirror pdfs but much more consistent shapes than the original not normalized tumor pdfs. A change of intensity values are observed for all the images compared to the mirror images.

two calibrates tumor image two and so on. The normalization corresponding to the tumor region leads to the pdfs in figure **??**. The normalized pdfs are easier to compared to the mirror image. The FLAIR pdf for the tumor shows an increase in normalized intensity values. The pdf corresponding to the DW shows a wider range of values in the GTV than in the healthy tissue. The pdf for the CT remains with a similar width but with less organized values, meaning there are more uneven values than in the mirror pdf.

Pdfs of the outer layers
The voxels inside the layers can either be in a location where the tumor regrows or in a region where the tumor does not regrows. The regrowth location is known and we are assuming that the organ and tissue locations in the brain remain static after the radiation therapy.

The pdfs concerning the normalized intensity values of the layers corresponding to locations where the recurrence appeared are displayed in figure 3.15. The pdfs concerning the non-recurrence coordinates of the layers are displayed in figure 3.16. The pdf for the CT displays a higher concentration of normalized intensity values near the mean for the recurrence region of the layers, this can be observed by noticing the higher peak nearing 1.5 in the pdf versus almost 1 for the non-recurrence CT pdf. The pdf corresponding to the DW image for the recurrence also displays a higher peak. Overall, the images display higher peaks in the pdfs involving the recurrence. However, the more drastic peak is observed when comparing the T1-Gd pdf; for the non-recurrence pdf the peak reaches almost 0.9 whereas for the recurrence the peak reaches approximately 2.0. Lastly, intensities of the medical images for the voxels corresponding to the layer (0-2 mm) are plotted and additionally the color of the dots represent whether a voxel correspond to a recurrence or non-recurrence location. This creates the segmentation plots displayed in figure 3.17. Unfortunately, a strong definite segmentation between the two populations (recurrence vs non-recurrence) is difficult to observe.

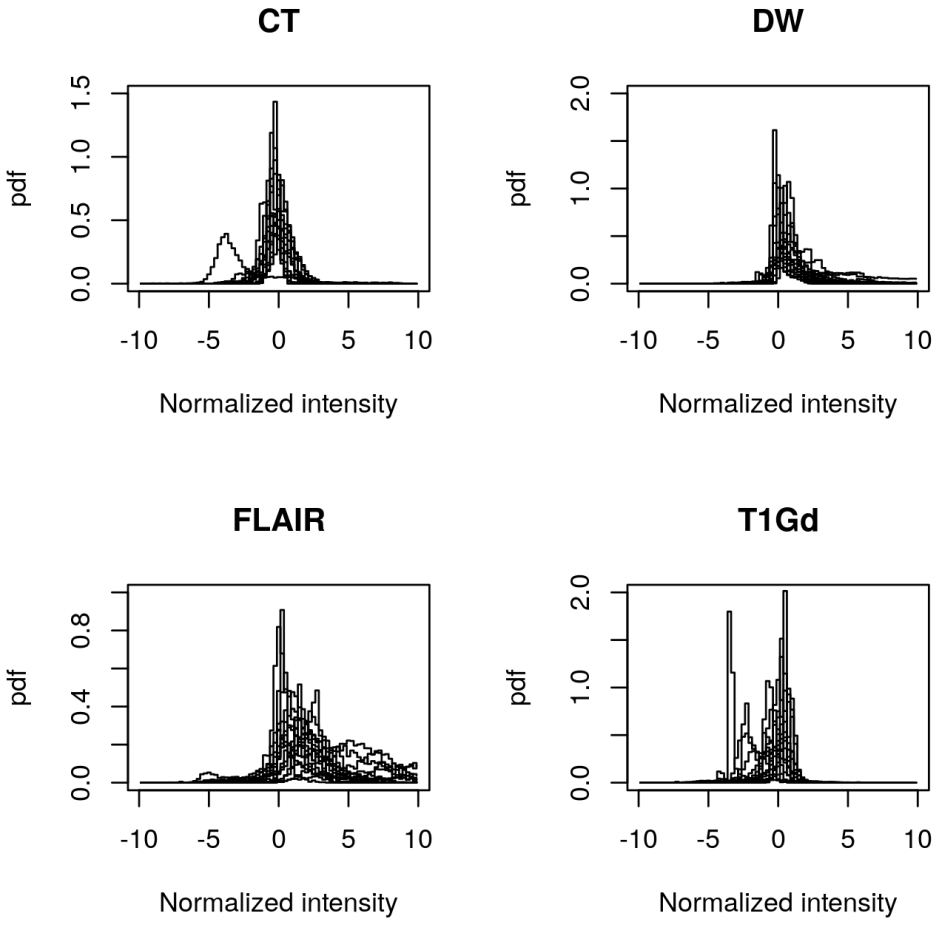


Figure 3.15: Pdfs of the normalized intensity values corresponding to the recurrence regions. The plots display higher peaks compared to the non-recurrence plots in figure 3.16.

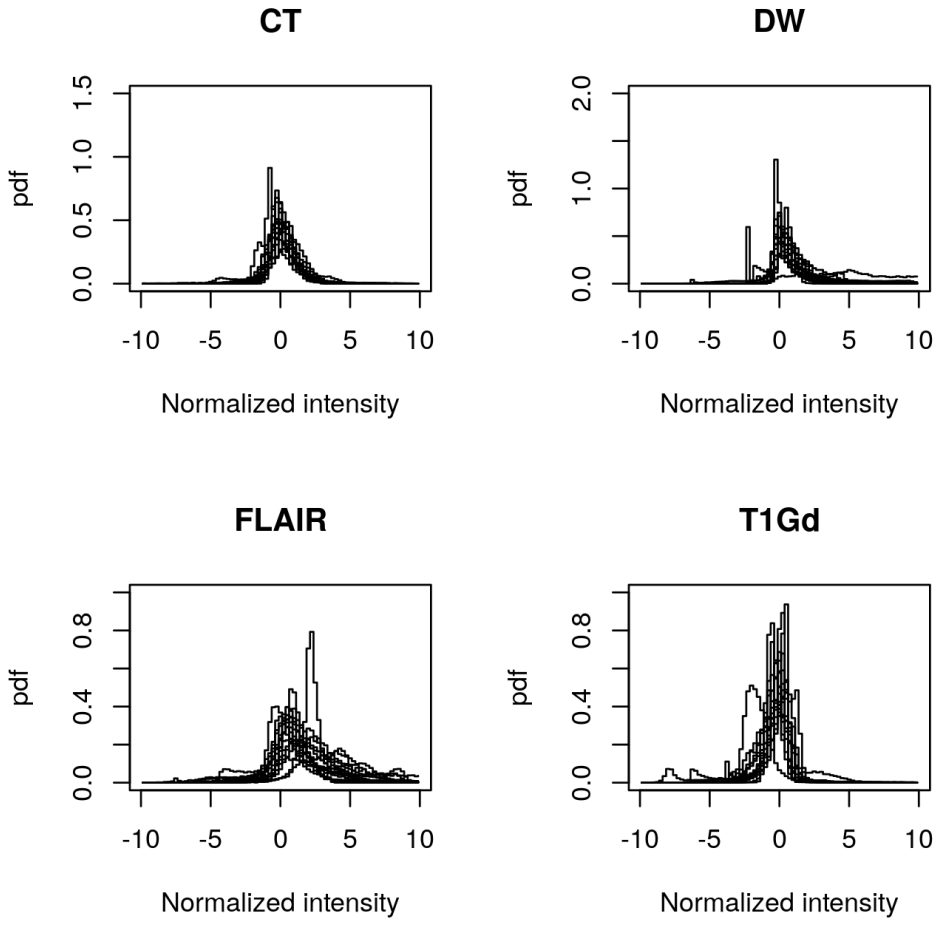


Figure 3.16: Pdfs of the normalized intensity values of corresponding to non-recurrence regions.

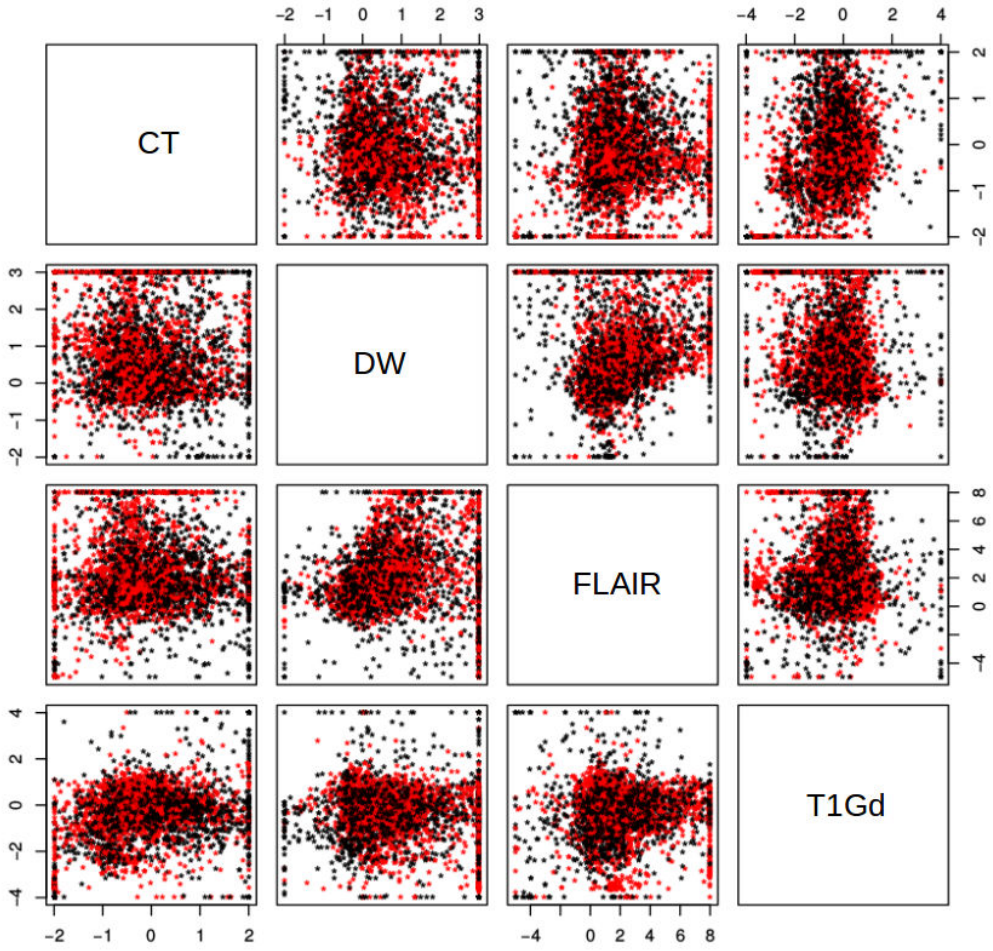


Figure 3.17: Segmentation plots of the medical images. The red points correspond to the recurrence and the black ones to the non-recurrence area. It is difficult to see a segmentation between the two populations.

3.2.4 Machine Learning

Multiple models, which involve Machine Learning techniques, were used with an overall goal of segmenting the recurrence and non-recurrence voxels (see figure 3.17) according to the medical image intensities.

Generalized linear models (GLMs)
The GLM models were developed using the R software glm() function. The GLM models are generated by specifying the variables, the family of response

function, as well as providing the data to be fit. The family item refers in this case to a binary response. The mathematical expression for the equation is given by,

$$link_{function}(Response(X)) = AX_1 + BX_2 + CX_3...etc \tag{3.1}$$

in which the X refers to the variables and the A, B, C are the equation coefficients which are estimated by the optimization algorithm in the glm() R function. The response comes from the data either in which a voxel, from the layer, is inside or outside the recurrence. Notice that the right hand side is linear with respect to the link function but as a whole the *Response(X)* is not linear. The pre-determined link function for the binomial response is the logit function

$$logit(p) = log\Big(\frac{p}{1-p}\Big) \tag{3.2}$$

which is the inverse of the logistic, commonly known as the sigmoid function, is often used for binary data fitting. The binary fitting is also called the logistic regression. Therefore the glm equation is,

$$logit(p(X_1, X_2, X_3)) = A + BX_1 + CX_2 + DX_3...etc \tag{3.3}$$

Since the logit is the inverse of the logistic or sigmoid function, then p is,

$$p(X_1, X_2, X_3) = sigmoid(A + BX_1 + CX_2 + DX_3...etc) \tag{3.4}$$

Full model

For our case, the glm equation for the full model is,

$$\begin{aligned} logit(p(\xi)) =& \{A + B(CT) + C(DW) + D(FLAIR) + E(T1Gd) \\ &+ F(CT \times DW) + G(CT \times FLAIR) \\ &+ H(DW \times FLAIR) + I(CT \times T1Gd) \\ &+ J(DW \times T1Gd) + K(FLAIR \times T1Gd) \\ &+ L(CT \times DW \times FLAIR) \\ &+ M(CT \times DW \times T1Gd) \\ &+ N(CT \times FLAIR \times T1Gd) \\ &+ O(DW \times FLAIR \times T1Gd) \\ &+ P(CT \times DW \times FLAIR \times TIGd)\} \end{aligned} \tag{3.5}$$

Reduced model

and the reduced glm model is,

$$logit(p(\xi)) = A + B(CT) + C(DW) + D(FLAIR) + E(T1Gd) \tag{3.6}$$

Where $\xi = [CT, DW, FLAIR, T1Gd]$. The constants A..P are the equation coefficients and they are the unknown parameters of our models. The A represents the intercept coefficient and for both models it was set to a value of zero. For some models the A coefficient was zero meaning the intercept is at zero. The variables CT, DW, FLAIR, T1Gd represent a voxel value of the medical image at a certain position (x,y,z).

One can note that the GLM model of a binomial response can be interpreted under the Bayesian Framework where the prior would be uniform for each parameter. An example of the work performed by the GLM model can be seen in figure 3.18.

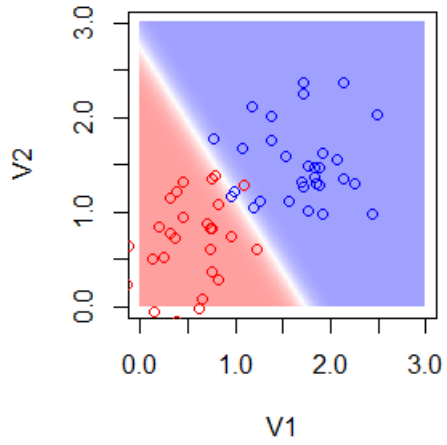


Figure 3.18: Illustration of the work performed by a binary response GLM model involving hypothetical variable V1 and V2. The GLM is capable of predicting for well differentiated data; the lower panel belongs to one type of response and the upper panel to another type of response. The GLM prediction is depicted by a solid line.

Decision tree models

Decision tree models were also used. A decision tree is a way of representing a decision strategy to be able to determine a class [53]. The author Ross mentions the importance that decision trees play in classification of objects [47]. The decision trees are part of the supervised machine learning methodologies.

We can use the analogy of an actual living tree in which the tree starts at the root then splits into main branches then subbranches and finally after the last branch we can find a leaf. As an analogy, the nodes where a branch

separates refers to a choice or the alternative(s), the leaf refers to the outcome. In mathematics the outcome can often be the solution of an equation which we are trying to solve.

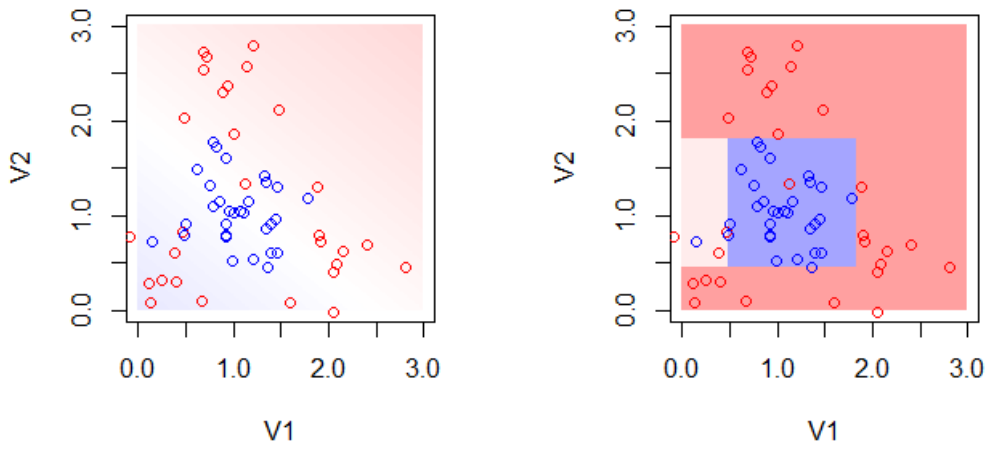
There are specific programming codes available in several programming software and languages for the decision tree modeling. The packages "tree" and "rpart" from the R-software were used for the decision tree modeling. An example of the three model equation used is written below as,

Tree model

$$Model_3(CT, DW, FLAIR, T1Gd) = tree\{response \sim CT + DW + FLAIR + T1Gd\} \tag{3.7}$$

The *tree* is a predefined function from the previous R packages mentioned. This function creates the decision tree model which uses the voxel intensity values (CT-scan plus the MRI images mentioned in the model) as input parameters. The response is binary; of the voxel being either inside or outside the recurrence area but the response of the equation is a probability.

Decision trees are more agile than GLM models in the sense that they are able to separate non linearly separable variables (figure 3.19). It is worth noting that they are prone to overfitting.



(a) GLM **(b)** Decision tree

Figure 3.19: Illustration of the agility of decision trees. The same points are drawn for both figures. (a) The GLM is not capable of making a correct prediction for such data. (b) The decision trees are capable of separating non linearly separable variables as can be seen by the square region.

Bootstrapping and random forest methods

Challenges come across when trying to develop a model robust enough, that

includes multiple parameters, is that the dataset can be drastically reduced. For instance, out of the 92 patients only around 20 had the desired characteristics for developing the recurrence modeling. The requirements includes having the CT-scan, DW-MRI, T1-Gd MRI (before and after treatment), and T2-Flair MRI. As well as having consistent image slices of about 2 mm in the z direction, also the data should include the recurrence contoured among other desired characteristics. The reduction of the number of patients encouraged the need to determine the uncertainty of the estimator parameter based on a reduced number of subjects. The answer to this issue was to implement bootstrapping methods. Bootstrapping allows to estimate the confidence interval of a parameter of interest. For instance, if we calculate the mean of a sample then the Bootstrapping method helps us determine the pdf of the mean parameter.

