Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 72 – 77

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Prediction in Radiotherapy Treatments: current trends

Ana Anacleto[a]*, Joana Dias[a,b]

[a]*Faculty of Economics, University of Coimbra, 3004512 Coimbra, Portugal*
[b]*Inesc-Coimbra, University of Coimbra, 3030290 Coimbra, Portugal*

Abstract

In this paper, a review of the existing state-of-the-art regarding prediction models in radiotherapy treatments is made. We focus the scope of the paper in data mining techniques used in the context of radiotherapy treatments. There are several data mining algorithms applied to datasets of cancer patients receiving radiotherapy treatments, with very distinct variables and heterogeneous features. The existing literature presents significant advantages in using data mining approaches to predict outcomes in this type of treatments, with an increasing adherence to their use and great potential to explore. Recent published studies are considered, followed by a discussion and some conclusions with the identification of possible future work.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Radiotherapy; Data Mining; predictions; radiotherapy treatments; cancer

1. Introduction

Cancer is one of the diseases of this century. Radiotherapy treatments are frequently chosen for patients with cancer, having an important role in loco regional tumors' control and increase in life expectancy[1]. In a radiotherapy treatment, an individualized treatment plan is considered that depends on the cancer type, the patient's clinical condition and disease development and taking into account several factors like demographic data (i.e.; gender, age, habits and profession), genetic risk and family history[2]. To be able to predict tumor response to radiotherapy treatments is one of the main challenges in cancer treatment[3], and predict it with precision will provide, to physicians, better tools for well-informed decision-making, concerning the benefits to patients and expected risks, allowing them to adjust and personalize the plans of treatment[4]. However, capturing the complexity that exists due

* Corresponding author. Tel.: +0-351-964-158-606
E-mail address: acsanacleto@hotmail.com

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.125

to all the interactions between heterogeneous variables is a very difficult task[4]. Actually, the same treatment may have different outcomes for patients with the same type of cancer[5]. The tumor control through radiation is affected by interactions of great complexity and involving relationships between tumor biology, the environment in which they developed, the radiation dose and the variables related to the patient[6], including radio sensitivity[7]. Prescription of the ideal treatment plan is not possible and the delivered treatment usually represents a compromise solution between acceptable dose for the tumor and minimization of complications in normal tissues[2]. The objective of this work is to present a brief and non exhaustive review of the current state of the art regarding the development and application of prediction approaches for radiotherapy treatments, trying to highlight the current different applications in this field. The review considers published papers referring to data mining techniques and published over the last 10 years. These papers were mostly collected from Google Scholar. The objective was to observe the datasets used, the choice of attributes, the choice of techniques and the achieved results. This paper is organized as follows: in the next section we shortly describe recent published works, considering the different aims. The state-of-the-art is discussed in section 3, and section 4 concludes, pointing out some possible future paths of research.

2. What to predict

Various authors have been using several techniques for building prediction models for radiotherapy treatments outcomes and complication risks for normal tissues. Usually the radiotherapy outcomes are characterized by tumor control probability (TCP) and the effects on adjacent tissues are quantified by Normal Tissue Complication Probability (NTCP)[8]. Considering survival analysis, the objective is to estimate the patient's survival probability during a period of time. It is one of the most challenging tasks physicians face and correct survival estimate in terminal patients helps preventing inadequate or unnecessary therapies and toxicity[9]. The survival expectation in patients with advanced metastatic cancer, for instance, significantly affects decisions concerning future treatment plan[10]. The best known models for survival predictions are based on conditional probabilities, but progress in areas such as knowledge discovery and data mining techniques have made possible the development of new and more powerful models[11].

