Development of clinically based prediction models using machine learning and Bayesian statistics

Oscar Daniel Zambrano Ramírez, Jean-Marc Fontbonne
Université Caen Normandie, Laboratoire de Physique Corpusculaire (LPC-Caen). France
8zambrano@gmail.com

Abstract

In this work, the framework for developing generic clinically based models is emphasized and illustrated with Bayesian statistics neurologic grade prediction models in order to exemplify the type of models that can be developed from a mathematical point of view. The models are based on clinical records of patients who underwent radiotherapy treatment due to glioblastoma which is an aggressive brain cancer. A first model requires as a parameter the neurologic grade of the patient before the treatment then predicts the grade after the treatment. A second, enhanced, model was developed with the aim of making the prediction more realistic and it uses the neurologic grade before the treatment as well, but it additionally depends on the Clinical Target Volume (CTV). Furthermore, with the aid of Bayesian statistic we were able to estimate the uncertainty of the predictions.

Key words: *learning; adaptive systems; statistics; clinical trials; prediction equations.*

Desarrollo de modelos de predicción basados clínicamente utilizando aprendizaje automático y estadísticas bayesianas.

Resumen

En este trabajo el marco teórico, para desarrollar modelos genéricos basados en datos clínicos, se enfatiza e ilustra con estadísticas bayesianas las cuales predicen grados neurológicos para ilustrar los tipos de modelos que se pueden desarrollar desde un punto de vista matemático. Los modelos se basan en datos clínicos de pacientes que se han sometido a radioterapia por causa de un glioblastoma, el cual es un cáncer de cerebro agresivo. Un primer modelo requiere como parámetro el grado neurológico del paciente antes del tratamiento y predice el grado después del tratamiento. Un segundo modelo, mejorado, fue desarrollado con el propósito de hacerlo más real, éste emplea también el grado neurológico antes del tratamiento; además depende del Volumen Blanco Clínico (CTV por sus siglas en inglés). Por último, con el uso de estadísticas bayesianas fue posible estimar la incertidumbre de las predicciones.

Palabras clave: *aprendizaje; sistemas adaptivos; estadística; ensayos clínicos; ecuaciones de predicción.*

Introduction

There has been an increase of cancer clinical data generation in the form of clinical records and imaging data. The rapid growth of clinical data is dramatically increasing due to the availability of electronic data. Hence, modelling for prognostics and therapeutic purposes is moving forward [1]. As a response, biophysical models based on clinical data mining and machine learning are increasingly being developed, with the aim of evaluating clinical effects of radiotherapy treatments. The rich oncology data is a well-known candidate to apply big data analytics in order to improve the cancer treatments [2]. Among common clinical data in oncology includes medical images and other records such as age, gender, grades, tumor size, just to name a few fields. Despite the rapid progress in machine learning and related techniques, there are still barriers for the implementation of machine learning models by clinicians. The barriers of understanding the complexity of machine learning methods by clinicians contributes to the slowdown of the implementation of the machine learning models [1]. Hence, in this work we proposed a nearly step by step guide to develop clinically based models for a wider audience extending beyond machine learning specialists.

Machine learning methods have been used to predict toxicity grades concerning gastro-intestinal and genito-urinary toxicities [3]. However, we were interested in applying machine learning methods to Glioblastoma tumors due to their very high aggressiveness and speed of evolution. We decided to use the well known Bayesian statistics to accomplish the machine learning process due to its useful evidence-based framework which helps move forwards towards a personalized medicine. Personalized patient care is increasingly becoming a trend [4, 5]. Bayes' theorem can be used as a mathematical tool to calculate a probability. In order to better exploit the Bayesian framework we could use a Markov Chain Monte Carlo (MCMC) method to generate random numbers and decide which values lead to a higher posterior likelihood in order to keep the value.

The Metropolis-Hastings (M-H) algorithm is a versatile MCMC method developed in the 1950s and generalized in the 1970s by Hastings [6]. The main idea behind the algorithm is to generate random numbers, as in Monte Carlo, then use those numbers for an iteration (such as using them as inputs for a probability function) which only depends on the previous iteration as in Markov Chains then the algorithm decides to keep or reject the value. If the probability of the iteration is higher than the previous probability then we keep the value otherwise it goes through an acceptance test, which states that if the value is bigger than some value generated by a uniform distribution from 0 to 1 then we keep that value.