Bootstrapping

The idea is to keep sampling from the same small dataset while allowing replacement to determine the confidence interval of the parameter of interest.

Sampling with replacement refers to allow to sample the same subject in a new sample. Detail explanation can be found in the literature such as in the following books by Chernick [11] and Kotz [30].

For the recurrence GLM modeling we performed about 100 bootstraps, each resulting in an independent prediction. Then a global prediction was obtained using the individual predictions. In a similar way about 100 bootstrapping iterations were performed for the decision tree modeling. The solution to the GLM are coefficients of an equation, however the solution to the tree models are analogous to the leaves of a tree in which different paths were taken; therefore we can not average the leaves. Notice that for a GLM model we can easily average the coefficients of the 100 models which resulted from the bootstraps for each GLM model. A method known as random forest was used instead of a single tree. The random forest consists in generating a large ensemble of trees and voting for the most popular classification [7]. More information about the random forest algorithm can be found in the following reference [38] which includes applications in the R software. Overall, the random forest allowed us to aggregate the predictions of the 100 bootstrappings into a single prediction. In figure 3.20, an example of the ensemble of predictions created by the random forest is depicted by the blue squares.

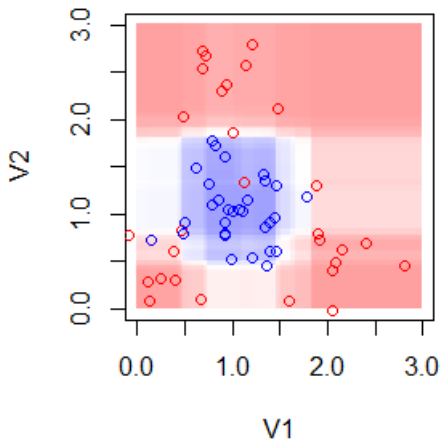


Figure 3.20: Random forest illustration. The ensemble of predictions are depicted by the square regions, and the blurriness of the squares represents the uncertainty which is inherent to the nature of the random forest.

3.2.5 Evaluation of models using Receiver Operating Characteristic (ROC) spaces

The Receiver Operating Characteristic (ROC) spaces were constructed for the purpose of evaluating the accuracy of the models. A ROC space is constructed by plotting the True Positive Rate (TPR) versus the False Positive Rate (FPR). The TPR concerns the number of times the model acurately predicts the locations where the tumor recurred while the FPR concerns the number of times the model accurately predicts a non-recurrence location. The mathematical equation to construct the TPR is,

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{3.8}$$

and the FPR is given by,

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{3.9}$$

where:

- **True Positive (TP):** Number of points where location of recurrence was correctly identify.
- **False Positive (FP):** Number of points where location of recurrence was incorrectly (or falsely) identify.

- **True Negative (TN):** Number of points where location of non-recurrence was correctly identify.
- **False Negative (FN):** Number of points where location of non-recurrence was incorrectly identify.
- **Positive (P):** Actual number of recurrence points.
- **Negative (N):** Actual number of non-recurrence points.

An ideal model would predict correctly most of the time meaning it would have a TPR approaching 1, and it would rarely predict a recurrence location where in reality it is a non-recurrence location. This would be equivalent as saying FPR should be low. The top left corner of a ROC space has these two conditions, high TPR and low FPR, hence an ideal model would predict in the top left corner of a ROC space plot.

3.3 Recurrence predictions

3.3.1 Recurrence modeling conditions

Multiple models were developed with the aim of improving the recurrence location prediction. For illustration purposes several conditions modelled are presented; they are specified in the illustration in figure 3.21. The 0-2 and 2-4 mm layers were chosen because they seem to make better predictions than the rest of the layers[3]. Lastly, in order to limit the imbalance between "recurrence" and "not recurrence" data, the same amount of recurrence and not recurrence data was randomly selected for each patient.

[3]Refer to section 3.2.2 for more information about the layers.

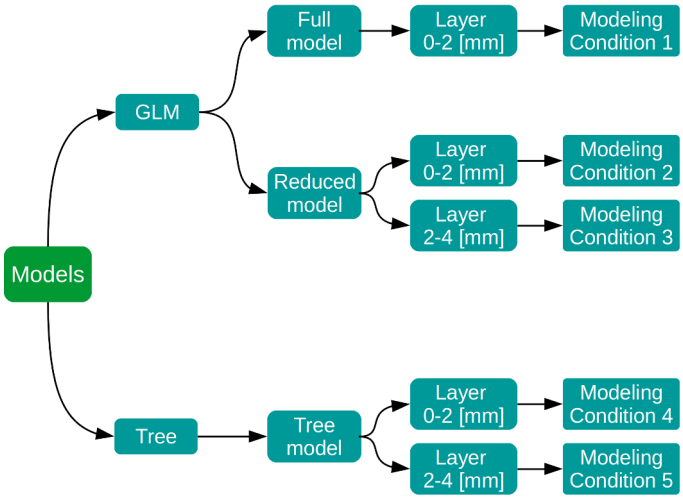


Figure 3.21: Examples of several conditions modelled. First the type of model is chosen either a GLM or a tree model then the layer to be used. The GLM model is composed of the full model and the reduced model.

3.3.2 Coefficients of the GLM models

In this section the coefficients of conditions, 1-3, involving the GLM models are presented. The first condition correspond to the full model and the last two conditions involve the reduced model. The average coefficients of equation 3.5 for condition 1 are written in table 3.1 and their values are identified with a blue vertical line in the coefficient histograms. The corresponding coefficients refer to the coefficients associated to a certain variable. For instance, the corresponding coefficient FLAIR for condition 1 refers to constant D in equation 3.5. Due to the bootstrapping techniques many coefficients were calculated. The histograms of the coefficients corresponding to condition 1, are shown in figure 3.22. A red vertical line is also drawn in the histograms to indicate the zero value with the intention to better visualize the distance to the average blue line.

Note

All the histograms in this section correspond to the ensemble of 100 iterations of the bootstrapping and the average blue line correspond to mean.

Overall, the further the blue line is from the red line the more weight the coefficient has on the equation. The purpose is to determine which variables are the most relevant in the recurrence prediction equation. The more complex variables composed of the FLAIR×T1Gd and the DW×FLAIR seem to have a lesser weight as can be observed at how the mean value is closer to zero than the rest of the coefficients as can be seen in figure 3.22.

From the histograms of the full model we observed that globally speaking the interaction between images play a lesser role, hence we developed the reduced model with the most relevant single coefficients (CT, DW, FLAIR, T1Gd). The histograms corresponding to the reduced model, condition 2 and 3, are displayed in figure 3.23 and 3.24 and they closely resembled each other. The corresponding coefficient for the CT, DW, and T1Gd appear to have a larger absolute value for condition 2 than for condition 3. For instance, the coefficient corresponding to the DW is -0.224 versus -0.161 and for the T1Gd is 0382 versus 0.286 (written in table 3.1). This means that the layer from 0 to 2 mm is the most relevant layer. In other words, the prediction power decreases with layer size suggests that pre-treatment images are more relevant for the beginning of the recurrence process, which is not really surprising. The higher FLAIR values indicate the swelling area where as the low CT values indicates where the surgery was made. Lastly, the low DW values agree with the initial findings of a previous medical thesis work [44].

Corresponding coefficient	Condition		
	1	2	3
Intercept	-0.277		
CT	-0.458	-0.544	-0.483
DW	-0.183	-0.224	-0.161
FLAIR	0.169	0.102	0.106
T1Gd	0.422	0.382	0.286
CT×DW	-0.109		
CT×FLAIR	-0.068		
DW×FLAIR	-0.015		
CT×T1Gd	0.094		
DW×T1Gd	-0.154		
FLAIR×T1Gd	0.003		
CT×DW×FLAIR	0.048		
CT×DW×T1Gd	-0.084		
CT×FLAIR×T1Gd	-0.087		
DW×FLAIR×T1Gd	-0.014		
CT×DW×FLAIR×T1Gd	0.032		

Table 3.1: Coefficients for the modeling conditions 1,2,3. Condition 1 involves the full model and conditions 2,3 involve the reduced model for the 2 mm and 4 mm layers.

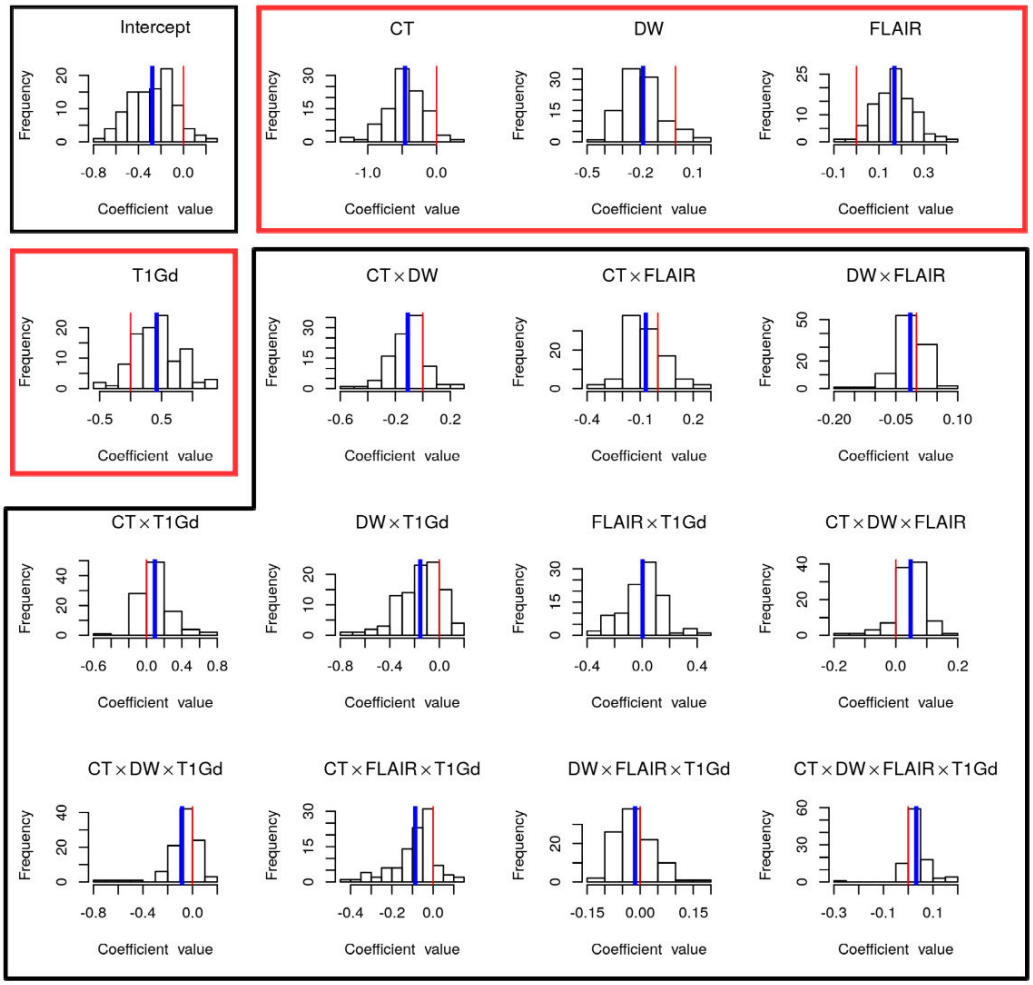


Figure 3.22: Coefficient histograms for Condition 1 involve a large number of coefficients which include all the interactions between the $CT, DW, FLAIR$, and $T1Gd$. Coefficients involving individual items (e.g., $FLAIR$ shown in the red boxes) seem to have a bigger impact than the complex items such as the coefficient corresponding to $DW \times FLAIR \times T1Gd$. The interaction coefficients are not statistically significant. One can conclude that interaction between images (black panel box) plays no role in the prediction of the recurrence.

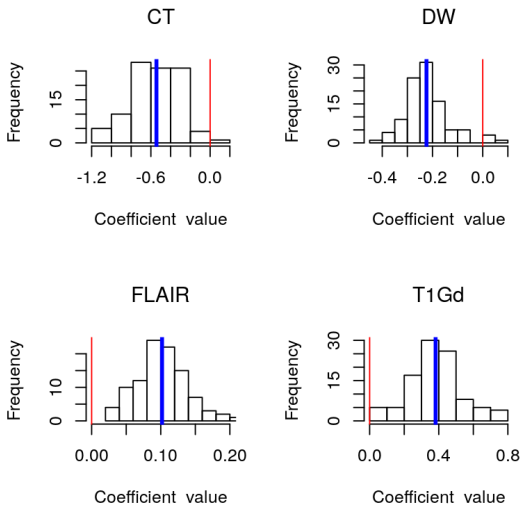


Figure 3.23: Coefficient histograms for Condition 2. The blue line depicts the mean and the red line corresponds to the value 0. The further the blue line is from the red line the more relevant the coefficient is.

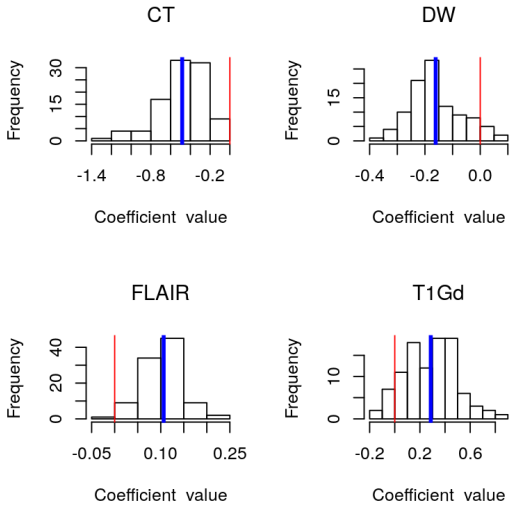


Figure 3.24: Coefficient histograms for Condition 3 which corresponds to the layer of 2-4 mm. The results of these coefficients seem consistently similar to those involving the 2 mm layer in figure 3.23.

3.3.3 Decision trees

The results of the tree models, conditions 4 and 5, are presented in this section. The tree model is described by equation 3.7. Examples of the decision trees for condition 4 are shown in figure 3.25 and 3.26. There are four variables in the models, the $ctt, flairt, adct$, and $t1gdt$. The variable names might sound a little bit unusual but this is because the tree is plotted with the actual names of the variables used in the programming code and these names were left in the tree to reduced a possible systematic error of typing the wrong variable name. Instead, a short table (3.2) is presented as a variable change guide. For instance, the *adct* variable name displayed in the tree diagrams refers to the *DW* variable. To read a decision tree first read

Medical Image	Actual variable in tree	Variable name
CT-scan	ctt	CT
DW-MRI	adct	DW
T2-Flair	flairt	FLAIR
T1-Gd	t1gdt	T1Gd

Table 3.2: Guide for the variable names of the tree models. As an example, the *ctt* variable displayed in the tree corresponds to the *CT* variable.

the variable at the top of the tree, for example take a look at the variable *ctt* in figure 3.25. If the *CT*[4] of a pixel is greater than 0.22 the branch to the left is followed, and if the *T1Gdt* is less than 0.54 and the *DW* (written as *adct*) is greater or equal to -0.74, then there is a probability of 0.23 that the position of the voxel corresponds to a location where the recurrence appears. The percentage 18% right next to the probability of 0.23 refers to how likely it is for the 0.23 probability to occur. These two values are in a box which are known as leaves of a decision tree. The tree leads to many leaves or possibilities. Notice that the percentages of the leaves must add up to 100% and can also be interpreted as the higher the percentage the higher the importance of the leaf.

From the second tree in figure 3.26, (which still corresponds to condition 4), we can see an interesting prediction of 0.94 probability of the location of the voxel to be located inside the recurrence region if the *CT* is less than -0.21 and *T1Gdt* is less than -1.1. Even though the prediction is quite good, the chances of occurring are very low, less than 8%. Therefore it is not a

[4]Notice that the ctt variable correspond to the CT variable.

likely prediction. A better predictive leaf would be the 0.80 probability leaf with 19% chances of occurring.

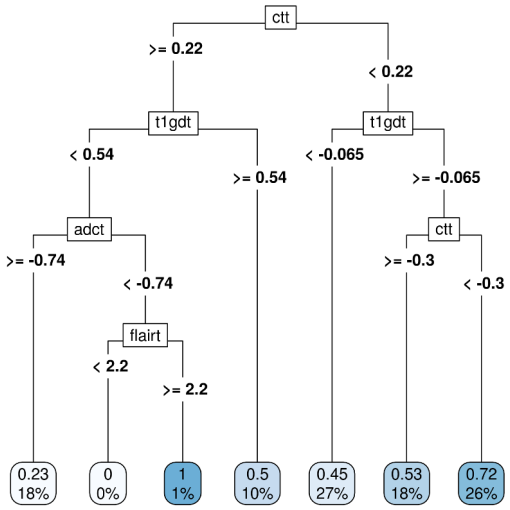


Figure 3.25: Decision tree 1 for condition 4 in which decision leaves are depicted at the bottom of the tree. The first leaf states that there is a 0.23 probability of a voxel, having the pixel values corresponding to the branch, to be inside the recurrence. The 18% right next to it is related to how likely it is for the 0.23 probability to occur.

A decision tree concerning condition 5 is shown in figure 3.27. The fourth leaf from left to right of the tree in figure 3.27 is one of the best predictions of the tree, but it is still not a good prediction. This leaf predicts a 0.6 probability of recurrence in a particular voxel satisfying the variable conditions of CT less than 0.0068, T1Gd greater than -1. It is perhaps the best prediction not because it gives the highest prediction of recurrence but because it is the best balance leaf between high prediction and occurrence rate (of 55%). Nonetheless, we can observe all the tree leaves (in the blue rectangles with round edges at the bottom of the trees) and not a single one is able to strongly predict the recurrence location with sufficient accuracy.