2.1. Normal Tissues Complications Probability (NTCP)

Gulliford et al.[2] explore the use of Artificial Neural Networks (ANN) to predict bladder and rectum complications after radiotherapy, in prostate cancer cases. The ANNs were trained using a group of 126 patients treated similarly. Results were encouraging, but demonstrated that the model precision was limited due to the low number of cases. In another research about performance impact of different statistical methods in the creation of NTCP prediction models for head and neck cancer, 185 cases and 21 variables were considered for xerostomia prediction[1]. The authors used the Stepwise Selection, Least Absolute Shrinkage and Selection Operator (LASSO) and Bayesian Model Averaging (BMA). It was verified that LASSO showed, both, the best performance and easiest interpretation results. In this study, it was verified that the number of variables is still small and more variables would be necessary to be able to consider details of the treatment plan and other prognostic factors. Langendijk et al[12] try to predict the swallowing dysfunction that could be easily used as prognostic tool in clinical practice for identification of risk groups. The dataset in study included 529 patients with head and neck cancers treated with radiotherapy treatments and that were alive and free from disease six months after the conclusion of the treatment. For model building, the authors used univariate and multivariate Linear Regression (LR) analysis techniques to study the association between the initial evaluation, treatment characteristic and risks of acquiring the swallowing dysfunction in 6 months. They defined a swallowing dysfunctions risk scale (TDRS – *Total Dysphagia Risk Score*), calculated by the sum of several risk factors values. Koiwai et al.applied TDRS, considering a sample of 47 patients with the same cancer type and that takes into consideration several risk factors[13]. To evaluate the capacity of prediction of TDRS they analyzed Receiver Operating Characteristic (ROC) curves and the respective Area Under the Curve (AUC) values. Results indicated that, in this context, TDRS is a valid measure to be considered for prediction of swallowing function complications. With the same objective of predicting swallowing complications, other authors used a sample of 96 patients with head and neck cancer diagnosis[14]. LR analysis was used to evaluate the relationships between dose-volume factors and swallowing

complications, showing that it is possible to reduce the toxicity and the long term swallowing complications. In a similar trend, Christianen et al.[15] applied LR to a set of 354 patients with head and neck cancer. The model performance measure used was AUC. Results showed that is difficult to establish a relationship between the dose-volume distribution of the organ at risk and the degree of complication. Support Vector Machines (SVM) were used with 3 datasets to assess if kernel based machine learning methods could improve models using institutional data and resampling methods[4]. The first dataset corresponds to 55 patients with head and neck cancer, the second and third datasets consist of, respectively, 52 and 45 patients with lung cancer. An independent set for evaluation was also used. One of the things that limited the capacity of the model prediction was the fact that relevant variables related with the patient and the disease were missing. SVMs are also used in a prediction model for lung lesions development[16]. Two separate datasets considering patients with lung cancer and treated with radiotherapy were used. The first dataset corresponds to the retrospective record of 219 patients and the second dataset, with 19 patients, serves as base to a prospective analysis. It was demonstrated that SVMs based models can improve the prediction of lung lesions development compared with the traditional LR methods.

2.2. Tumor Control Probability (TCP)

Several Data Mining techniques were used in order to discover hidden relationships between the prognostics variables for dose-volume dosimetry and several radiobiological processes in patients with lung cancer and to generalize data not yet applied prospectively for the purpose of improving TCP prediction[6]. This approach was motivated by the extraordinary increase in the availability of the specific biological and clinical patient's data due the progress in genetics and image technology. A dataset of 56 patients and a total of 23 were considered for the TCP model. Both LR and SVM were tested, and the authors concluded that SVM presented a better performance for TCP prediction possibly due to the capacity of dealing with nonlinear and complex interactions between variables.

2.3. Breathing movements

Some authors have tried to predict breathing movements with the objective of improving treatment delivery, making it possible to synchronize the radiation beam incidence with the moving target. Breathing movements' patterns are, inherently, complex by nature[17]. To predict lung cancer tumors displacement movement caused by breathing, ANNs were used[18]. Breathing movements' data were collected during 60 sessions for samples of five patients and evaluated the prediction accuracy of the Sinusoidal Model and the Adaptive Filter Model algorithms[18]. In general, the breathing movements' prediction done with the Adaptive Filter Model presented a better performance than the Sinusoidal Model. For 10 patients with lung cancer, the Auto-Regressive and Moving Average Model (ARMA) was evaluated for the prediction of irregular breathing movements[19]. It was possible to improve the delivery precision of real-time motion compensation radiotherapy.