Materials and methods

Clinical data of about 90 patients suffering from glioblastoma, a very aggressive type of brain cancer, was obtained from an oncology center in France. The database includes patient characteristics and outcomes of the treatment such as gender, age, location of tumor, surgery information, use of temozolomide as an adjuvant in radiotherapy, MRI and CT images, tumor recurrence locations, hematology grade, neurologic grade before and after the treatment, dates regarding the treatment, tumor size related items such as CTV, as well as other relevant information, concerning the treatment and follow up. For many of the patients, an initial treatment was performed in the standard 30 fractions, 2 Gy each for a total of 60 Gy in the tumor area.

We used this database to illustrate the methodology of developing clinically based prediction models using Bayesian statistics and machine learning. To exemplify the types of models that can be built we utilized the neurologic grade before and after the treatment and the CTV as parameters to predict the probability of developing a certain neurologic grade after the treatment and the uncertainty of this probability. Two main models were developed, the first model requires as input parameter the neurologic grade of the patient before the first treatment, and the second model requires additionally the CTV and both models aim to predict the neurologic grade after the treatment.

To accomplish the learning process, we used Bayes' theorem as a mathematical tool to determine the probability of a parameter given data. The process we used to develop clinically based prediction models is represented in figure 1:

- An initial set of data containing information about patients is known.
- Clinical observations are gathered.
- Known parameters, patient parameters, are passed down to initialize the computer model.
- Model parameters are initially guess for the model.
- The error between the predictions and the clinical observation is determined by data fitting process.
- he parameters are evaluated, for instance by a decision-making algorithm such as the Metropolis-Hastings (M-H), and passed down as model parameters.
- The decision-making algorithm is ran thousands of times in order to explore the posterior joint probability density function (pdf) of the model parameters.

Once correctly initialized, the model parameters posterior pdf can be used to infer the model prediction for any new patient, given its planed treatment. Moreover, the M-H algorithm gives us both, model parameters uncertainty and prediction uncertainty, thus providing a way to check the confidence level of our prediction.

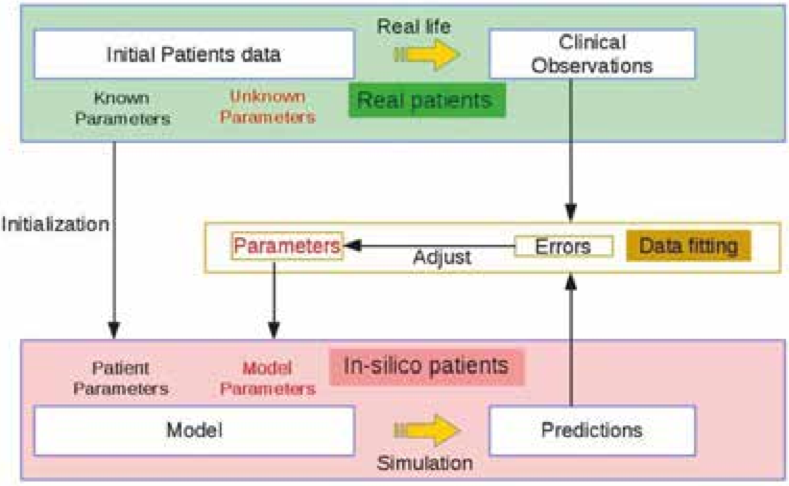


Figure 1. Clinically based model building diagram.

We implemented the M-H decision making algorithm to make the machine learning predictions, and the posterior function from Bayes' theorem was used as the proposal function in the M-H algorithm. The general Bayes theorem yields,

$$Posterior(parameters|data) = \frac{Likelihood(data\ |parameters) \cdot Prior(parameters)}{\Pr(data)}$$

The procedure for developing our first model is as follows:

- Construct the posterior function by constructing the likelihood function and defining the prior function based on previous knowledge of the prediction.
- Use the posterior function for the M-H algorithm.
- Verify the stability of the algorithm by Markov-Chains for instance.