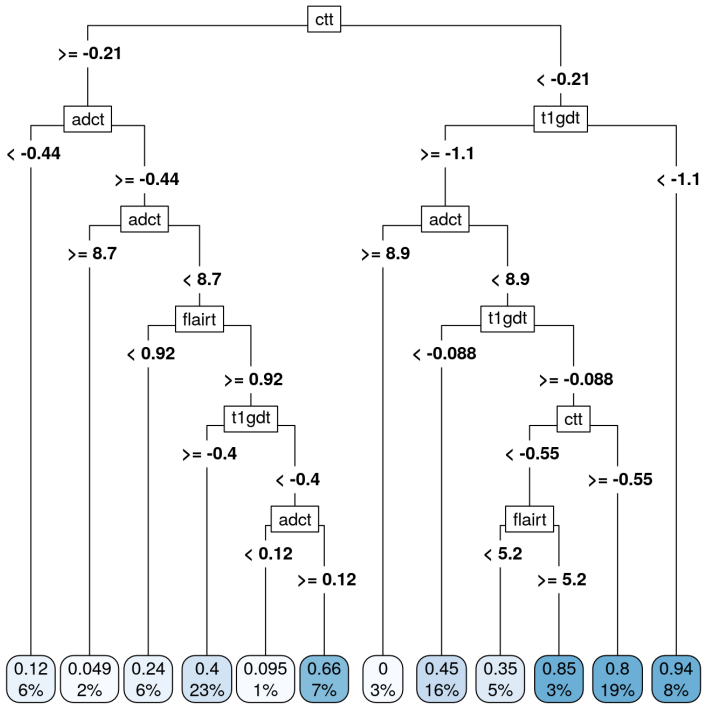


Figure 3.26: Another example of decision tree for condition 4. The last leaf predicts a 0.94 probability of occurrence but the 8% right next to it indicates that the 0.94 probability is not likely to occur.

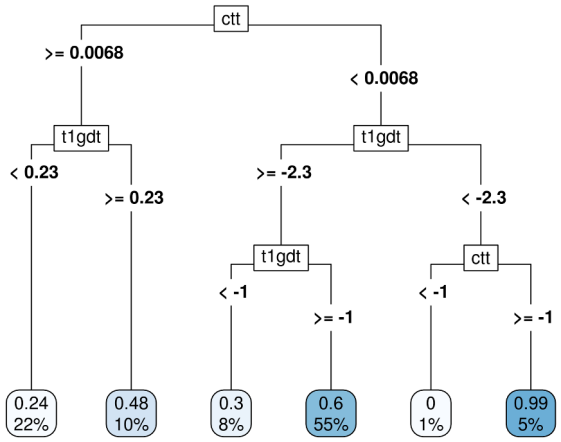


Figure 3.27: Decision tree for condition 5. Even though the fourth leaf from left to right makes the fines prediction from all the trees presented it is still not sufficiently likely to occur. That leaf predicts a 0.6 probability of recurrence to appear with 55% for this leaf to occur.

3.3.4 Receiver Operating Characteristic (ROC) space

The ROC space for condition 1 is shown in figure 3.28 and for condition 2 and 3 are shown in figure 3.29. Overall the ROC space for the reduced model in figure 3.29 show a slightly higher TPR than those corresponding to the full model in figure 3.28. Condition 2 seems to make a slightly better prediction than condition 3 meaning, layer 0-2 mm is more relevant.

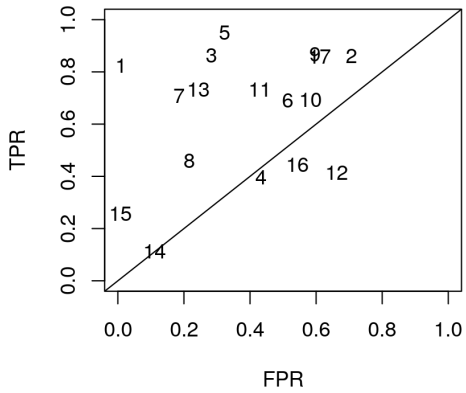
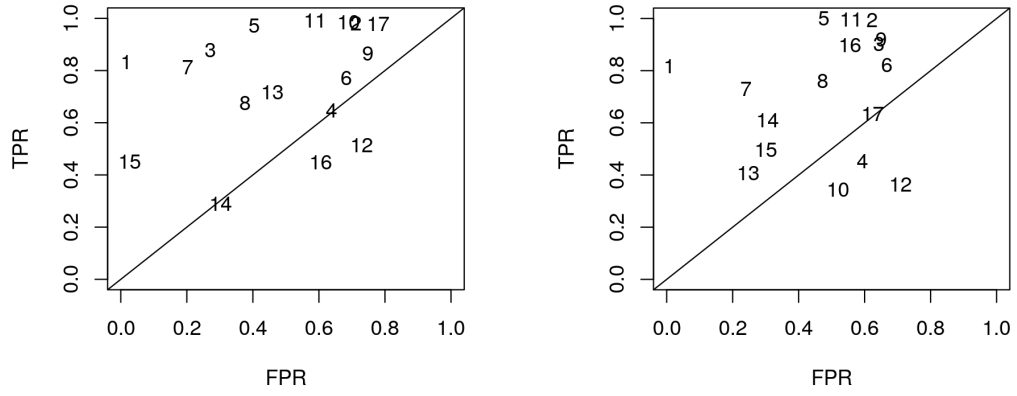


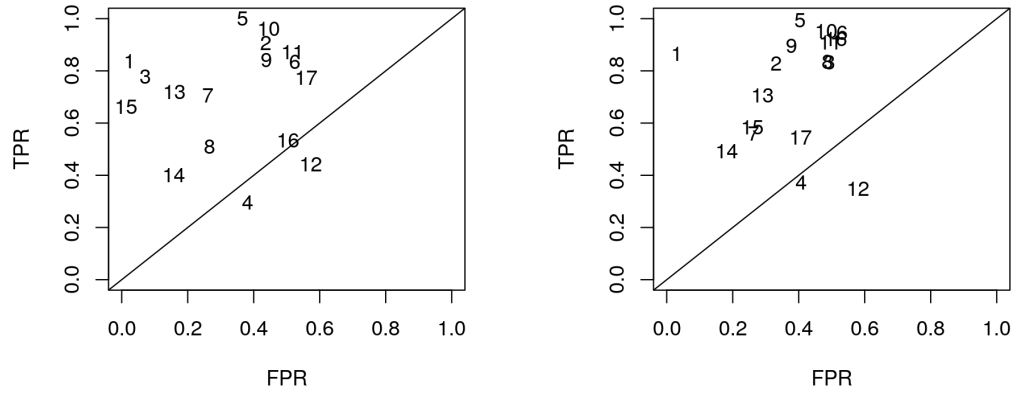
Figure 3.28: Receiver Operating Characteristic space (ROC) for full GLM model, condition 1. A reasonable amount of positive predictions are made but the model also wrongly predicts.

The ROC space compare the accuracy of the prediction of 17 patients, each number inside the ROC space represents a patient. Patients 1 and 7 are some of the best predicted patients since their TPR are high and the FPR relatively small. It means that the location of the recurrence was correctly predicted with low false alarm predictions. However, as can be seen in the plots, the majority of patients are concentrated in a TPR from 0.4-0.90 with a FPR from 0.0-0.70. The models do not predict with sufficient confidence the recurrence location.



(a) Condition 2 **(b)** Condition 3

Figure 3.29: Receiver Operating Characteristic (ROC) spaces for the reduced GLM model, conditions 2 and 3. Each number on the plot represents a patient. The ideal prediction would be the top left corner meaning high correct predictions and low incorrect predictions.



(a) Condition 4 **(b)** Condition 5

Figure 3.30: Receiver Operating Characteristic (ROC) spaces for the tree model, conditions 4 and 5. The tree ROC space show an improvement compared to the GLM models but despite such improvement, a strong reliability of the prediction can not be established.

The decision tree ROC space for conditions 4 and 5 are presented in figures 3.30. The ROC space for the tree model compared to the GLM models seem to be more reliable since the FPR is smaller and the TPR prediction is slightly higher and more constrained in the region ranging from about 0.3 to 1.0. For instance, we can see that the ROC space for patient 14 is slightly better for the tree model. Despite the improvement in the model, it is hard to establish

a strong reliability on the tree models as well. These ROC spaces allow us to quickly evaluate the model predictions of the 17 patients in a single graph.

3.3.5 Prediction maps

In this section, visual representations of the results of the models are presented using image slices. Patient number listed as 7 in the previous ROC spaces is used to construct the prediction maps. The slide $z = 30$ is displayed for the CT-scan, DW or Diffusion MR, the T2-Flair MR, and the T1-Gd MR (figure 3.32 and 3.33).

Figure 3.31 illustrates the prediction maps. First, the recurrence models outputs as a result a probability value ranging from 0 to 1 in which the higher the value the higher the probability that the voxel belongs to a recurrence region according to the model. When the model predicts a probability higher than 0.5 then it is interpreted as "yes" the voxel is in the recurrence region, if it is less than 0.5 then the voxel is not in the recurrence region. Since the recurrence region is known, then we can compare the (computer) model prediction with the actual findings in real life from the medical database. The True Positive (TP), refers to the correct prediction of a recurrence region shown in red in figure 3.31. The green TN refers to correct prediction of non-recurrence regions. In other words the red (TP), and green (TN) are the correct predictions and the blue (FP) and purple (FN) are the wrong predictions.

The prediction map in figure 3.32 is the prediction when using condition 2. If we take a look at the diffusion image (DW-MR) we can see that there is a recurrence (In french, *recidive*) contour which is mostly inside the GTV and some peaks of the recurrence surpasses the tumor. The left top corner peak colored in purple incorrectly predicts that there is no recurrence where as the peak on the right bottom corner of the recurrence correctly predicts the recurrence in red. The blue color on the bottom left and top of the GTV contour can be interpreted as the incorrectly detected recurrence. The green region refers to the region where there is no recurrence and the model correctly predicted the non-recurrence region.

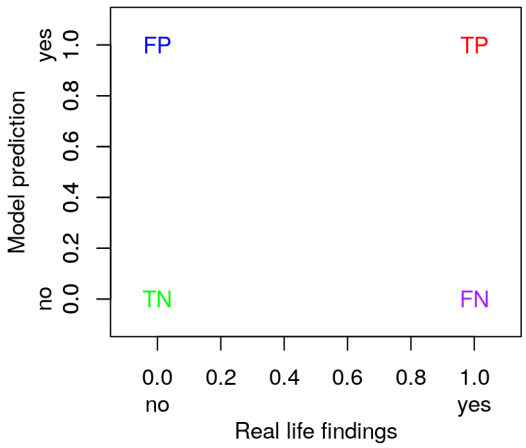


Figure 3.31: Prediction map guide. The True Positive (TP) in red and the True Negative (TN) in green correspond to the correct predictions. The incorrect predictions are the False Positive (FP) in blue and the False Negative (FN) in purple.

The prediction map in figure 3.33 uses condition 4 (which correspond to the tree model) for the prediction results. The tree and GLM prediction maps resemble considerable well each other. Nonetheless, from this $z = 30$ slide the maps seem to slightly better predict when using the tree model but it is dubious. This argument is based on the observation that the tree model better predicts the top left corner of the recurrence region since this region is red for the tree prediction map where as for the GLM model the color is purple. On the negative side, the right bottom peak of the recurrence is less well predicted in the tree model which can be observed since for the tree model is half red, half purple where as for the GLM is entirely in red.

Evidently, conclusions can not be made using slices of a single patient. The prediction maps generated served mostly as a guide to visualize the results. A further discussion concerning the recurrence prediction based on pre-treatment imaging is addressed in the conclusion section.

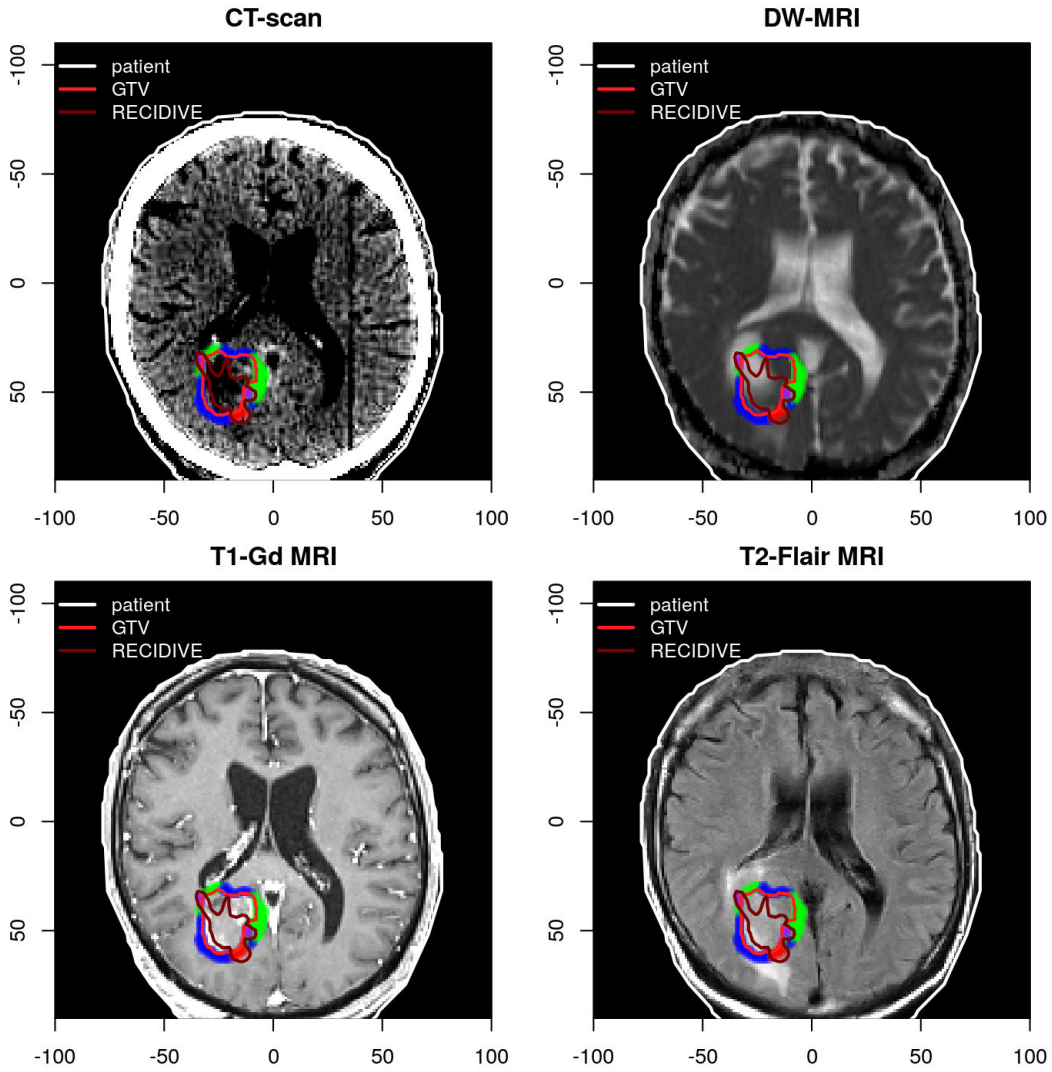


Figure 3.32: Prediction map, involving the GLM model, overprinted on the CT-scan, DW, T1-Gd, and T2-Flair. The voxels corresponding to a layer are displayed in colors surrounding the GTV contour. The red color means correct prediction of recurrence, green means correct prediction of non-recurrence. The false recurrence prediction is depicted by blue and the erroneously overlooked recurrence is in purple.

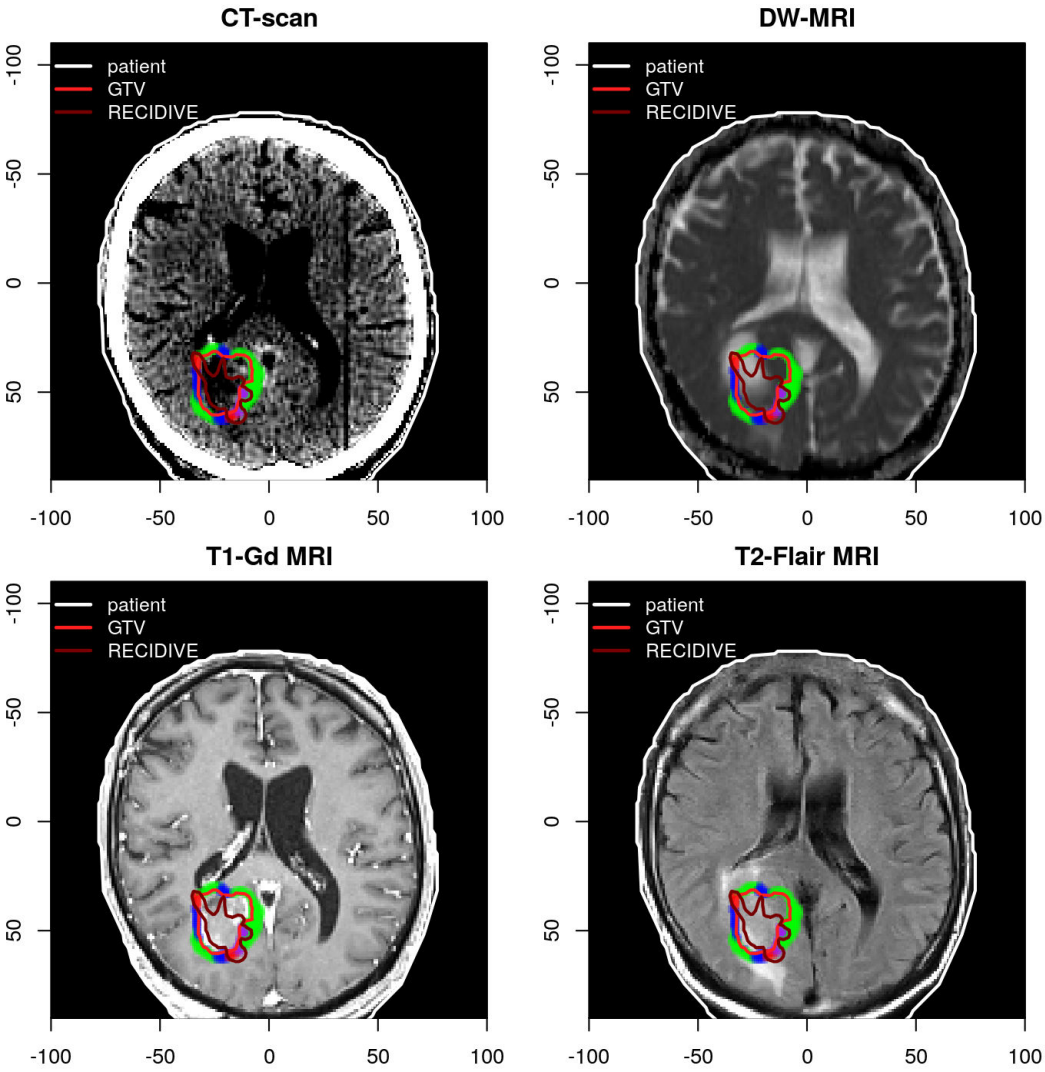


Figure 3.33: Prediction map, involving the tree model, overprinted on the CT-scan, DW, T1-Gd, and T2-Flair. The correct prediction of the recurrence does not seem to have improved in comparison to the GLM model prediction in figure 3.32. However, the correct identification of non-recurrence area (green color) has significantly improved.