2.4. Survival

Delen et al.[11] developed classification models based in ANNs, Decision Trees (DT) and LR for prediction of survival capacity in breast cancer patients. A dataset with more than 200.000 cases and 17 variables was used. The DT model presented the best performance. Gao et al. compared 9 different data mining techniques for survival rate prediction at 5 years[20]. Two groups of colon-rectal cancer patients' data were used, coming from two different sources. The algorithms used were Back Propagation Network (BP), Radial Basis Function (RBF), General Regression Neural Network (GRNN), Adaptive-Network Based Fuzzy Inference System (ANFIS), SVM, Bayesian Networks (BN), Naive Bayes (NB), Classification And Regression Tree (CART) and the LR. The purpose was to evaluate the models' precision when compared with TNM and the performance measure used was AUC. The first dataset was composed of more than 10 000 registers and 20 variables randomly selected from a dataset with more than 36 000 registers. The second dataset was composed by approximately 760 registers and 14 variables randomly selected from a different dataset with more than 1 500 registers. It was verified that the models

built with Data Mining algorithms showed slightly more accuracy when compared with the TNM model, however, the application of the TNM system is simpler.

2.5. Tumors classification

Given the importance of a correct tumor classification, a research was developed by Huml et al.[21], proposing the use of data mining techniques, based on Minkowski Functionals, combined with images created based on Atomic Force Microscopy (AFM) methodology, done from histopathological samples. AFM became a widely used technique for the characterization of biological samples with nanometric resolution. The idea was to be able to increase accuracy in the determination of specific tumors characteristics and to identify, without ambiguity, the tumors of degree II from the remaining ones. A sample of 113 records obtained from 14 patients with brain cancer was used. This approach allowed a great precision in tumor classification and it could offer new elements for more objective diagnoses.

2.6. Radiotherapy Effectiveness

In the case of multi-form glioblastoma a pilot study was made where a model that tries to predict the radiotherapy effectiveness was developed[22]. A dataset of 9 patients with this cancer type and that received radiotherapy was considered. The authors used classic linear-quadratic (LQ) model to define a proliferation model extension and invasion of the gliomas, including the radiotherapy effects. It was the first model built to predict radiobiological parameters in human beings.

3. Discussion

Datasets are a crucial raw material that are the base for all the studies referred to in the previous section. The importance of well-structured records is nowadays widely recognized. The International Agency for Research on Cancer (IARC) recognized, in its biennial report, the need to improve the provision of global information on this disease and to increase the population coverage regarding high quality records, particularly in the developing countries[23]. Some important steps have been given for same agencies, as the CancerData.org initiative, in trying to collect clinical cancer data all over the world to provide to researchers' communities global shareable databases. Some of the features pointed as essential to obtain good records are the record exhaustiveness, the accuracy and its timely availability[24]. Typically, clinical data are collected in the course of patient care, many times in a manual way, while the necessary research data are forgotten or left for second plan. Therefore, the clinic databases can present characteristics that hinder the application of data mining tools. The data may have missing values or noise, be imprecise, redundant or inconsistent[11]. Problems associated with clinical data gathering and availability can be minimized with investments in data collection processes, since this field of research has shown to bring several advantages[11]. As can be seen by the existing literature, most of the existing works are based on small datasets. Several authors refer difficulties in model building and generalization due to the small data dimension and available variables[2,6,18,20,22]. Having enough data to feed data mining models is crucial if we want to obtain higher quality results leveraging the possibility of knowledge retrieval and generalization. The attributes that are available in the datasets is another important feature that has to be considered, since, in general, the variables that are most commonly found in the literature are those related with sociodemographic data, description of tumor characteristics, radiation doses and adjuvant treatments, associated complications, treatments modality and other features related to the patient's general state. Having a wider set of attributes makes possible the use of variable selection approaches, that will allow a better selection of prediction variables that, hopefully, will bring more and better insights regarding the potential relationships that exist between dependent and independent variables[1]. That is clinically important because the variables are information bearers about potentials causes and relationships that allows to find patterns and to define accurate models[4]. Data dissemination in this research field can be quite complex due to concerns regarding the confidentiality and privacy of the data[25]. It's very important to balance the demands of research communities, data access policies and comply with the ethical and legal procedures to maintain the confidentiality of respondents and data providers[26].The creation of synthetic databases to protect

individual records can be a way of overcoming some of the existing problems[27]. Synthetic databases aim at preserving the data information content and, at the same time, protecting the confidentiality, ensuring that synthetic data records are drawn from a model that fits statistically to the original data[26]. These data have the same statistical properties as the original data, but contain dummy information[28].