Here for instance, we are interested in assessing the probability of developing a given neurologic grade after a Glioblastoma radiotherapy treatment. Knowing the patient grade before treatment, the likelihood for developing any given grade is a multinomial pdf (telling us the number of observations of a categorical variable, Gr$_i$, i.e. the final grade, given the frequency for observing this final grade, f$_i$). The prior probability (the knowledge we have for the frequencies) can be informative (extracted from previous knowledge) or non- informative (uniform priors for instance or best, Jeffrey's priors). Here, a uniform prior distribution was used since in this case we did not find any previous publication dealing with the prediction of the neurologic grade after the Glioblastoma treatment based on the initial grade. This way, we know that the posterior (the pdf of the parameters, given the observed patients final grade) is simply a Dirich let distribution which should be the target of the M-H algorithm. In this simple case, computations can be made by hand. The role of M-H algorithm is clarified for more complex cases.

For the second, enhanced model, we used additionally the CTV. We graph the grade vs the size of CTV and proposed a function to fit this data: we proposed a sigmoid function with unknown parameters a and b. Again the M-H algorithm was used to find those values as their posterior pdf cannot be computed by hand. The equation was used to construct the likelihood function for the posterior function in the Bayes' theorem. That is, Bayes' theorem was used to calculate a probability and the algorithm for the proposal of random values and for decision making to keep those vales which increase the likelihood of the parameters.

Results

From the glioblastoma database, 41 patients started with no neurologic grade then after the first (since several adjuvant treatments attempts were performed due to the aggressive nature of glioblastoma tumors) radiotherapy treatment 27 remained with no neurologic grade. However, 13 patients developed a grade of one while 1 patient developed a high grade of 3 as it can be observed in figure 2. On the other hand, most patients who started with a grade of one remained with the same grade. Using the data in figure 2 we calculated the probability of developing a certain neurologic grade after the first treatment given the patient started with a given grade. Evidently, there is not enough subjects in our data to produce statistically significant predictions. However, we use the data as an example of what type of models we can develop.

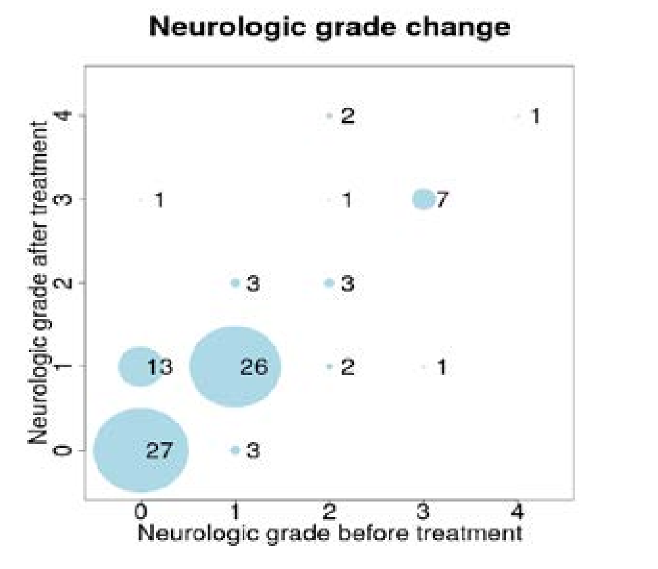


Figure 2. Neurologic grade change, by number of patients, before and after the first radiation therapy treatment.

Applying Bayes' theorem in the M-H algorithm in a machine learning type of process we were able to obtain the results in figure 3 which predicts the probability of developing a neurologic grade after the initial treatment for patients who started with no neurologic grade before the treatment. According to our predictions a patient has 60 % chances of remaining with no neurologic grade and about 30 % chances of developing a grade of 1. The probability calculated of developing grade three or four are very slim and is difficult to tell due to the limited number of subjects. The vertical value corresponds to the likelihood of the probability and the width of the graph represents the uncertainty of the prediction; the wider the curve the greater the uncertainty. With the results of this graph we can illustrate how to find a parameter, in this case the probability of a neurologic grade, and the confidence of that parameter.

Evidently this model is a simplistic model nonetheless the main idea is to keep enhancing the model in order to move forward towards a more realistic prediction. Therefore, we decided to develop the enhance model which shows that the prediction highly depends on the size of the tumor as can be seen in figure 4. The black solid circles in this figure represents the size of the CTV for each of the patients. In this graph we can see that a patient who started with a CTV below 200 cm^3 has less probability of developing a grade of one compared to those patients who started with a CTV above200 cm^3. The solid black circles in the lower part of the graph corresponds to those patients who started with no neu-