3.4 Conclusions

The results obtained in this work concur with a previous thesis study [44], in which both the non-recurrence and recurrence peritumoral display higher ADC values compared to healthy brain (for our case, the mirror area in the brain). The ADC intensities of the recurrence regions have been found to have lower values than the non-recurrence peritumoral area; the results show a small but intriguing difference. In other words, the recurrence region displays a small decrease in ADC, however there is an increase in FLAIR values. A study from the literature found that the ADC and FLAIR values decreased 9.5% ($p < 0.001$) and 9.2% ($p < 0.001$) respectively in the recurrence peritumoral edema versus peritumoral non-recurrence regions [9]. Our findings are in agreement with respect to the ADC but contradict the FLAIR values. In that study they suggest that using multi-parametric logistic model seem to better predict the recurrence region than a single intensity value alone. In our work, we have performed multi-parametric value (DW, T2-Flair, T1-Gd, and CT intensity values). Indeed, the multi-parametric modeling better predicts the recurrence. However, after performing several multiparametric GLM and decision trees, the models still can not provide a definite answer. Therefore this led us to conclude that we consider improbable for models to be able to predict the recurrence with absolute certainty only by using the current MRI data (before and after the treatment). Several possibilities exist: the difference in intensities values in the recurrence is too small which makes it difficult to perceive with current imaging resolution, intermediate MRIs are needed for the evolution of the tumor and recurrence, the different calibration of MRIs from the different clinics are creating a bigger disruption than previously anticipated, lastly the possibility that pieces of additional information are missing to be able to predict the recurrence.

Structure mapping visual representation tools

4.1 Introduction

The purpose of this chapter is to try to understand why the previous recurrence machine learning models did not work sufficiently well. This is accomplished by developing visual representation tools in which the intensity values on the surface of tumor like structures are plotted onto 2D maps. The goal of these maps is to try to identify possible patterns of intensity values corresponding to the recurrence versus non-recurrence locations. If clear patterns are not detected then that would suggest the current medical data is insuffi-

cient to be able to construct strong competent recurrence prediction models.

The analysis of 3D structures are central issues in the medical imaging area [39]. The analysis of 3D structures includes a wide variety of applications in medical physics. For instance, the study of the brain in neurology and the analysis of proteins in molecular biology. That is, 3D structure analysis applications include the mapping of the brain [33] and proteins onto a unit sphere [48].

The mapping of the surface of 3D structures is referred as embedding or surface parameterization [48]. The embedding of brain structures provides a way to compare brains which could potentially be used for detecting abnormalities in the brain. The advantage of embedding the brain onto a unit sphere is that the surface of the sphere can be plotted onto a 2D map for a rapid analysis. Such convenience is quickly observed by noticing the usefulness of a 2D regular world map in which a nearly spherical shape is mapped onto a 2D map. Such world map allows for a rapid identification of geographical locations, weather related forecast illustrations, among innumerable other applications. The convenience of displaying information of 3D structures using a portable 2D display is perhaps the biggest advantage. That is precisely the representation tools we are looking for; to be able to represent the surface of the expanded tumor structure onto a 2D map.

The visual representation tools originated as a by-product of the recurrence analysis work presented in the previous chapter. So far, the current method for presenting the recurrence prediction is done by choosing a single slide of a medical image such as a MRI slide and display the predictions on the 2D image. However, there are many slides meaning that only a small part of the prediction is observed at a time. Hence a better visualization tool would be invaluable.

Previous methods of visualization refer to the risk maps in previous chapter

Instead of displaying the recurrence using a single medical image brain slide we would ideally like to display all the recurrence and non-recurrence intensity values in a single portable 2D map. Notice that the new visualization tools are additional tools to analyze the recurrence and not a substitution of previous methods of visualization used in the previous chapter.

4.2 Projecting the surface of a 3D structure onto a unit sphere then to a 2D map

A road map illustrating the steps taken for the development of the visualization tools is shown in figure 4.1. The first main step is to choose a suitable structure of interest such as a tumor or an expanded tumor. The second main step is to created a mesh covering the surface of the 3D structure[1]. With the (x,y,z) coordinates of each vertex on the mesh we can calculate the latitude and longitude angles implementing concepts of the heat diffusion equation. Each vertex of the mesh is associated to a latitude and longitude angle. Obtaining the angles can be tricky; a detailed description is provided in this chapter. Following, the calculated latitude and longitude angles are used to project the 3D surface onto a unit sphere. Lastly, a 2D map of the surface of the sphere is created.

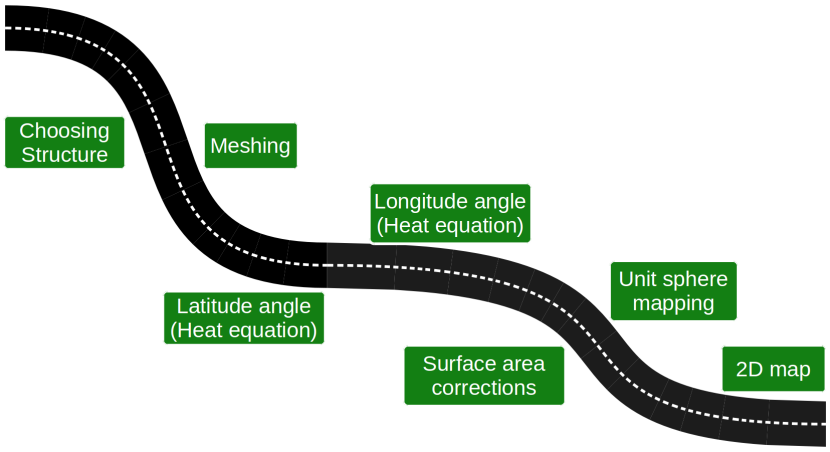


Figure 4.1: Road map of the main steps taken to develop the visualization tools; the final objective is to create a 2D map.

4.2.1 Mesh of the 3D structure

One crucial requirement for creating the mesh is that the surface of the tumor structure belongs to the genus 0 surface category meaning the structure does not have any holes. Fortunately, a good amount of the surfaces of tumors

[1]Triangular meshes were constructed for all the meshes in this work.

seem to be suitable structures of the genus 0 category. Using the contours, provided by the medical images in the Glioblastoma database, we were able to select the voxels corresponding to the inside or outside of a contoured structure of interest such as the tumor. The voxels inside the 3D contour were assigned a value of 1 and voxels located on the outside were assigned the value of 0. With this information we can reconstruct tumor structures. Examples of 3D reconstructed tumor structures are displayed in figure 4.2.

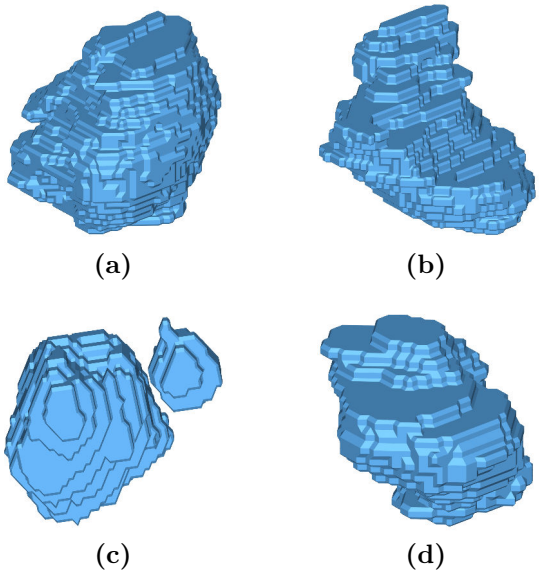
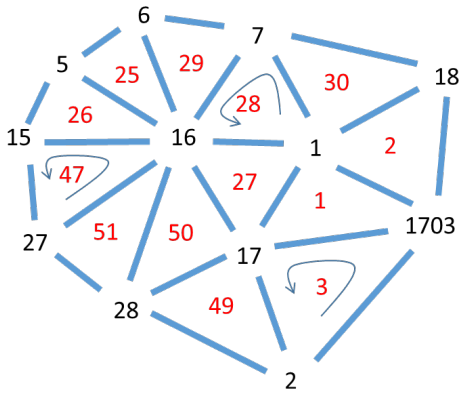


Figure 4.2: Illustration of 3 dimensional tumor structures, (a),(b), and (d) are suitable structures for the meshing process but (c) is not allowed since it is composed of two structures.

The Glioblastoma tumor structures have a highly variable form as can be observed. For instance, tumor (a) has a much more spherical looking shape compared to the rest of the tumors, where as (b) has somewhat the shape of a boot with the long top and thick and flatten bottom. The tumor structure (c) is not a suitable structure to be parameterized since it is composed of two GTV volumes; a main one and a smaller one at the top right hand side of (c). The tumor 3D structures visually illustrates the kind of structures that are suitable for the meshing process.

A triangular mesh of the outer surface of the structure was created using the function "contour3d" from the package "misc3d" of the R software which uses the Marching Cubes algorithm. The algorithm calculates triangle



(a)

Triangles of a vertex
17) 1 3 27 49 50
Vertices of triangles
1) 1 17 1703
3) 2 1703 17
27) 17 1 16
49) 2 17 28
50) 17 16 28
Neighboring vertices of a vertex
17) 1 2 16 28 1703

(b)

Figure 4.3: Representation of the triangular meshing (a) and the contents of the mesh (b). The mesh data contains the triangles of a vertex, vertices of triangles, and the neighboring vertices of a particular vertex.

vertices implementing linear interpolation [40]. This algorithm is very useful for processing 3D medical imaging data for the purpose of creating a mesh on the surface of a structure. Henceforth, we used the algorithm for creating the surface mesh using triangular parameterization.

The function "contour3d" was used to obtain a list of vertices *V1*, *V2*, *V3* in which each vertex contains (x, y, z) coordinates; this is the pre-mesh. This pre-mesh was used to create a new mesh which instead of containing a list of the vertices containing coordinates, is composed of a list of triangles of vertices, the vertices of the triangles, and the neighboring vertices of each vertex. A representation of the meshing and the contents of the mesh data are displayed in figure 4.3. For instance, vertice 17 contains triangles 1,3,27,49, and 50. Triangle 1 is composed of vertices 1,17, and 1703. Lastly, the neighboring of vertices of a vertex are written down. For example, vertex 17 has immediate vertices 1,2,16,28, and 1703. The mesh was re-organized in this manner because it was needed for ease of manipulating the mesh data.

A real triangular mesh plotted on the surface of the tumor structure, figure 4.2 (d), is displayed in figure 4.4. The visible small triangles covering the surface of the tumor structure create the triangular facets and they are about equal size but due to the angle of perspective some facets look smaller

than the rest.

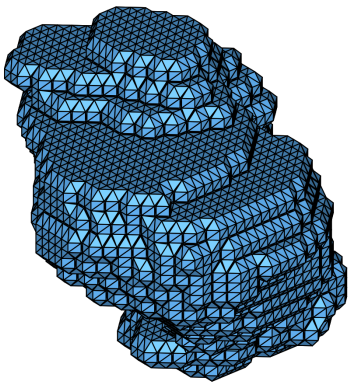


Figure 4.4: Illustration of a triangular mesh covering the surface of the tumor structure shown in figure 4.2 (d). The small triangular facets composing the mesh are clearly visible.

4.2.2 Latitude and longitude angles of the vertices of the mesh

We used as a foundation the algorithm described by Brechbühler et al., [6] for solving for the latitude and longitude angles associated to each vertex of the triangular mesh. Their work provides one of the most practical algorithms for performing the embedding of complex surfaces of genus 0 onto a unit sphere.

Latitude angle calculation

The latitude angles were found using the premise of the distribution of heat in which a hot point is arbitrarily chosen which we called North Pole then another point as far away as possible was also chosen but this time it would be a cold point which we named South Pole. The North Pole was assigned a temperature of 1 where as the South Pole was assigned a temperature of 0. The idea is to construct a set of equations that model the heat distribution by solving for a temperature value at each vertex of the triangular mesh in descending heat from the hottest point to the coldest point. The value of the solved temperature is directly related to the value of the latitude angle.

Key point between temperature and latitude angle!

The equation to model the heat distribution is given by the Laplace equa-

tion,

$$\bigtriangledown^2\theta = 0 \qquad BC \left| \begin{array}{c} \theta_N = 1 \\ \theta_s = 0 \end{array} \right. \tag{4.1}$$

After setting the temperature Boundary Conditions (BC) of North Pole = 1, and South Pole = 0. then the heat equation becomes,

$$A\theta = b \tag{4.2}$$

This equation was used to solve for the temperature value at each of the vertices of the mesh. In the equation, θ is a matrix of number of rows equals to the total number of vertices of the mesh and of one column; θ represents the temperature values (the values of the latitude angles)[2]. A is a matrix that associates distances to vertices and matrix b is an auxiliary matrix and it is of the same size as matrix θ. The expanded matrix version of equation 4.2 is,

$$\begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,j} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i,1} & A_{i,2} & \cdots & A_{i,j} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_i \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_i \end{pmatrix} \tag{4.3}$$

The index i and j represents the number of vertices of the triangular mesh of an structure. The index j is of the same size of i. The elements of matrix A are associated to a vertex and the diagonal values indicate the sum of all the distances of the neighboring vertices to the vertex of interest. The indexes (i, j) of the diagonal elements indicate the vertex of interest. The θ_i values are the variables to solve for, and they are related to $vertex_i$. For instance, θ_1 is the latitude angle associated to vertex one. A guided construction of equation 4.3 is given below to better understand.

Let us use the first row of A to illustrate the construction of matrix A. The index $i = j = 1$ of the diagonal element $A_{1,1}$ indicates that the vertex of interest of row one is vertex one. $A_{1,2}$ is associated to vertex two, $A_{1,3}$ is associated to vertex 3 and similarly for all the first row. $A_{1,j}$ is the distance from vertex j to the vertex of interest, which in this row is vertex one. That is, $A_{1,2}$ is the distance from vertex 2 (indicated by $j = 2$) to the vertex of interest which is vertex one. $A_{1,j}$ is assigned a value of zero if vertex j is not a neigboring vertex of the vertex of interest. In a similar manner we find that the vertex of interest of the second row is vertex two which is identified by

[2]The latitude angles are equals to $\theta \times 2\pi$.

the indexes ($i = j$) of the diagonal element given by $A_{2,2}$. Then the distances of the neighboring elements j to the vertex of interest are calculated for the rest of the columns of the second row. Similarly, all the other rows are filled in the same way.

The single row matrix b is the remaining matrix to be filled. First, the vertex index i of the direct neighbors of the South Pole are identified. Then, the value of 1 (which represents the maximum temperature value) is assigned to the elements b_i only if vertex i is a direct neighbor of South Pole and the value zero is assigned for all the rest of the elements of b_i.

Lastly, the system of linear equations in 4.3 is solved for the latitude angles θ_i. The filling of matrix A, using the real structures such as tumors, results in a massive matrix but mostly empty. Solving for equation 4.3 for a huge matrix becomes complex. The solution implemented was to define A as a sparse matrix and solve equation 4.3 using the sparse matrix solvers in R. The reason why matrix A is so big is because the size of A depends on the number of total vertices of the triangular mesh and the mesh of real structures consist of a huge number, several thousands, of vertices. However, most elements in A are zero values since for each row only the distances $A_{i,j}$ associated to the immediate neighbors of the vertex of interest are calculated and the rest of the elements are assigned the value of zero as stated in the algorithm that we are following [6]. This situation leads to a matrix in which most elements are zero hence defining A as a sparse matrix is necessary to be able to solve equation 4.3.

An illustration of the heat distribution temperature (values of the latitude angles) values at each of the vertices of the triangular mesh is shown in the top left corner of figure 4.5. The temperature is represented in colors ranging from the red color, hottest temperature of 1, to the blue color representing the coldest temperature of zero. The North Pole is displayed with a small red sphere on the upper part of the structure and the South Pole is at the bluest bottom part. The nearly uniform green color clearly depicts that the latitude angle θ is not well distributed as can be also confirmed by the top right plot in which most of the θ values are concentrated at a value of about 0.6. The step of making data more uniformly distributed was then required. Such step was performed by normalizing data by means of the empirical cumulative distribution function: $\theta_N = ecdf(\theta) \times \theta$. Where θ_N is the normalized latitude. The normalized latitude values are plotted in the left bottom side of figure 4.5. The heat distribution can be seen to have greatly improved by the well differentiated rainbow colors on the the 3D

structure and by the wider range of distribution of values displayed on the bottom right side in figure 4.5.