4. Conclusion

The ability of predicting correctly the future outcomes of radiotherapy treatments can be an important tool in medical decision making. Data mining techniques are shown to be well suited in this context, especially with the growing existence of available datasets. The literature points to an increasing adherence of these new techniques and a huge potential in using it, being possible to find significant advantages in their application when compared with traditional approaches. There are several open paths for future research. Being able to create synthetic databases that mimic the real medical records would bring several advantages to the research community, since new methodologies could be developed without the dependence of available huge databases. Furthermore, the use of ensemble methodologies could also improve the accuracy of prediction models. At the end, it will be important to develop tools that can be used in clinical practice, helping the medical doctor prescribing the best treatment for each patient and relying on accurate models that are capable of estimating the radiotherapy treatment outcome both in terms of the tumor and the preservation of organs at risk.

Appendix A. Table with a summary of the referred papers

Table 1. Prediction Type Synthesis

Prediction Type	Methods	Performance Measure	Resampling	Datasets Characteristics
Tumor Classification[21]	Minkowski Functionals	Minkowski Measures		Patients: 14 Records: 113
Radiotherapy Effectiveness[22]	LQ		Leave-One-Out of Cross-Validation	Patients: 9
Survival[11]	ANN, DT and LR	Precision, Sensibility and Specificity		Records: 200000 Atributes:17
Survival[20]	BP, RBF, GRNN, ANFIS, SVM, BN, NB, CART and LR	AUC	Cross-Validation	Records: 10000+760 Atributes:20+14
Breathing movements[18]	ANN			Patients: 3
Breathing movements[17]	Sinusoidal Model and Adaptive Filter Model	Precision Error		Patients: 5
Breathing movements[19]	ARMA	Precision Error		Patients: 10
TCP[6]	LR and SVM	Spearman Correlation	Leave-One-Out of Cross-Validation and Bootstrap	Patients: 56 Atributes:23
NTCP[2]	ANN	Sensibility and Specificity	Train and test	Patients: 119
NTCP[1]	Stepwise Selection, LASSO and BMA		Train and test	Records: 185 Atributes:21
NTCP[12]	LR			Patients: 529
NTCP[4]	SVM	Matthews Correlation Coeficient	Cross-Validation	Patients: 55+52+45
NTCP[13]	TDRS	AUC and ROC		Patients: 47
NTCP[14]	LR			Patients: 96
NTCP[15]	LR	AUC	Bootstrap	Patients: 354
NTCP[16]	SVM	Matthews Correlation Coeficient	Leave-One-Out of Cross-Validation and Bootstrap	Patients: 219+19