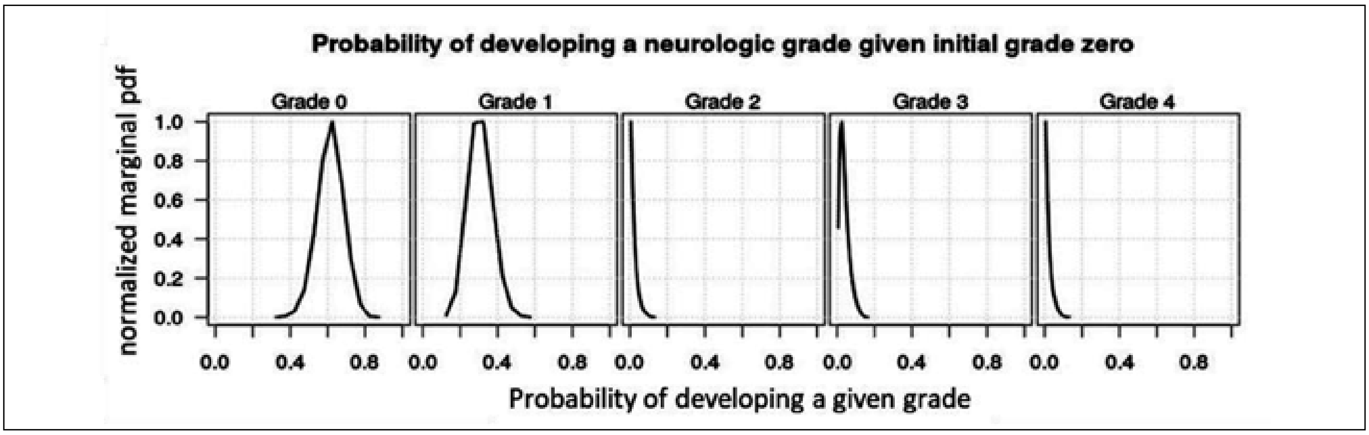


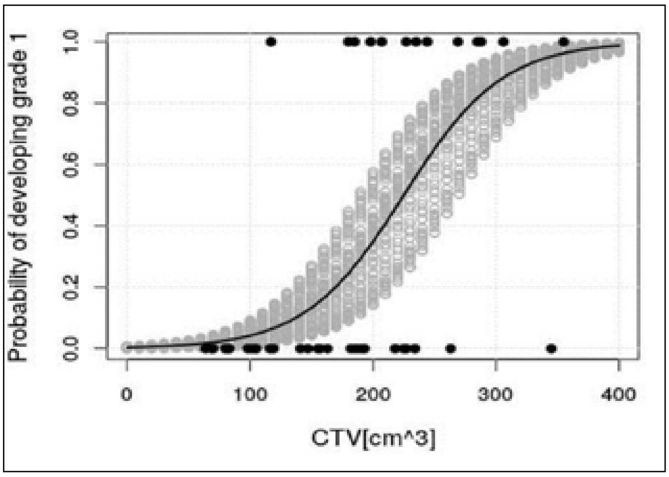
Figure 3. Probability of developing a neurologic grade after the first treatment given patients started with a grade zero before the treatment.

rologic grades and patients who started with a grade of one are drawn in the upper part of the graph. The following sigmoid function was used to fit the experimental data were a and b are unknown parameters calculated using the M-H algorithm.

$$f.sig(CTV) = \frac{1}{1 + e^{-(CTV-a)\cdot b}}$$

The solid black curve represents the sigmoid function draw with a highly likely value for a (225cm^3) and b (0.025cm^{-3}) parameters.

Figure 4. Probability of developing neurologic grade one based on CTV size for patients who started with no neurologic grade before the first treatment.



The marginal pdf of the a and b parameters calculated with the M-H algorithm are shown in figure 5. These curves are drawn with the purpose of illustrating that there can be other possible solutions. The horizontal axis corresponds to the value of the parameter and the vertical one to the normalized probability of this value.

From the prediction point of view, the joint pdf produced by the M-H algorithm gives us a clear understanding of the consequences of ongoing stochastic process. Pulling randomly in the joint pdf gives us the whole family of functions that could credibly describe the clinical outcomes, given all the already treated patients. On figure 4, the gray circles next to the main solid line represents another 100 sigmoid curves drawn with the different a and b parameters from our simulations. This helps to precisely define the confidence level that we can have in our prediction.