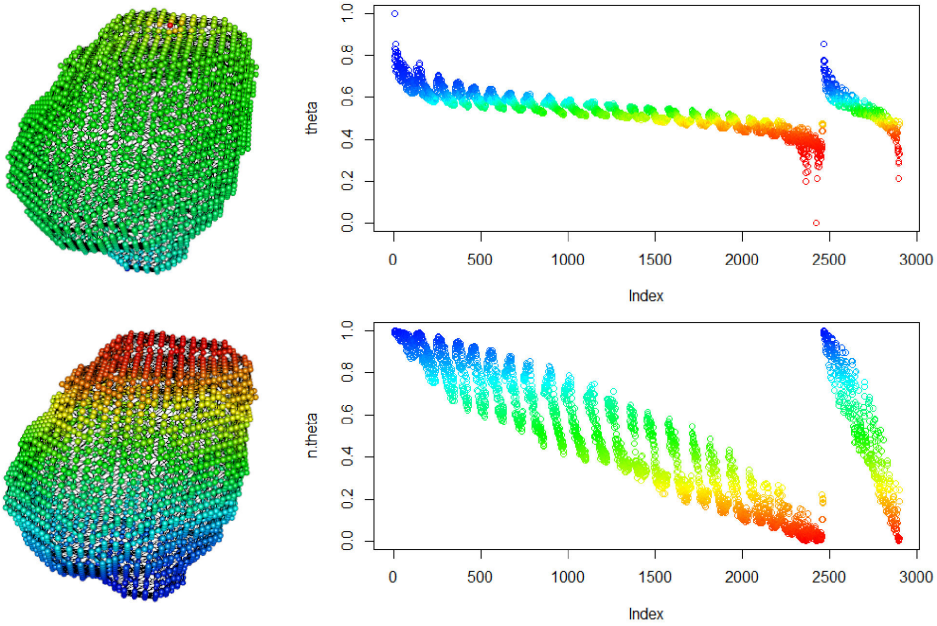


Figure 4.5: Latitude angle heat distribution illustration. The heat distribution is represented in colors in which the hottest spot is represented in red (North Pole) and the coldest point in blue (South Pole). The top figures shown a non-satisfactory distribution in which most latitude θ values are concentrated somewhere around the 0.6 value. After performing a normalization procedure, the heat distribution greatly improves (bottom part) as seen by the well differentiated rainbow colors on the 3D image and by the wider distributed range of normalized latitude θ_n values.

Longitude angle calculation

The second angle that we need in order to achieve the spherical parameterization is the longitude angle. The longitude angle is a bit more complex to calculate but the heat distribution premise remains the same.

The first step is to create a departure line by finding the line of descending temperature from the North Pole to the South Pole which is equivalent as saying descending value of latitude angle. Following, another path is created right next to the departure line, almost parallel to it, which we are going to call it the arrival line. A point from the departure line is selected, and assigned a minimum temperature of 0, and its closest arrival point is

selected as well but assigned a maximum temperature value of 1. A path is created between the departure and arrival points in which the heat is evenly distributed. The process is performed for all the departure points in a cyclic manner resulting in a temperature value at each vertex of the mesh.

Summarizing the important steps to calculate longitude angles:

- Find the departure and arrival lines
- Paths are created between departure points and their immediate arrival points
- The heat is evenly distributed between the paths by the heat distribution equation
- The heat distribution equation is solved using system of linear equations in a similar way as for the latitude angle, θ, heat distribution problem.
- The even distribution of the heat can be verified by plotting the temperature values (values of the longitude angles) on each vertex.

The following Laplace equation models the distribution of heat,

$$\bigtriangledown^2\phi = 0 \qquad BC \left| \begin{array}{l} \phi_{0^+} = 0 \\ \phi_{0^-} = 1 \end{array} \right. \tag{4.4}$$

After setting the Boundary Condition (BC) of $\phi_{0+} = \phi_{0-} + 1$ the heat equation yields,

$$A'\phi = b' \tag{4.5}$$

The boundary condition refers to the cyclic process of evenly distributing heat between the departure points and the arrival points in which a sudden change of temperature happens between the arrival point (temperature of 1) and the newly departure point for a new cycle. The value of 1 refers to the maximum temperature. In a similar manner as for the latitude angle, a distance related matrix A' and the auxiliary matrix b' are filled.

An example of a departure line is shown in figure 4.6(a) with the color magenta and points on the arrival line are depicted as green spheres immediately on the right side of the departure line. Then, after solving the system of linear equations we obtained the longitude angle, ϕ, values. These values are plotted in figure 4.6(b); the arrival and departure lines are not plotted in (b) but the drastic drop of temperature from deep red to blue clearly indicates it.

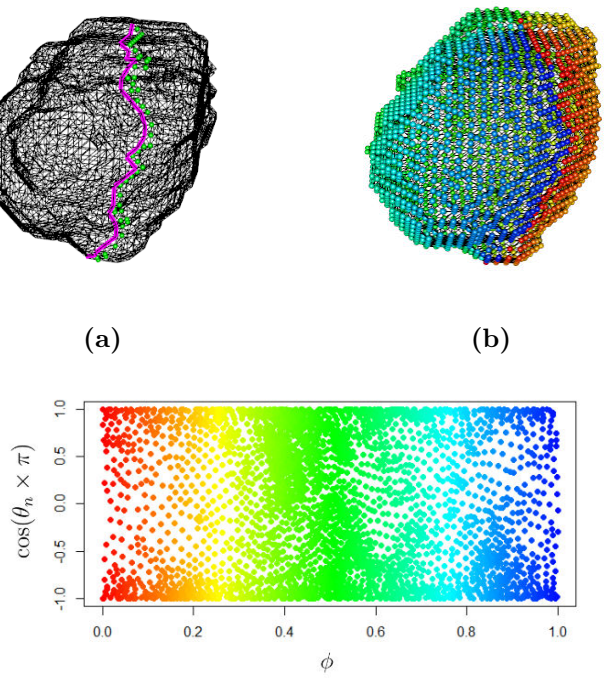


Figure 4.6: Illustrations concerning the longitude angle heat distribution. The departure or date line (magenta color) and the arrival line (green spheres on the right side of the magenta line) are depicted in (a); the departure line follows a path of descending latitude angle values. The even heat distribution around the surface can be seen in (b). Lastly, a Mercator projection of (b) is displayed in (c) which shows a non-homogeneous distribution of points.

Notice the heat distribution difference concerning the values of the latitude angles in figure 4.5 in which the heat is distributed from the North Pole to the South Pole as opposed to around the surface as it is in the longitude angle heat distribution problem.

A Mercator projection of the surface of the sphere was also created since we have the polar angles, latitude and longitude, associated to each vertex of the mesh on the surface of the sphere. Out of the hundreds of existing map projections, the Mercator projection provides a relatively straight forward mapping. To construct the Mercator projection we can use the following

equations [52],

$$x_{Mercator} = \phi \tag{4.6a}$$

$$y_{Mercator} = \ln\left[\tan\left(\frac{1}{4}\pi + \frac{1}{2}\theta\right)\right] \tag{4.6b}$$

Where, θ is the latitude angle (or polar angle) and ϕ is the longitude angle (or azimuth angle). A Mercator projection of the surface in figure 4.6(b) is shown in (c). Unfortunately the 2D map does not provide a satisfactory enough even distribution of points but it does show a correct heat distribution in the ϕ direction ranging from blue to red. Therefore an optimization must be performed which is addressed in the following section.

Lastly, an illustration of some latitude (red) and longitude (green) lines are depicted on the surface of a structure in figure 4.7. The bold thicker green line represents the date line. The illustration allows us to better visualize the problem.

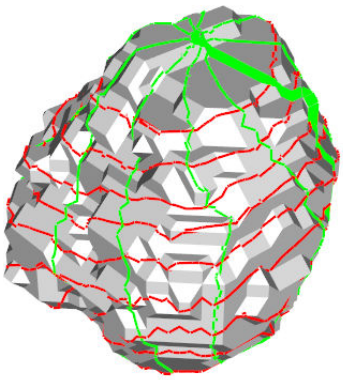


Figure 4.7: Illustration of latitude and longitude lines on the surface of a 3D structure. The red horizontal lines represent the latitude lines and the green vertical lines represent the longitude lines. The thicker bold green line depicts the date line.

4.2.3 Optimized Mollweide projections

The latitude, θ, and longitude, ϕ, angles previously calculated are essential for the spherical mapping. By converting these polar coordinates to Cartesian coordinates we can plot the surface of a 3D complex structure of interest onto a unit sphere; that is called spherical parameterization. For a sphere of radius 1, the conversion yields,

$$x = \sin(\theta\pi)cos(\phi 2\pi), \tag{4.7a}$$
$$y = \sin(\theta\pi)\sin(\phi 2\pi), \tag{4.7b}$$
$$z = \cos(\theta), \tag{4.7c}$$

This means that the vertices (V_i) of a mesh covering the surface of the complex 3D amorphous structure are transformed to Cartesian coordinates (x_i, y_i, z_i). Each pair of latitude and longitude angles are associated to a unique set of x, y, z values; this is called bijection. A spherical parameterization using equations 4.7(a,b,c) is shown in figure 4.8. The wired frame shows the uneven distribution of the parameterized vertices. A histogram of the surface area created by the vertices of the mesh is also plotted which clearly shows the serious discrepancies in the surface areas. The mean value of the surface area is depicted by the red vertical line in the histogram. Following, a Mollweide projection of the spherical mesh was created and it is depicted in the top right corner of figure 4.8. The Mollweide projection was used since it provides the area preserving advantage and it is widely used. The equations used for the Mollweide projection are [25],

$$x_{Mollweide} = (2\sqrt{2}/\pi)\phi\cos(\alpha) \tag{4.8a}$$
$$y_{Mollweide} = \sqrt{2}\sin(\alpha) \tag{4.8b}$$

where the angle α is defined by solving,

$$2\alpha + \sin(2\alpha) = \pi\sin(\theta)$$

similarly as before, θ represents the latitude and ϕ the longitude. In order to solve the Mollweide projection equations we used a range from $\pi/2$ to $-\pi/2$ for the latitude and a range from $-\pi$ to π for the longitude. The color of the Mollweide projection in figure 4.8 indicates the size of the surface areas; the colors in increasing size are red, yellow, green, blue. The small surface areas are highly concentrated near the poles which can be observed by the red color at the poles, whereas near the equator the color is mostly green and blue indicating a larger surface area.

The issue of uneven distribution of surface areas illustrated in figure 4.8 was addressed by developing an optimizer. This is an issue because the uneven surface areas would mean that we could not generate a reliable 2D conformal map. The approach taken for the optimization was to shrink the biggest triangles composed of 3 vertices on the 2D Mollweide map. We apply the idea of energy minimization in which the energy is analogous to the area of the triangles. The idea is that by compressing the area of the biggest triangles the smallest triangles increase. The reduction of the area of the biggest triangles is done gradually in a loop by performing hundreds, of iterations. For this operation, we select all triangles whose surfaces are greater than 4pi/N. For each triangle, we compute its gravity center and reduce the length from the mesh to the gravity center by 10%. We observed a gradual improvement of the homogeneity of the size of the triangles. For the specific cases in this work we used around 300 iterations for the optimization.

The optimized results are shown in figure 4.9. The even surface areas can be observed on the surface of the sphere. The average of the surface areas is depicted by the red vertical line in the histogram and we can see that the values are highly concentrated around the mean. The vertices of the surface of the sphere are plotted on the Mollweide map with the color representing the area of the surface areas. The even distribution of the green color clearly shows a great improvement in the even size of the surface areas. We can conclude that the optimized version displays a reliable 2D projection conformal mapping.

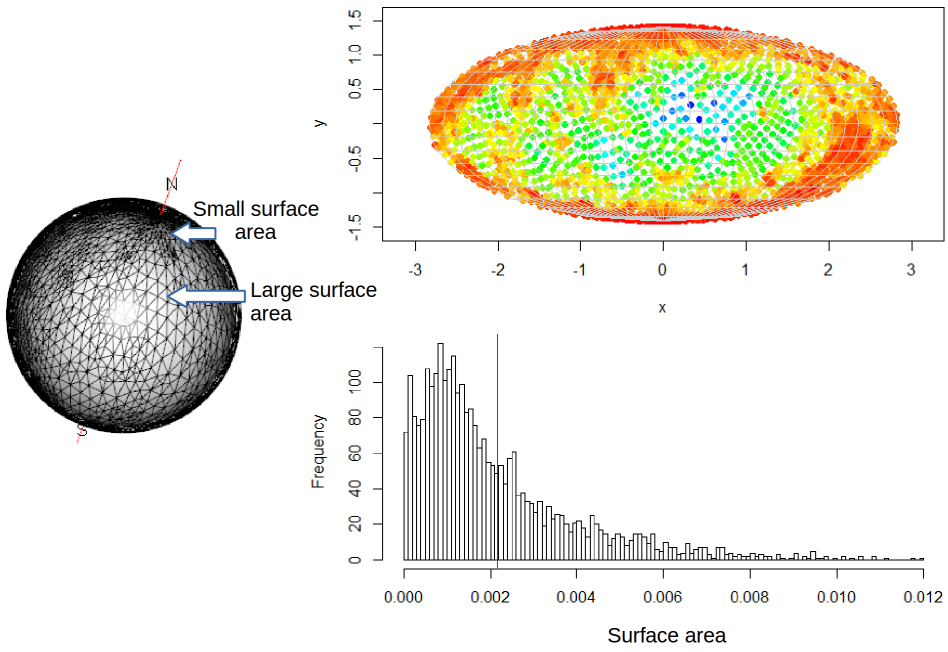


Figure 4.8: Spherical parameterization illustration done by converting the latitude and longitude angles into Cartesian coordinates using the set of equations 4.7(a,b,c). The surface areas created by the vertices are depicted on the surface of the sphere and they vary greatly. A histogram of the surface areas shows the uneven surface area distribution; the mean value is depicted by the red vertical line. The vertices of the mesh are plotted on a Mollweide map on the top right corner with the color displaying the surface area; red means small surface area up to the color blue meaning the largest surface area.

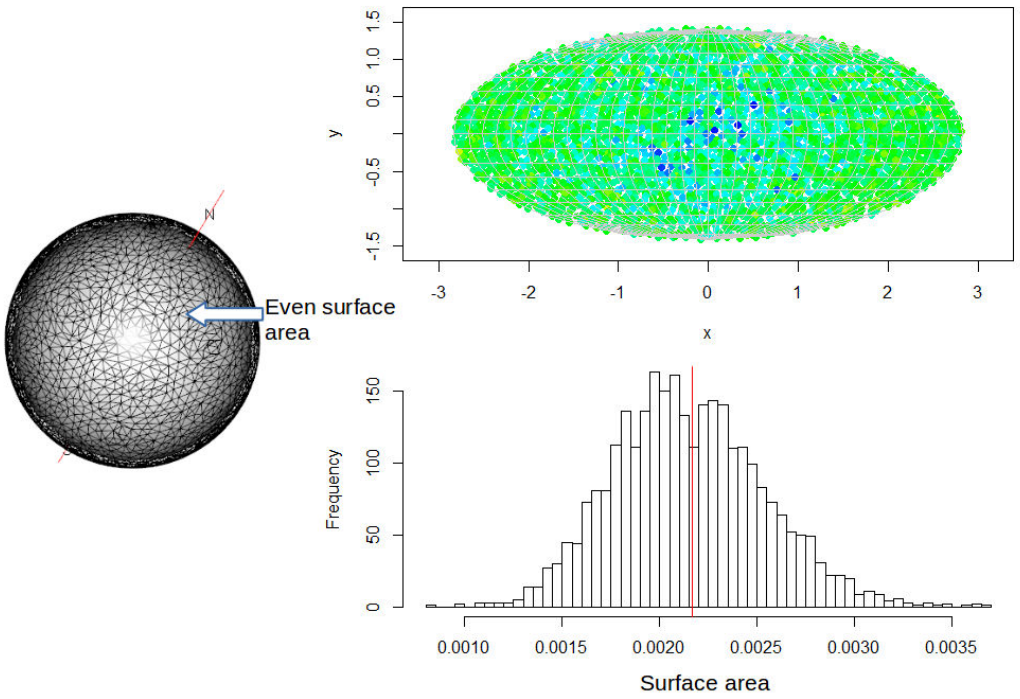


Figure 4.9: Illustration of the optimized spherical parameterization. The surface areas created by the vertices are depicted on the surface of the sphere which display even areas. A histogram of the surface areas shows the even surface area distribution around the mean depicted by the red vertical line. The vertices of the mesh are plotted on a Mollweide map on the top right corner with the color associated to the surface area in which the green color clearly shows the even distribution of the areas.

4.3 Projection results

Mollweide projections first example
The optimized Mollweide projections of an expanded surface corresponding to the intensity values of the MRIs and the CT, for a case where the recurrence is strongly linked to the medical images, are shown in the figure 4.10. Certain locations of the expanded surface correspond to locations where the tumor reappear after the main treatment. A binary Mollweide projection is presented in sub-figure (a) in which the location of recurrence is represented by the salmon color and the non-recurrence location is represented by the turquoise bluish color. The importance of this projection is that it clearly depicts the location of the recurrence.

The Mollweide map using the DW-MRI image is depicted in sub-figure (b). The color range specifies the ADC values, red representing low values where blue represents higher values. Overprinted latitude lines every 20 degrees and longitude lines every 30 degrees are shown on the Mollweide maps. Mollweide maps were also obtained in a very similar manner for the T2-Flair MRI, T1-Gd MRI, and the CT scan. The Mollweide map for the T2-Flair image shows a much more homogeneous distribution of values with a cold spot (represented by red) on the upper part of sub-figure (c). The following two Mollweide projections, the T1-Gd in sub-figure (d) and the CT-scan in (e), are interesting because they show a very prominent clear spot of low values at the very center corresponding to the recurrence location.

Mollweide projections second example
We have seen some Mollweide projections in which they clearly depict a change of intensity values in locations corresponding to the recurrence. However, it is often not the case. In this second example Mollweide projections, in which it is not possible to detect a clear link by a visual analysis, are depicted in figure 4.11. Only the DW-MRI projection in sub-figure (b) displays lower intensity values (red) corresponding to the recurrence region. However, the multiple regions with low intensity values diminishes the possible prediction capability of the image.

Mollweide projections third example
The third and last example of Mollweide projections are shown in figure 4.12. The projections do not seem to be able to show any difference in intensity values corresponding to the recurrence. The recurrence seem to be covering a great portion of the surface of the expanded tumor since most of the 2D map in sub-figure (a) is depicted in the salmon color corresponding to the

recurrence. Perhaps the fact that the recurrence is so big but not localized in a single spot makes it more difficult for a clear change in intensities to take place.