References

1. Xu C-J, van der Schaaf A, Schilstra C, Langendijk J a., van't Veld A a. Impact of Statistical Learning Methods on the Predictive Power of Multivariate Normal Tissue Complication Probability Models. *Int J Radiat Oncol*. 2012;82(4):e677–e684.
2. Gulliford SL, Webb S, Rowbottom CG, Corne DW, Dearnaley DP. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. *Radiother Oncol*. 2004;71:3–12.
3. Ogawa K, Murayama S, Mori M. Predicting the tumor response to radiotherapy using microarray analysis (Review). *Oncol Rep*. 2007;18:1243–1248. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17914580.
4. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. *Phys Med Biol*. 2009;54(December 2008):S9–S30.
5. Lambin P, Roelofs E, Reymen B, et al. "Rapid Learning health care in oncology" - An approach towards decision support systems enabling customised radiotherapy. *Radiother Oncol*. 2013;109(1):159–164.
6. Naqa I El, Deasy JO, Mu Y, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol*. 2010;49(November 2009):1363–1373.
7. Lambin P, van Stiphout RGPM, Starmans MHW, et al. Predicting outcomes in radiation oncology-multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013;10:27–40.
8. El Naqa I, Bradley J, Blanco AI, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. *Int J Radiat Oncol Biol Phys*. 2006;64(4):1275–1286.
9. Gripp S, Moeller S, Bölke E, et al. Survival prediction in terminally ill cancer patients by clinical estimates, laboratory tests, and self-rated anxiety and depression. *J Clin Oncol*. 2007;25(22):3313–3320.
10. Fairchild a, Debenham B, Danielson B, Huang F, Ghosh S. Comparative multidisciplinary prediction of survival in patients with advanced cancer. *Support Care Cancer*. 2014;22(3):611–7.
11. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif Intell Med*. 2005;34:113–127.
12. Langendijk J a., Doornaert P, Rietveld DHF, Verdonck-de Leeuw IM, René Leemans C, Slotman BJ. A predictive model for swallowing dysfunction after curative radiotherapy in head and neck cancer. *Radiother Oncol*. 2009;90(2):189–195.
13. Koiwai K, Shikama N, Sasaki S, Shinoda A, Kadoya M. Validation of the total dysphagia risk score (TDRS) as a predictive measure for acute swallowing dysfunction induced by chemoradiotherapy for head and neck cancers. *Radiother Oncol*. 2010;97(1):132–135.
14. Caglar HB, Tishler RB, Othus M, et al. Dose to Larynx Predicts for Swallowing Complications After Intensity-Modulated Radiotherapy. *Int J Radiat Oncol Biol Phys*. 2008;72(4):1110–1118.
15. Christianen MEMC, Schilstra C, Beetz I, et al. Predictive modelling for swallowing dysfunction after primary (chemo)radiation: Results of a prospective observational study. *Radiother Oncol*. 2012;105(1):107–114.
16. Spencer SJ, Almiron Bonnin D, Deasy JO, Bradley JD, El Naqa I. Bioinformatics methods for learning radiation-induced lung inflammation from heterogeneous retrospective and prospective data. *J Biomed Biotechnol*. 2009:14.
17. Vedam SS, Keall PJ, Docef a, Todor D a, Kini VR, Mohan R. Predicting respiratory motion for four-dimensional radiotherapy. *Med Phys*. 2004;31:2274–2283.
18. Isaksson M, Jalden J, Murphy MJ. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. *Med Phys*. 2005;32:3801–3809.
19. Ren Q, Nishioka S, Shirato H, Berbeco RI. Adaptive prediction of respiratory motion for motion compensation radiotherapy. *Phys Med Biol*. 2007;52:6651–6661.
20. Gao P, Zhou X, Wang ZN, et al. Which is a more accurate predictor in colorectal survival analysis? nine data mining algorithms vs. the TNM staging system. *PLoS One*. 2012;7(7):1–8.
21. Huml M, Silye R, Zauner G, Hutterer S, Schilcher K. Brain tumor classification using AFM in combination with data mining techniques. *Biomed Res Int*. 2013;2013.
22. Rockne R, Rockhill JK, Mrugala M, et al. Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach. *Phys Med Biol*. 2010;55:3271–3285.
23. *INTERNATIONAL AGENCY FOR RESEARCH ON CANCER - Biennial Report*.; 2013. Available at: http://governance.iarc.fr/SC/SC50/Biennial Report 2012-2013.pdf.
24. Nóbrega SD, Paulino CD. Registos de Cancro; Controlo de Qualidade; Exactidão; Validade; Fiabilidade . *Inst Nac Estatística*. 2001:1–22.
25. Drechsler J, Reiter JP. Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata. *J Am Stat Assoc*. 2010;105(492):1347–1357.
26. Penny R, Graham P, Young J. *Methods for Creating Synthetic Data*.; 2008.
27. Gupta V, Miklau G, Polyzotis N. Private Database Synthesis for Outsourced System Evaluation. *AMW*. 2011:1–12.
28. Gray J, Sundaresan P, Englert S, Baclawski K, Weinberger PJ. Quickly generating billion-record synthetic databases. *ACM SIGMOD Rec*. 1994;23(2):243–252.