Discussion

One big challenge impeding the development of validated algorithmic models is the size of the data. Data set is often not sufficiently large enough to validate algorithms [7]. Another challenge is accessing oncology data since delicate issues are in play such as patient privacy and anonymity concerns. The challenges of privacy concerns are mentioned increasingly in the literature [8]. We were confronted with time consuming tasks for ac-

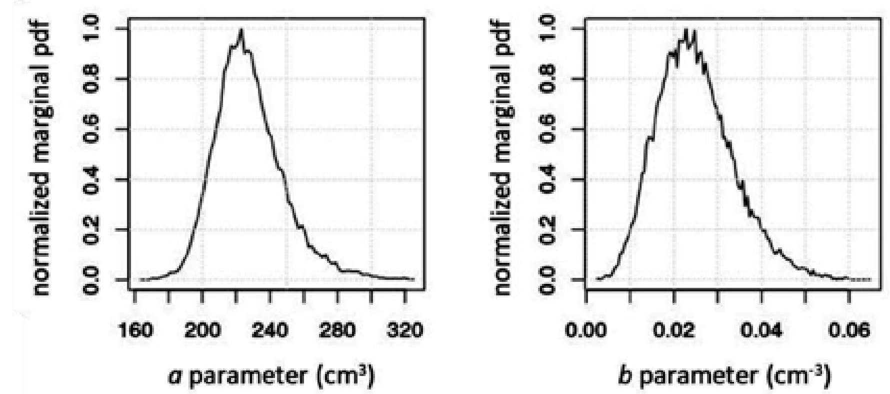


Figure 5. Marginal pdf of a and b, normalized to their maximum, to complete the proposed sigmoid function.

cessing a wider range of parameters for our database in order to be in line with the anonymity criterion.

We are aware of the limitation of our neurologic grade data in figure 1, and evidently there is not sufficient neurologic grade data in our study to produce statistically significant grade predictions. However, the main objective was to use the data to create examples of clinically based prediction models based on machine learning and Bayesian statistic to illustrate a versatile methodology for building these models hoping it aids to clarify concepts. That is, this works deals with the problematic of lowering the barriers of understanding the machine learning methodology, which is an important issue mentioned in the literature [1]. Therefore, the importance of illustrating the methodology of comprehensive model building since many times it is not quite well understood by non-experts in the subject. Understanding this methodology moves research forwards towards personalized medicine. Hence, our point of view is in agreement that machine learning tools have the potential to personalized medicine [4]. Lastly, it is worth highlighting the usefulness of the Bayesian framework to this work and due to its practicality, we would expect the strong continuation of the revival of this framework as mentioned by other authors in the literature [9].

Conclusions

The current trend of using concepts of machine learning will only keep increasing and the development of complex and sophisticated algorithms will drive the process. Simple clinically based prediction models were built using machine learning and Bayesian statistic in order to lucidly exemplify the model building methodology. In the process, we have made the contribution of correlating the neurologic grade prediction, after the first treatment of a glioblastoma treatment, following a simplified method hoping a wider audience would be able to follow before getting involved in to more complex machine learning processes.

Bibliographic References

[1] KANG J, SCHWARTZ R, FLICKINGER J, BERIWAL S. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. Int J Radiat Oncol Biol Phys. 2015; 93(5): 1127-1135.

[2] EL NAQA I. Perspectives on making big data analytics work for oncology. Methods. 2016; 11: 32-44.

[3] PELLA A, et. al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. Medical physics. 2011; 38: 2859-2867.

[4] SESEN MB, NICHOLSON AE, BANARES-ALCANTARA R, et. al. Bayesian networks for clinical decision support in lung cancer care. PLOS ONE. 2013; 8(12): e82349.

[5] CASTANEDA C, et. al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma. 2015; 5: 4.

[6] CHIB S, GREENBERG E. Understanding the metropolis-hastings algorithm. The American Statistician. 1995; 49(4): 327-335.

[7] BIBAULT JE, GIRAUD P, BURGUN A. Big data and machine learning in radiation oncology: state of the art and future prospects. Cancer Letters. 2016; 382: 110-117.

[8] LUSTBERG T, et. al. Big data in radiation therapy: challenges and opportunities. Br J Radiol. 2017; 90(1069): 20160689.

[9] ADAMINA M, TOMLINSON G, GULLER U. Bayesian statistics in oncology: a guide for the clinical investigator. Cancer. 2009; 115: 5371-5381.

Recibido: 13 de febrero de 2018
Aceptado: 29 de mayo de 2018