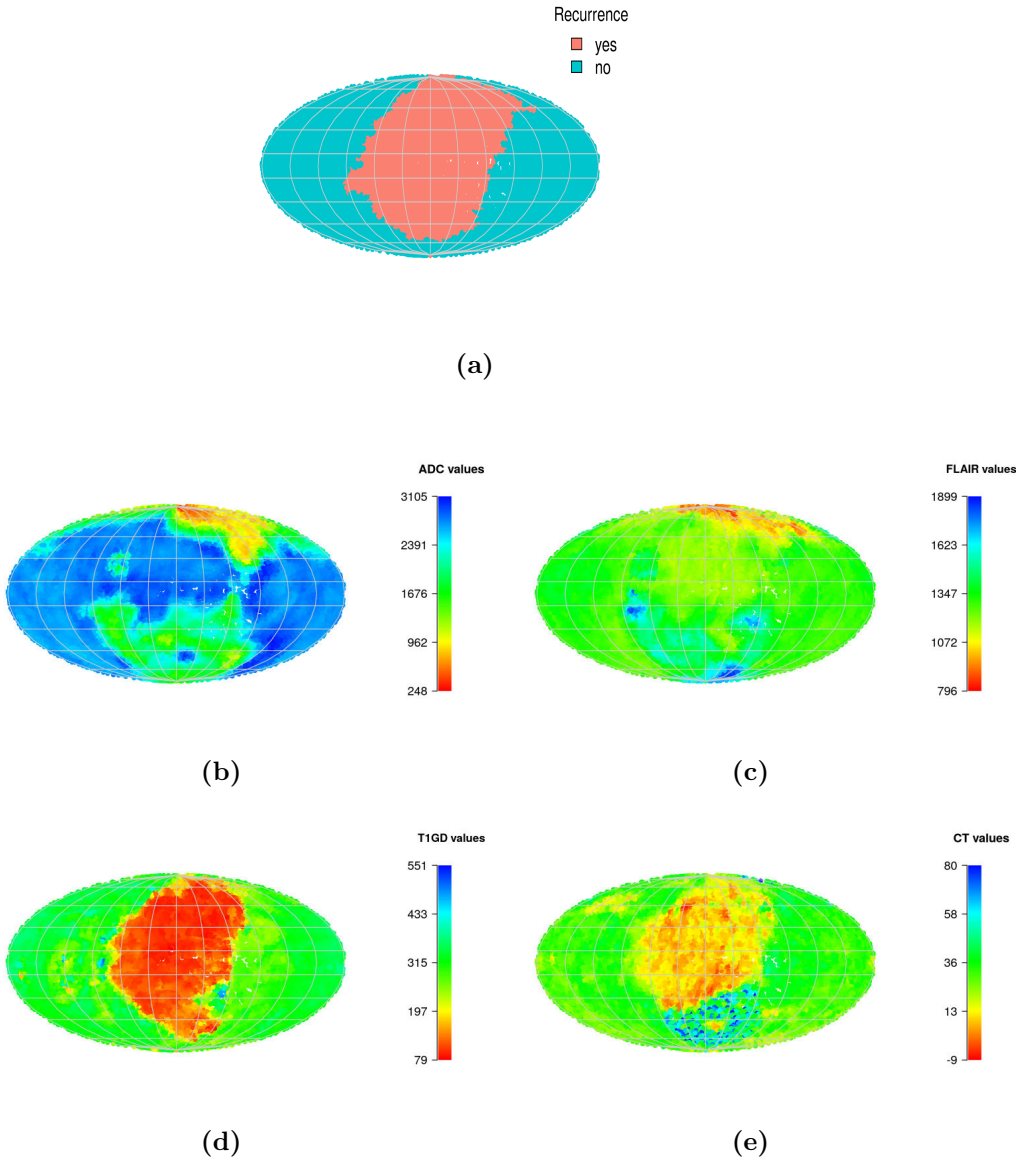
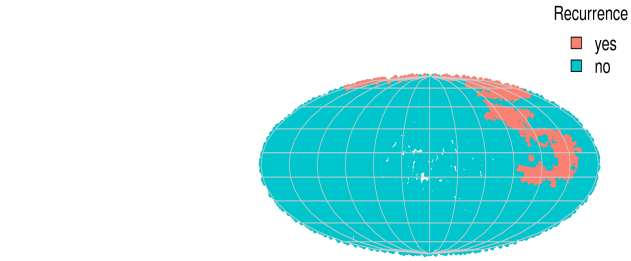
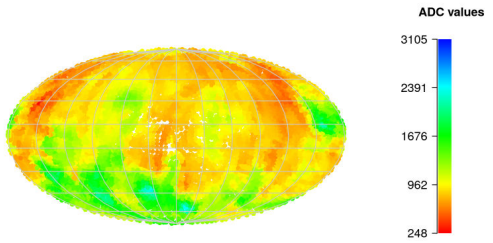


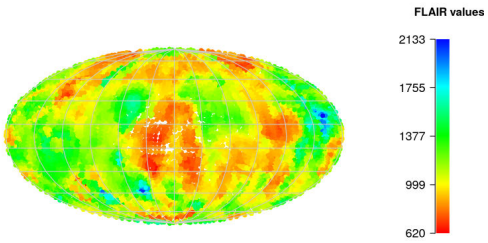
Figure 4.10: Mollweide projections of the surface of expanded tumor (GTV + 2 mm) structure of example one: (a) for the recurrence versus non-recurrence locations in which the recurrence positions are displayed in salmon color and non-recurrence positions in turquoise, (b) using the ADC values of the DW-MRI image, (c) the T2-Flair MRI image values, (d) T1-Gd MRI values, (e) CT scan values. Projections (d) and (e) show a clear difference in pixel values corresponding to the recurrence location.



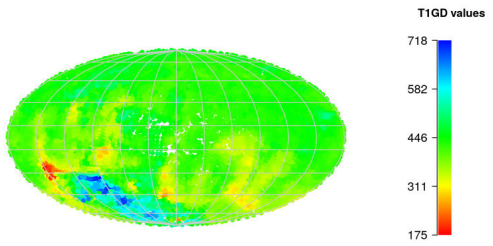
(a)



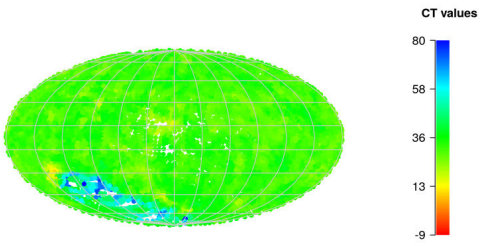
(b)



(c)



(d)



(e)

Figure 4.11: Mollweide projections of the surface of expanded tumor (GTV + 2 mm) structure of example two: (a) for the recurrence versus non-recurrence locations in which the recurrence positions are displayed in salmon color and non-recurrence positions in turquoise, (b) using the ADC values of the DW-MRI image, (c) the T2-Flair MRI image values, (d) T1-Gd MRI values, (e) CT-scan values. Projections do not show a change in pixel intensity values corresponding to the recurrence location with the exception of a lower intensity spot on the DW-MRI projection but its relevance is diminished because there are too many other low intensity spots.

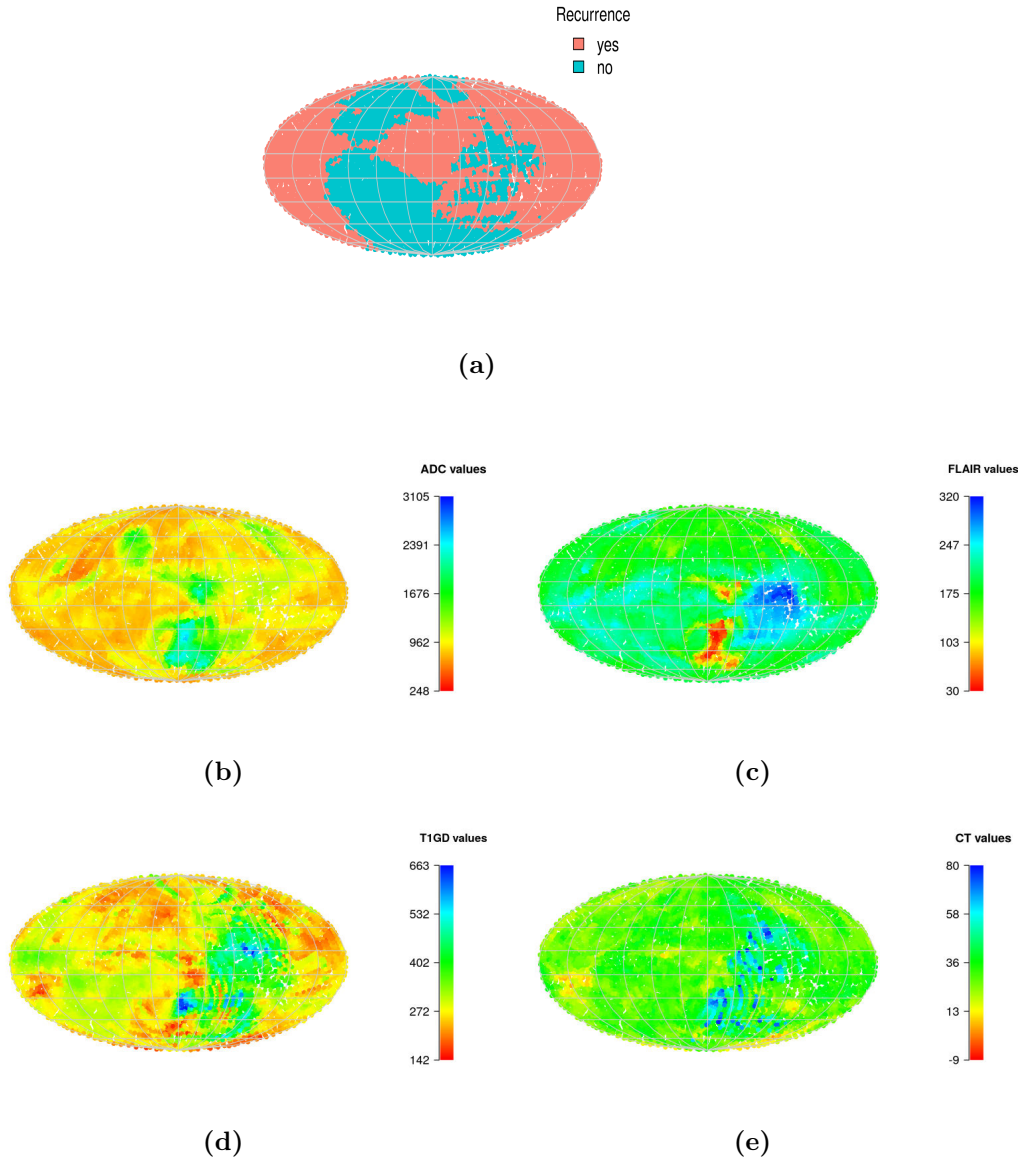


Figure 4.12: Mollweide projections of the surface of expanded tumor (GTV + 2 mm) structure of example three: (a) for the recurrence versus non-recurrence locations in which the recurrence positions are displayed in salmon color and non-recurrence positions in turquoise, (b) using the ADC values of the DW-MRI image, (c) the T2-Flair MRI image values, (d) T1-Gd MRI values, (e) CT-scan values. None of the projections display any pattern, of the intensity values, corresponding to the recurrence location.

4.4 Conclusions

The 2D mapping of the surface of a structure of interest such as the tumor or the expanded tumor provides a valuable visual representation tool especially to quickly identify patterns in the intensity values concerning the medical images. The visualization representation tools developed allowed for the creation of the Mollweide projections which were used for the analysis of recurrence link to the intensity values of the images. While some of the 2D maps depicted interesting correlations between the recurrence and the intensity values, it does not completely provide a definitive answer concerning the link between intensity values and recurrence location. However, it can be concluded that occasionally some medical images show a tendency to display a discrepancy of pixel values where the tumor will reappear.

The developed tools are not restricted for cancer related purposes since they can be applied for other purposes as well, as long as sufficient medical data is available, such as the analysis of well defined organ structures which can be obtained by contouring techniques using the medical images. One possible application could be for neurologic analysis studies. We can expect an increase in the usage of visual representation tools, similar to those presented in this work, which will gain more attention in the medical field since the amount and quality of medical images keep increasing.

Conclusions

The three main sections composing this work were the Bayesian framework for modeling clinical data, the tumor recurrence analysis, and the development of structure mapping visual representation tools. The data of Glioblastoma brain cancer was used for developing prediction models such as neurologic grade and tumor recurrence predictions. The methodology for modeling oncology data for the ambition of directing towards personalized medicine was emphasized.

Bayesian framework for modeling clinical data
The Bayesian framework developed emphasized three useful inference objectives of the Bayes' theorem. The first objective was to predict a parameter of interest by means of finding its pdf. The pdf of the parameter allows for the prediction of the value of the parameter and its uncertainty. The pdf contains all the possible values of the parameter and the likeliness of each of those values. Often the parameter was needed to complete a prediction model which was used to make predictions for a new data set. The use of the newly identified pdf of the parameter for making predictions for a new data set is the second inference objective. The comparison of multiple models is the third and last Bayesian inference objective. The comparison of models allows for identifying the best predicting model.

Bayes' theorem is central for developing clinical based models. The theorem works as a bridge between the unknown parameters of interests and

the observed data. In simple words Bayes' theorem states that in order to obtain a parameter given some observed data we need to multiply the likelihood function created by the observations, times the previous belief of the parameter of interest. The model development was exemplified with practical examples. The advantage of using conjugate prior relations was emphasized as well. The strong solving capabilities of the numerical approach was highlighted. The numerical approach of the Bayesian framework allows for a flexible and robust methodology for the development of a broad range of models. For instance, neurologic prediction models were developed, first a simpler model was constructed then a more complex one was added which clearly demonstrates the flexibility of the framework which allows to keep increasing the complexity of the models.

The neurologic simple versus complex model highlighted the drastic change in prediction when the additional parameter of CTV size was added to the prediction model; initially the simple model only predicted using the neurologic grade before the treatment. The grade prediction greatly depended on the CTV; the greater the CTV size the greater the probability of developing a higher neurologic grade after the main treatment.

Tumor recurrence analysis
A reduced form of the Bayesian approach, the generalized linear model (GLM), was used to analyze the possibility of predicting the recurrence location based on medical imaging. The first step was to expand the tumor in layers of 2 mm at a time; the layers that had a more profound impact in the predictions were the first two layers. The intensity values pertaining to the recurrence and non-recurrence locations, inside the layers, were identified and the intensities of the medical images at those locations were selected. The predictive models used those pixel intensity values.

The medical imaging included the DW, T2-Flair, and T1-Gd MRI sequences and the CT-scans. The models predict the location of the recurrence given a certain intensity value corresponding to a voxel. First a complete model was created which included all the MRI sequences and the CT-scan as well as all their possible combinations. For instance one variable was related to DW-MRI, another to T2-Flair, and there were also the variables corresponding to the possible combinations such as CT×DW×FLAIR. The complete model stated that for the most part the single variables (e.g., CT) were more important for the prediction than the composite variables such as CT×DW×FLAIR. After eliminating the less important variables we proceeded to construct the reduced model which makes a slightly better predic-

tion.

We used Machine Learning tools in the form of decision trees in addition to the GLM models with the objective of making a prediction based on discriminatory case analysis instead of a linear response as in the link function of the GLM. The discriminatory cases are depicted by leaves of the decision trees in which the path of a branch must be followed to arrive at a particular leaf. The decision tree model improved the recurrence prediction compared to the GLM models. However, the improvements were not sufficient to obtain an indisputable accurate recurrence prediction model. The accuracy of the predictions were corroborated by using the Receiver Operating Characteristic (ROC) spaces in which the correct and incorrect recurrence predictions were easily depicted. Additionally, prediction maps were shown on a single MRI slice.

Structure mapping visual representation tools
The development of a new method of visualization analysis was desired because of the lack of reliable predictions of the GLM and tree models, as well as the limitation of the prediction maps. This new method was the structure mapping and visual representation tools. The method consisted of projecting the expanded tumor structure onto a unit sphere and subsequently onto a 2D display in the form of the Mollweide projection.

The first step was to choose a suitable tumor structure with no holes. Then a triangular mesh based on the Marching cubes algorithm was created. With the aid of an ingenious method developed by Brechbühler et al., we were able to calculate a latitude and longitude angle for each vertex of the mesh. These polar angles were used for the unit sphere mapping and the 2D plotting. The unit sphere mapping revealed uneven surface areas created by the vertices. Therefore an optimization method was developed with the purpose of creating even surface areas. The newly created optimized vertices were used to create the Mollweide maps.

The Mollweide maps depicted the pixel intensity values of the surface of the expanded structure. We used the pixel values corresponding to DW, T2-Flair, and T1-Gd MRI sequences and the CT-scans to create the Mollweide maps. For some cases, the 2D maps revealed a change in pixel intensity corresponding to the tumor recurrence location. The change is more evident for the T1-Gd, and CT-scan and partially for the T2-Flair. However, a clear recurrence pixel intensity pattern is not often the case. Hence, making a conclusion of a definitive strong link between pixel intensity and recurrence location is not currently possible. In spite of that, we did find a link.

The inability of a solid recurrence prediction is unlikely related to the quality of the modeling instead we believe it is do to external reasons. Several possibilities exist for the lack of accuracy in the predictions. The first strong possibility is that additional intermediate imaging is necessary for the construction of the models not just the pre and post treatment imaging which it is currently the case. A more precise tumor recurrence evolution could be reconstructed with such additional imaging. Another possibility, is that the current MRI resolution is not sufficient. It is likely that the resolution of medical imaging keeps improving and in that case similar GLMs and decision trees models could be used to test the resolution effect. One other strong possibility exists which is that medical imaging alone is not sufficient to predict recurrence location. A personal opinion is that there are missing pieces of information, hence the current medical imaging alone would not suffice.

Prespectives
The generic methodology developed in this work provides a foundation for a wide range of clinically based models. Different models could quickly be built for other parameters of interest. The modeling is not limited to applications to Glioblastoma. Other types of tumors could be analyzed as well. In fact, the methodology could even be used for a wider range of modeling applications even for non-cancer related applications such as finance. For instance, here we were interested in finding neurology toxicity parameters but in finance we might be interested in finding investment risk parameters. An important strength of the developed methodology is such flexibility which we can see by the range of possible applications.

We expect the usage of 3D reconstructed surfaces to keep increasing for a wide variety of purposes which includes applications for diagnostics or even for the training of medical professionals. Along the same lines, the 2D Mollweide mapping could potentially be used for the analysis of others diseases or neurologic studies. It can be used for other applications that also involve the pixel intensity variations around the surface of specific structures.

Lastly, this work serves as a guide for the development of clinically based models. A project called Plateforme de Modélisation pour la Radiothérapie (PMRT) is currently being developed, in collaboration with multiple institutes, at the Laboratoire de Physique Corpusculaire (LPC) at the city of Caen in Normandy France. The project consists of creating a modeling platform for the purpose of developing models using oncology clinical data from several clinics and hospitals. Big data modeling, such as the PMRT

project, could greatly benefit the personalized medicine approach. In order to improve personalized medicine, in France, a straightforward but privacy friendly policy for accessing oncology clinical data for research purposes must be implemented.

Résumé détaillé

Introduction

L'amélioration constante des traitements oncologiques a conduit à une augmentation significative des données médicales sous forme de dossiers électroniques. L'augmentation de la puissance de calcul et la nécessité d'adopter une approche personnalisée pour améliorer les soins aux patients mènent à l'élaboration de modèles prédictifs fondés sur des données cliniques. La médecine s'evolue vers une médecine personnalisée [8]. Un traitement personnalisé est particulièrement intéressant, car la réponse biologique de chaque patient peut varier considérablement malgré l'administration de traitements identiques. Par conséquent, l'adaptation des traitements pourrait améliorer considérablement les résultats. Des modèles prédictifs sont en cours d'élaboration pour intégrer les quantités considérables de données oncologiques [5]. Le potentiel de l'utilisation des techniques d'apprentissage machine sur des données volumineuses est une voie prometteuse vers la personnalisation des soins aux patients. De telles techniques peuvent être utilisées pour aider les cliniciens à prendre des décisions plus éclairées en matière de traitement en se fondant sur les données cliniques antérieures concernant les patients [51]. Les approches bayésiennes peuvent être utilisées pour apprendre à partir des données cliniques d'une manière continue permettant l'acceptation de l'incorporation de nouvelles données menant à une approche d'apprentissage continu.

L'objectif principal est de développer des modèles biophysiques pour prédire les effets cliniques en trouvant des modèles à partir de dossiers cliniques déjà documentés de patients, y compris des images médicales. En développant ces modèles de prédiction, nous pouvons trouver des paramètres qui jouent un rôle important dans les résultats cliniques. La mise en lumière de ces paramètres importants cachés conduit à une approche personnalisée.

La base de données Glioblastomes

Les données cliniques utilisées dans ce travail ont été obtenues du centre de lutte contre le cancer "Centre François Baclesse" situé en Normandie, France. La base de données comprend les dossiers cliniques recueillis tout au long de l'histoire médicale d'environ 90 patients souffrant d'un glioblastome. Le glioblastome est un type de cancer du cerveau très agressif qui résiste généralement aux traitements, y compris la chimiothérapie et la radiothérapie. Le glioblastome affecte les patients à différents âges, mais touche surtout les patients plus âgés. Les patients présentent souvent une surexpression du R-EGF, des mutations du PTEN (MMAC1) [28] . Les glioblastomes sont l'un des cancers les plus vascularisés et les plus invasifs, les pronostics ne se sont pas améliorés depuis des décennies [1]. Il reste encore beaucoup de travail à faire pour améliorer les pronostics à partir de la compréhension des mutations génétiques qui y sont associées, ainsi que la détection très précoce. La classification des tumeurs cérébrales va du grade I au grade IV, ce dernier étant le plus agressif [27] . Un glioblastome multiforme est un cancer de grade IV, c'est donc un cancer très agressif. Dans "la classification des tumeurs du système nerveux central de l'Organisation mondiale de la santé de 2016", le rapport mentionne que la classification utilise désormais des paramètres moléculaires en plus de l'histologie [41].

Pour cette étude, une quantité fixe de données a été acquise, mais un objectif à long terme serait de poursuivre l'ajout des données à la base de données en tant qu'étude continue plutôt qu'étude rétrospective. Par conséquent, à long terme, des dossiers cliniques supplémentaires pourraient être ajoutés. Les données recueillies sont divisées en deux types, les entrées de données et les images médicales. Les entrées de données se réfèrent aux champs qui sont enregistrés et entrés dans l'ordinateur comme les symptômes et le sexe du patient, les entrées de données globales se réfèrent à la plupart des autres champs qui ne sont pas des images médicales. Les entrées de données peuvent être quantitatives ou qualitatives et se présenter sous différents formats. Le sexe est masculin ou féminin, l'âge est un nombre, les complications peuvent être écrites dans des échelles de grades. Certaines des entrées de données peuvent également être vraies ou fausses, par exemple si une chirurgie a été pratiquée ou non. Le type d'ablation chirurgicale a été enregistré : soit aucune extraction chirurgicale ou simple biopsie, soit exérèse partielle ou complète de la tumeur. La base de données contient également plusieurs IRM et images de tomodensitométrie différentes, la dosimétrie 3D appliquée et les contourages opérés par le radiothérapeute oncologue. La base de données contient plusieurs séquences d'IRM différentes, y compris les

séquences classiques T1-Gd, T2-Flair, et la séquence de diffusion DW-MRI. Les séquences d'IRM se réfèrent au type spécifique d'image d'IRM réalisée et leur signification spécifique est expliquée plus loin dans le troisième chapitre.

Même si les antécédents médicaux des patients dans notre base de données peuvent varier selon les individus, nous pouvons en illustrer une chronologie typique. La représentation chronologique des antécédents médicaux est illustrée à la figure 1.2. Généralement, le patient passe par une consultation médicale au cours de laquelle les données sont enregistrées, par exemple les symptômes et l'âge sont notés. Des examens médicaux sont effectués, pouvant inclure des examens d'imagerie médicale. Un diagnostic est fait à l'aide des examens médicaux et le patient est traité dans certains cas en commençant par la chirurgie, puis la radiothérapie est planifiée, suivie du le traitement en radiothérapie et en chimiothérapie. La base de données contient les tomodensitogrammes, la dosimétrie 3D, le contour des zones ciblées et des organes à risque, etc. Des complications et plusieurs effets secondaires sont également enregistrés. En raison de la nature agressive du glioblastome, un traitement de rattrapage peut être effectué.

Formalisme Bayésien pour la modélisation de données cliniques

La première partie porte sur l'élaboration d'un cadre bayésien pour la modélisation des données cliniques. Le chapitre commence par une explication complète et intuitive de l'approche bayésienne qui nous permet d'élaborer des modèles de prédiction fondés sur des données cliniques. L'un des principaux avantages de l'approche bayésienne est que l'incertitude des paramètres est accessible.

Une construction complète et intuitive du théorème de Bayes a été décrite à partir de concepts statistiques de base. Nous avons procédé à la construction de la probabilité en proposant un a priori et en déterminant les preuves. Proposer un a priori peut s'avérer difficile car il est difficile d'énoncer clairement ce en quoi on croyait auparavant au sujet d'un paramètre d'intérêt. Il a été démontré que la détermination de la distribution a posteriori des paramètres d'intérêt est essentielle pour les cas pratiques. Les trois principaux objectifs d'inférence accessibles au moyen du formalisme Bayésien sont présentés : l'estimation de paramètres habituellement en trouvant leur distribution a posteriori, la prédiction d'un nouvel ensemble de données en utilisant la distribution a posteriori des paramètres, la comparaison des modèles.

Plusieurs modèles, impliquant la base de données sur le glioblastome,

ont été développés en utilisant le cadre bayésien qui inclut des modèles de prédiction de grade neurologique. L'approche numérique pour résoudre le théorème de Bayes est soulignée car elle permet la modélisation d'un large éventail de situations ; le processus de résolution implique des techniques d'apprentissage machine. Des exemples réels de données cliniques de patients atteints de glioblatome ont été utilisés pour développer la méthodologie de construction du modèle. Le modèle de prédiction neurologique améliorée démontre la forte dépendance à la taille du CTV alors que le modèle simple prédit la même probabilité pour tout CTV puisqu'il ne tient pas compte de cette quantité. Le potentiel du calcul numérique, par exemple en utilisant l'algorithme de Metropolis-Hastings, en particulier pour les grands ensembles de données, a été souligné.

Ce travail n'a pas pour but de fournir une interprétation médicale des résultats, il est axé sur l'approche mathématique de la modélisation à partir de données cliniques réelles. Par exemple, l'utilisation de CTV dans le modèle amélioré au lieu de GTV était motivée par une meilleure corrélation avec le résultat observé. Cette observation suggère que le résultat est lié au traitement plutôt qu'au patient, mais ce n'est qu'une proposition, pas une démonstration. La méthodologie démontre le potentiel de la capacité prédictive des statistiques bayésiennes ainsi que l'approche de l'apprentissage machine. Cette approche peut être particulièrement utile pour les grandes bases de données qui s'orientent vers une approche prédictive personnalisée.

Prédiction de la zone de récidive de la tumeur

La deuxième partie applique un cas réduit du cadre bayésien dans lequel des modèles linéaires généralisés (GLM) ont été construits pour explorer les corrélations des images médicales de prétraitement sous forme de résonance magnétique (IRM) et de tomodensitométrie (CT) avec la récidive de la tumeur. Les séquences d'IRM utilisées sont DW-MRI, T2-Flair et T1-Gd. Puis, d'une manière similaire, des modèles plus complexes utilisant des arbres de décision, issus de techniques d'apprentissage machine, ont été utilisés pour révéler d'éventuelles corrélations cachées avec la récidive que les modèles GLM sont incapables de trouver. L'évaluation du modèle GLM et du modèle d'arbre est illustrée (figure I) à l'aide d'espaces ROC et comparée à certaines courbes ROC que l'on trouve dans la littérature [9]. Les résultats de nos espaces ROC nous permet d'évaluer l'exactitude des prédictions pour chaque patient. Alors que les courbes ROC dans la littérature ne donnent qu'un résultat global.

Les valeurs d'intensité des images correspondant aux endroits de la surface externe de la tumeur et son image miroir du côté opposé du cerveau ont été analysées et comparées. Les images miroirs ont également été utilisées à des fins de normalisation. Des couches (d'environ 2 mm) ont été créées en augmentant normalement autour de la tumeur. Une comparaison des valeurs d'intensité a été effectuée indépendamment pour chacune des différentes couches. Cela a été possible parce que la position des zones de récidive et de non-récurrence, à l'intérieur des couches, est connue à partir des données d'images médicales. La pertinence et les limites de l'étude sont également abordées.

Les résultats obtenus dans ce travail concordent avec une étude de thèse précédente [44], dans laquelle la récidive péritumorale montre des valeurs d'ADC (DW-MRI) comparativement plus basses qu'en dehors de la zone de récidive. Dans ce travail, les résultats montrent une différence faible mais statistiquement significative. Ces résultats sont comparés avec la littérature dans laquelle une étude très récente suggère que les valeurs ADC et FLAIR ont diminué respectivement de 9,5% ($p < 0{,}001$) et 9,2% ($p < 0{,}001$) dans les régions de récidive de l'œdème péritumoral par rapport aux régions péritumorales non récurrentes [9]. Cette étude suggère que l'utilisation d'un modèle logistique multi-paramétrique semble mieux prédire la région de récurrence qu'une seule valeur d'intensité. Dans notre travail, nous avons effectué des ajustements multiparamétriques (DW, T2-Flair, T1-Gd et valeurs d'intensité de CT) dans lesquelles les valeurs T1-Gd et CT semblent plus pertinentes pour la prédiction de la récidive. Cependant, après avoir ajusté plusieurs modèles multiparamétriques et arbres de décision, les modèles ne peuvent toujours pas fournir une réponse définitive. En particulier, si l'évolution que nous observons sur l'ADC est conforme à [9], nous trouvons une évolution inverse pour FLAIR. Cela nous a donc amenés à conclure que nos modèles sont incapables de prédire la récidive avec une certitude absolue en utilisant uniquement les données actuelles de l'IRM (avant et après le traitement). Il existe plusieurs possibilités. La différence d'intensité de la récidive est trop faible, ce qui la rend difficile à percevoir avec la résolution d'imagerie actuelle. Des IRM intermédiaires sont nécessaires pour l'évolution de la tumeur et de la récidive. Les différents calibrages des IRM des différentes cliniques créent une perturbation plus importante que prévu auparavant. Enfin il est possible que des informations supplémentaires manquent pour parvenir à une prédiction de la récidive.

Représentation en mapping sphérique de la surface tumorale

La troisième partie concerne le développement d'outils de cartographie sphérique et de représentation cartographique dérivés de la nécessité d'analyser visuellement le lien entre la récidive tumorale et les valeurs d'intensité de la récidive. Nous avons donc cherché ici à obtenir une représentation plane du volume tumoral sur laquelle les coordonnées correspondent à une position précise dans l'espace et l'intensité correspond à la valeur des images en ces coordonnées. L'illustration décrivant la cartographie d'une structure d'intérêt est illustrée sur la figure II. L'objectif est d'obtenir des cartes 2D pour analyser l'intensité des pixels correspondant au point de récurrence à trouver pour un motif. Les cartes de Mollweide montrent un modèle clair, mais c'est rarement le cas.

La première étape pour développer ces outils consiste à construire un maillage couvrant la structure de la tumeur expansée dans lequel chacun des sommets a des coordonnées cartésiennes connues. Un couple d'angles, latitude et longitude, sont associés à chaque sommet en suivant un algorithme développé par Brechbühler et al [6] qui est basé sur le concept de diffusion thermique. Avec les angles de latitude et de longitude, nous avons pu créer une carte sphérique de la tumeur étendue. Puis des cartes 2D, de la surface de la représentation sphérique, sous forme de projections de Mercator et de Mollweide ont été construites. Les cartes 2D permettent une analyse directe des valeurs d'intensité des lieux de récurrence et de non-récurrence. La cartographie 2D de la surface d'une structure d'intérêt telle que la tumeur ou son extension fournit un outil de représentation visuelle précieux, notamment pour identifier rapidement des modèles dans les valeurs d'intensité concernant les images médicales.

Les outils développés ne sont pas limités à des fins liées au cancer, ils peuvent également être utilisés à d'autres fins, pourvu que l'on dispose de données médicales suffisantes, comme l'analyse de structures d'organes bien définies qui peuvent être obtenues par des techniques de contournage utilisant les images médicales. Une application possible pourrait être pour des études d'analyse neurologique. On peut s'attendre à une utilisation plûs courant des outils de représentation visuelle, semblables à ceux présentés dans ce travail, qui attireront davantage l'attention dans le domaine médical puisque la quantité et la qualité des images médicales ne cessent de croître.

Perspectives

La méthodologie générique mise au point dans le cadre de ce travail jette les bases d'une vaste gamme de modèles cliniques. Différents modèles pourraient rapidement être construits pour d'autres paramètres d'intérêt. La modélisation ne se limite pas aux applications au glioblastome. D'autres types de tumeurs pourraient également être analysés.

D'un point de vue technique, la reconstruction 3D d'autres organes ou parties du corps est réalisée de la même manière que la reconstruction tumorale réalisée dans ce travail. Au cours des prochaines années, on s'attendrait à ce que l'utilisation de surfaces reconstruites en 3D à l'aide d'images médicales soit de plus en plus utilisée à des fins diverses. Dans le même ordre d'idées, la cartographie 2D de Mollweide pourrait potentiellement être utilisée pour l'analyse d'autres maladies ou des études neurologiques. Elle peut être utilisée pour d'autres applications qui impliquent également les variations d'intensité des pixels autour de la surface de structures spécifiques.

Enfin, ces travaux servent de guide pour le développement de modèles cliniques. Un projet appelé Plateforme de Modélisation pour la Radiothérapie (PMRT) est en cours d'élaboration, en collaboration avec plusieurs instituts, au Laboratoire de Physique Corpusculaire (LPC) de Caen. Le projet consiste à créer une plateforme de modélisation dans le but de développer des modèles utilisant des données en oncologie provenant de plusieurs cliniques et hôpitaux. La modélisation de grandes quantités de données, comme dans le projet PMRT, pourrait grandement profiter à l'approche de la médecine personnalisée. Afin d'améliorer la médecine personnalisée, en France, une politique simple mais respectueuse de la vie privée pour accéder aux données cliniques en oncologie à des fins de recherche doit être mise en place.

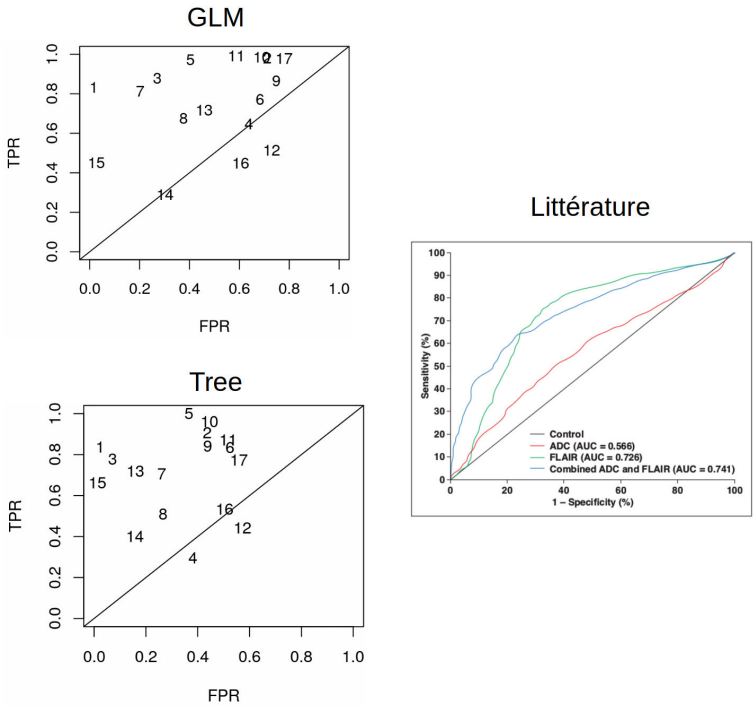


Figure I: Les espaces ROC (Reciever Operating Characteristic) pour le modèle GLM et l'arbre sont indiqués sur le côté gauche de la figure dans laquelle chaque nombre représente un patient. L'espace ROC de l'arbre présente un TFP inférieur à celui du modèle GLM. Sur le côté droit, une courbe ROC d'un modèle de la littérature est montrée. Les courbes ROC de la littérature et nos résultats spatiaux ROC ne se contredisent pas. Les espaces ROC présentés montrent clairement que le modèle prédit un haut TPR (True Positive Rate) pour certains patients alors qu'il échoue de manière drastique pour d'autres. D'autre part, la courbe de ROC peignée présente une évaluation générale et ne permet pas de séparer les prédictions de ROC pour chaque patient.

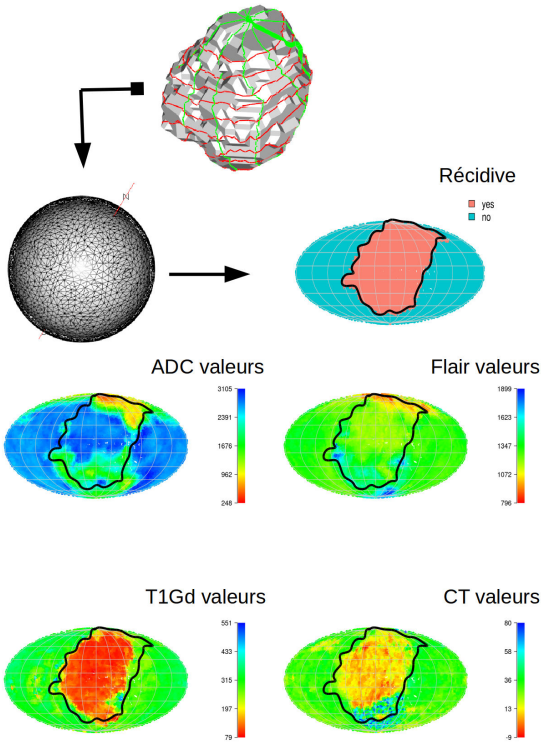


Figure II: Illustration décrivant la cartographie 2D de la surface d'une structure d'intérêt. Tout d'abord, une structure tumorale est normalement dilatée d'une épaisseur d'environ 2 mm vers l'extérieur. Les angles de latitude et de longitude sont obtenus pour les sommets d'un maillage. Les lignes de longitude et de latitude sont tracées sur la surface de la structure tumorale élargie. La surface de la structure est paramétrée de façon sphérique, puis une projection de Mollweide 2D montre l'emplacement de la récurrence. Les valeurs d'intensité, pour les valeurs ADC, FLAIR, T1Gd et CT correspondant au lieu de récurrence sont représentées en couleurs. Un lien avec le domaine est démontré, mais c'est rarement le cas.

Mathematical expressions

Bernoulli's distribution

$$P(x) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0 \end{cases} \tag{A.1}$$

this can also be written in more compact manner,

$$P(x) = \theta^x (1 - \theta)^{1-x} \tag{A.2}$$

Where x represents a data value $x = 0$ represents failure, $x = 1$ represents success and θ represents the probability of a success.

Binomial Distribution,

$$\begin{aligned} P(n|N) &= \binom{N}{n} \theta^n (1 - \theta)^{N-n} \\ &= \frac{N!}{n!(N-n)!} \theta^n (1 - \theta)^{N-n} \end{aligned} \tag{A.3}$$

N represents the total number of trials, and n represents the number of success events and θ represents the probability of a success.

The Beta distribution,

$$P(\theta) = \frac{(1 - \theta)^{\beta-1} \theta^{\alpha-1}}{B(\alpha, \beta)} \tag{A.4}$$

Where $B(\alpha, \beta)$ is the beta function serving to normalize,

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

and the Beta can also be express using the gamma function that is;

$$\beta(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{A.5}$$

Proof of the Beta-Binomial conjugate prior
Using Bayes' theorem 2.3 Letting the likelihood function to be a binomial distribution and the prior a Beta distribution as well as recalling that the evidence function is to normalize to 1, then we get,

$$\begin{aligned}
P(\theta|D) &= \frac{P(D|\theta)P(\theta)}{P(D)} \\
&= \left\{\frac{1}{P(D)}\right\}\left\{\binom{N}{n}\theta^n(1-\theta)^{N-n}\right\}\left\{\frac{1}{\beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right\} \\
&= \left\{\frac{1}{P(D)}\binom{N}{n}\frac{1}{\beta(\alpha,\beta)}\right\}\left\{\theta^n(1-\theta)^{N-n}\right\}\left\{\theta^{\alpha-1}(1-\theta)^{\beta-1}\right\} \\
&= \big\{C\big\}\big\{\theta^n(1-\theta)^{N-n}\big\}\big\{\theta^{\alpha-1}(1-\theta)^{\beta-1}\big\} \\
&= C\,\theta^{n+\alpha-1}(1-\theta)^{N-n+\beta-1} \\
&= C\,\theta^{\alpha_1-1}(1-\theta)^{\beta_2-1}
\end{aligned} \tag{A.6}$$

in which $\alpha_1 = (n+\alpha)$ and $\beta_1 = (N-n+\beta)$ and C is only a normalizing factor even though it looks daunting! We know that to normalize a beta distribution we can use the inverse of the beta function (A.5). Therefore the posterior function is,

$$\begin{aligned}
P(\theta|D) &= C\,\theta^{\alpha_1-1}(1-\theta)^{\beta_2-1} \\
&= \frac{1}{\beta(\alpha_1,\beta_1)}\theta^{\alpha_1-1}(1-\theta)^{\beta_2-1}
\end{aligned} \tag{A.7}$$

Therefore we have proofed that the posterior function is also a beta function!

Diagnostics of Monte Carlo Markov Chains (MCMC)
First, let us understand that a MCMC was created by stipulating a certain number of iterations and a certain step size. The purpose of the values of the MCMC, in this work, is to cover all the possible values of the parameter of interest (θ). A tool to diagnose MCMC is the Auto-Correlation Function (ACF) which inspects the correlation of the chain with itself. Figure 2.12 shows the ACF for each of the corresponding MCMC.

For instance let us analyze AC θ_2. The original MCMC is displayed one unit (+1) or lag one, creating a new chain of lag one. Similarly many (200) displacements are created. The ACF for the first couple of lags is 1 which means it is 100% correlated to itself which is logical since the lag one chain is almost identical to the original one. On the other hand as the lag increases the ACF drastically decreases, which indicates the chain is no longer correlated. For example, it can be seen that the ACF is nearly zero when the lag is about 100 for AC θ_2; this result of the ACF help us diagnose the MCMC. Such result can be interpreted as, the first iteration of the MCMC is no longer correlated at all to the 100 iteration. The faster the ACF descent the better. It is important because if the ACF value does not descent fast enough we would need to consider increase the simulation step size. Overall, the diagnostics tool of the ACF tells us that the obtained MCMC are acceptable simulations.

Bibliography

[1] Tercia Rodrigues Alves, Flavia Regina Souza Lima, Suzana Assad Kahn, Denise Lobo, Luiz Gustavo Feijó Dubois, Rossana Soletti, Helena Borges, and Vivaldo Moura Neto. Glioblastoma cells: A heterogeneous and fatal tumor interacting with the parenchyma. *Life Sciences*, 89(15-16):532–539, October 2011.

[2] Roland Bammer. Basic principles of diffusion-weighted imaging. *European journal of radiology*, 45(3):169–184, 2003.

[3] Stewart Bates. Progress towards personalized medicine. *Drug Discovery Today*, 15(3-4):115–120, February 2010.

[4] Stefan Bauer, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes. A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine and Biology*, 58(13):R97–R129, July 2013.

[5] Jean-Emmanuel Bibault, Philippe Giraud, and Anita Burgun. Big Data and machine learning in radiation oncology: State of the art and future prospects. *Cancer Letters*, 382(1):110–117, November 2016.

[6] Ch. Brechbuhler, G. Gerig, and O. Kubler. Parametrization of closed surfaces for 3-d shape description. *Computer Vision and Image Understanding*, 61(2):154 – 170, 1995.

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

[8] Christian Castaneda, Kip Nalley, Ciaran Mannion, Pritish Bhattacharyya, Patrick Blake, Andrew Pecora, Andre Goy, and K Stephen Suh. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, 5(1), December 2015.

[9] Peter D. Chang, Daniel S. Chow, Peter H. Yang, Christopher G. Filippi, and Angela Lignelli. Predicting Glioblastoma Recurrence by Early Changes in the Apparent Diffusion Coefficient Value and Signal Intensity on FLAIR Images. *American Journal of Roentgenology*, 208(1):57–65, January 2017.

[10] Chengshui Chen, Mingyan He, Yichun Zhu, Lin Shi, and Xiangdong Wang. Five critical elements to ensure the precision medicine. *Cancer and Metastasis Reviews*, 34(2):313–318, June 2015.

[11] Michael R. Chernick and Robert A. LaBudde. *An introduction to bootstrap methods with applications to R*. Wiley, Hoboken, NJ, 2011. OCLC: 751021626.

[12] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[13] Joseph A. Cruz and David S. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, 2:117693510600200, January 2006.

[14] Mary F. Dempsey, Barrie R. Condon, and Donald M. Hadley. Measurement of tumor "size" in recurrent malignant glioma: 1d, 2d, or 3d? *American Journal of Neuroradiology*, 26(4):770–776, 2005.

[15] Issam El Naqa. Perspectives on making big data analytics work for oncology. *Methods*, 111:32–44, December 2016.

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[17] Andrew Gelman. Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics. *Statistical Science*, 24(2):176–178, May 2009.

[18] Michael Goitein. The utility of computed tomography in radiation therapy: An estimate of outcome. *International Journal of Radiation Oncology• Biology• Physics*, 5(10):1799–1807, 1979.

[19] Joseph V. Hajnal and et al. Use of Fluid Attenuated Inversion Recovery (FLAIR) Pulse Sequences in MRI of the Brain. *Journal of Computer Assisted Tomography*, 16(6):841–844, December 1992.

[20] John W. Henson, Paola Gaviani, and R. Gilberto Gonzalez. MRI in treatment of adult gliomas. *The lancet oncology*, 6(3):167–175, 2005.

[21] Daniel D. Hoggarth, editor. *Stock assessment for fishery management: a framework guide to the stock assessment tools of the Fisheries Management Science Programme*. Number 487 in FAO fisheries technical paper. Food and Agriculture Organization of the United Nations, Rome, 2006. OCLC: 255237023.

[22] Ali Işın, Cem Direkoğlu, and Melike Şah. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Computer Science*, 102:317–324, 2016.

[23] Jose-Miguel Bernardo James O. Berger. On the development of the reference prior method. Technical Report 91-15C, April 1991.

[24] Nihar Jana, Anirban Basu, and Prakash Narain Tandon, editors. *Inflammation: the Common Link in Brain Pathologies*. Springer Singapore, Singapore, 2016.

[25] Ruirui Jiang, Hongbin Zhu, Wei Zeng, Xiaokang Yu, Yi Fan, Xianfeng Gu, and Zhengrong Liang. Bladder wall flattening with conformal mapping for MR cystography. page 76250E, San Diego, California, USA, March 2010.

[26] Nida S. Khan, Asma S. Larik, Quratulain Rajput, and Sajjad Haider. A BAYESIAN APPROACH FOR SUSPICIOUS FINANCIAL ACTIVITY REPORTING. *International Journal of Computers and Applications*, 35(4), 2013.

[27] Paul Kleihues, Peter C. Burger, and Bernd W. Scheithauer. The new WHO classification of brain tumours. *Brain Pathology*, 3(3):255–268.

[28] Paul Kleihues and Hiroko Ohgaki. Primary and secondary glioblastomas: From concept to clinical diagnosis. *Neuro-Oncology*, 1(1):44–51, 1999.

[29] Kinuko Kono, Yuichi Inoue, Keiko Nakayama, Miyuki Shakudo, Michiharu Morino, Kenji Ohata, Kenichi Wakasa, and Ryusaku Yamada. The role of diffusion-weighted imaging in patients with brain tumors. *American Journal of Neuroradiology*, 22(6):1081–1088, 2001.

[30] Samuel Kotz, editor. *Breakthroughs in statistics. 2: Methodology and distribution*. Springer series in statistics Perspectives in statistics. Springer, New York Berlin, 3. print. edition, 1993. OCLC: 311789218.

[31] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[32] K. Krabbe, P. Gideon, P. Wagn, Ulla Hansen, C. Thomsen, and F. Madsen. MR diffusion imaging of human intracranial tumours. *Neuroradiology*, 39(7):483–489, 1997.

[33] Frithjof Kruggel. Robust Mapping of Brain Surface Meshes onto a Unit Sphere. In *ISBI*, pages 201–204. Citeseer, 2007.

[34] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, 1st edition, 2010.

[35] D. Le Bihan. Diffusion/perfusion MR imaging of the brain: from structure to function. *Radiology*, 177(2):328–329, 1990.

[36] Denis Le Bihan and Mami Iima. Diffusion Magnetic Resonance Imaging: What Water Tells Us about Biological Tissues. *PLOS Biology*, 13(7):e1002203, July 2015.

[37] Eva Lu Lee and Laurel Westcarth. Neurotoxicity associated with cancer therapy. *Journal of the advanced practitioner in oncology*, 3(1):11, 2012.

[38] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

[39] Xiuwen Liu, John Bowers, and Washington Mio. Parametrization, alignment and shape of spherical surfaces. In *VISAPP (1)*, pages 199–206, 2007.

[40] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques - SIGGRAPH '87*, pages 163–169, Not Known, 1987. ACM Press.

[41] David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, 131(6):803–820, June 2016.

[42] Tim Lustberg, Johan van Soest, Arthur Jochems, Timo Deist, Yvonka van Wijk, Sean Walsh, Philippe Lambin, and Andre Dekker. Big Data in radiation therapy: challenges and opportunities. *The British Journal of Radiology*, 90(1069):20160689, January 2017.

[43] Philip Mayles, Alan Nahum, and Jean-claude Rosenwald. Handbook of Radiotherapy Physics: Theory and Practice. In *Handbook of Radiotherapy Physics: Theory and Practice*, pages 649–653. Taylor & Francis, United States of America, 2007.

[44] Emmanuel Meyer. *Evaluation des facteurs prédictifs clinico-radiologiques de récidive des glioblastomes : étude préalable à la constitution d'une plateforme de modélisation en radiothérapie.* Thesis in medicine, UNIVERSITÉ de CAEN., Caen, France.

[45] Travis B. Murdoch and Allan S. Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

[46] M. Nakahara, K. Ericson, and B. M. Bellander. DIFFUSION-WEIGHTED MR AND APPARENT DIFFUSION COEFFICIENT IN THE EVALUATION OF SEVERE BRAIN INJURY. *Acta Radiologica*, 42(4):365–369, 2001.

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

[48] Sahand Jamal Rahi and Kim Sharp. Mapping complicated surfaces onto a sphere. *International Journal of Computational Geometry & Applications*, 17(04):305–329, 2007.

[49] Christian P. Robert, Nicolas Chopin, and Judith Rousseau. Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, 24(2):141–172, May 2009.

[50] H Roux and J Lavieille. Imagerie par résonance magnétique nucléaire et rhumatologie. pages 8–33. Laboratoires Ciba-Geigy.

[51] M. Berkan Sesen, Ann E. Nicholson, Rene Banares-Alcantara, Timor Kadir, and Michael Brady. Bayesian Networks for Clinical Decision Support in Lung Cancer Care. *PLoS ONE*, 8(12):e82349, December 2013.

[52] John P. Snyder. The space oblique Mercator projection. *Photogramm. Eng. Remote Sensing*, 44:585–596, 1978.

[53] Paul E. Utgoff. Incremental induction of decision trees. *Machine learning*, 4(2):161–186, 1989.

[54] Theodore K. Yanagihara and T. J. Wang. Diffusion-weighted imaging of the brain for glioblastoma: Implications for radiation oncology. *Appl Radiat Oncol*, pages 5–13, 2014.

Bayesian statistics and modeling for the prediction of radiotherapy outcomes. An application to glioblastoma treatment

A Bayesian statistics framework was created in this thesis work for developing clinical based models in a continuous learning approach in which new data can be added. The objective of the models is to forecast radiation therapy effects based on clinical evidence. Machine learning concepts were used for solving the Bayesian framework. The models developed concern an aggressive brain cancer called glioblastoma. The medical data comprises a database of about 90 patients suffering glioblastoma; the database contains medical images and data entries such as age, gender, etc. Neurologic grade predictions models were constructed for illustrating the type of models that can be build with the methodology. Glioblastoma recurrence models, in the form of Generalized Linear Models (GLM) and decision tree models, were developed to explore the possibility of predicting the recurrence location using pre-radiation treatment imaging. Following, due to the lack of a sufficiently strong prediction obtained by the tree models, we decided to develop visual representation tools to directly observe the medical image intensity values concerning the recurrence and non-recurrence locations. Overall, the framework developed for modeling of radiation therapy clinical data provides a solid foundation for more complex models to be developed.

Key Words: Modeling, Bayesian statistics, Glioblastoma, Machine Learning, Tumor recurrence.

Utilisation des statistiques bayésiennes et de la modélisation pour la prédiction des effets de la radiothérapie. Application au traitement du glioblastome

Un cadre statistique bayésien a été créé dans le cadre de cette thèse pour le développement de modèles cliniques basés sur une approche d'apprentissage continu dans laquelle de nouvelles données peuvent être ajoutées. L'objectif des modèles est de prévoir les effets de la radiothérapie à partir de preuves cliniques. Des concepts d'apprentissage machine ont été utilisés pour résoudre le cadre bayésien. Les modèles développés concernent un cancer du cerveau agressif appelé glioblastome. Les données médicales comprennent une base de données d'environ 90 patients souffrant de glioblastome ; la base de données contient des images médicales et des entrées de données telles que l'âge, le sexe, etc. Des modèles de prévision neurologique ont été construits pour illustrer le type de modèles qui sont obtenus avec la méthodologie. Des modèles de récidive du glioblastome, sous la forme de modèles linéaires généralisés (GLM) et de modèles d'arbres de décision, ont été développés pour explorer la possibilité de prédire l'emplacement de la récidive à l'aide de l'imagerie préradiothérapie. Faute d'une prédiction suffisamment forte obtenue par les modèles arborescents, nous avons décidé de développer des outils de représentation visuelle. Ces outils permettent d'observer directement les valeurs d'intensité des images médicales concernant les lieux de récidive et de non-récurrence. Dans l'ensemble, le cadre élaboré pour la modélisation des données cliniques en radiothérapie fournit une base solide pour l'élaboration de modèles plus complexes.

Mots clés : Modélisation, Statistique bayésienne, Glioblastome, Apprentissage automatique, Récidive.