10-30-2014

A Decision Support Model for Personalized Cancer Treatment

Florentino Antonio Rico-Fontalvo
University of South Florida, fricofon@mail.usf.edu

A Decision Support Model for Personalized Cancer Treatment

by

Florentino Antonio Rico-Fontalvo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Industrial Engineering
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Major Professor: Grisselle Centeno, Ph.D.
Steven Eschrich, Ph.D.
Javier Torres-Roca, M.D.
Ali Yalcin, Ph.D.
Jose Zayas-Castro, Ph.D.

Date of Approval:
October 30, 2014

Keywords: Supervised Learning, Fuzzy Logic, Systems Biology, Gene Expression,
Rectal Cancer, Random Forest

DEDICATION

I want to dedicate this dissertation to my parents: Martha Cecilia Fontalvo-Rivera and Florentino Antonio Rico-Calvano: you are the source of my drive, inspiration and motivation. Thank you Mom and Dad, none of this could not have been possible without your support, love and encouragement. This is a tribute to both of you.

ACKNOWLEDGMENTS

I would like to thank my mentor and major advisor, Dr. Grisselle Centeno, who has been with me through this journey. You believed in me from the very beginning, and it has been an honor working with you. Your role as an advisor goes beyond this dissertation, I have no words that can express my gratitude. I sincerely thank you for giving the opportunity of achieving my greatest academic accomplishment.

I cannot express enough thanks to my committee for their continued support and encouragement. Dr. Zayas-Castro, thank you for being an integral part of my academic journey. I will always be grateful for your mentoring, academic, professional and financial support I received from you during the last years, I will always carry your advice in the journey I have ahead. Dr. Yalcin, thank you for your friendship and the learning opportunities I had with you. Dr. Eschrich and Dr. Torres-Roca, I offer my sincere appreciation for sharing and contributing so much to my research, thank you for opening a new world of knowledge and believing in my potential, my heartfelt thanks.

The completion of this work could not have been possible without the support of Gloria Latter, Liz Conrad and Catherine Burton. Thank you for your help and having patience with me during this dissertation journey.

To my brother and sisters: Jorge, Heidy and Ximena, you are an inspiration and my role models. My nephews and niece: Camilo, Alejandro and Juliana, I always carry you in my mind.

Finally, to Santiago: my deepest gratitude. It was your encouragement when the times got rough that kept me going. Thank you for your patience and unconditional support, know that it is much appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This work is motivated by the need of providing patients with a decision support system that facilitates the selection of the most appropriate treatment strategy in cancer treatment. Treatment options are currently subject to predetermined clinical pathways and medical expertise, but generally, do not consider the individual patient characteristics or preferences. Although genomic patient data are available, this information is rarely used in the clinical setting for real-life patient care. In the area of personalized medicine, the advancement in the fundamental understanding of cancer biology and clinical oncology can promote the prevention, detection, and treatment of cancer diseases.

The objectives of this research are twofold. 1) To develop a patient-centered decision support model that can determine the most appropriate cancer treatment strategy based on subjective medical decision criteria, and patient's characteristics concerning the treatment options available and desired clinical outcomes; and 2) to develop a methodology to organize and analyze gene expression data and validate its accuracy as a predictive model for patient's response to radiation therapy (tumor radiosensitivity).

The complexity and dimensionality of the data generated from gene expression microarrays requires advanced computational approaches. The microarray gene expression data processing and prediction model is built in four steps: response variable transformation to emphasize the lower and upper extremes (related to Radiosensitive and Radioresistant cell lines); dimensionality reduction to select candidate gene expression probesets; model development using a Random Forest algorithm; and validation of the model in two clinical cohorts for colorectal and esophagus cancer patients.

Subjective human decision-making plays a significant role in defining the treatment strategy. Thus, the decision model developed in this research uses language and mechanisms suitable for human interpretation and understanding through fuzzy sets and degree of membership. This treatment selection strategy is modeled using a fuzzy logic framework to account for the subjectivity associated to the medical strategy and the patient's characteristics and preferences. The decision model considers criteria associated to survival rate, adverse events and efficacy (measured by radiosensitivity) for treatment recommendation. Finally, a sensitive analysis evaluates the impact of introducing radiosensitivity in the decision-making process.

The intellectual merit of this research stems from the fact that it advances the science of decision-making by integrating concepts from the fields of artificial intelligence, medicine, biology and biostatistics to develop a decision aid approach that considers conflictive objectives and has a high practical value. The model focuses on criteria relevant to cancer treatment selection but it can be modified and extended to other scenarios beyond the healthcare environment.

CHAPTER 1: INTRODUCTION

1.1 Background

Rectal cancer is a disease in which malignant cells form in the tissues of the rectum [1]. The rectum is part of the colon and is located in the gastrointestinal track; thus, its position in the pelvis poses additional challenges in treatment when compared with colon cancer (see Figure 1) [2].

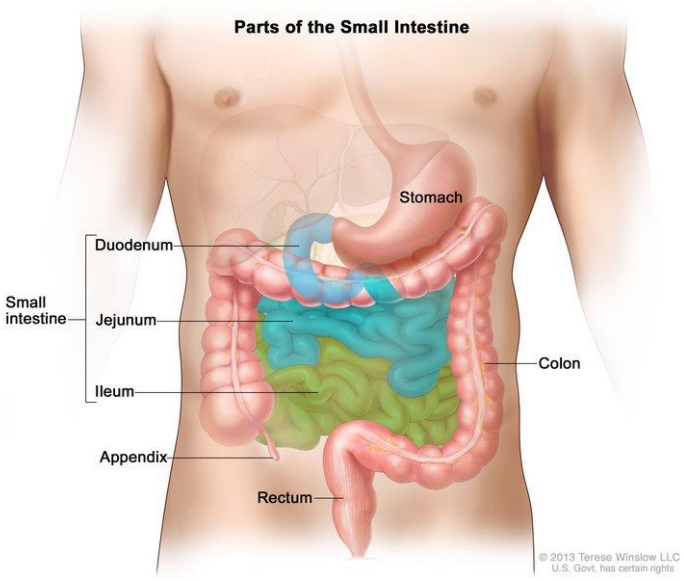


Figure 1 Diagram of colon and rectum. National Cancer Institute ©2013 Terese Winslow

Colorectal cancer is the third most common cancer diagnosed in both men and women in the United States. According to the American Cancer Society, 96,830 new cases of colon cancer and 40,000 new cases of rectal cancer were reported in 2014 [2]. However, rates have been declining by 3% per year in men and by 2.3% per year in women since 1998. This trend has been attributed to the detection and removal of precancerous polyps as a result of

colorectal cancer screening [3] . Overall, only 39% of colorectal cancer patients diagnosed between 1999 and 2006 had localized-stage disease, for which the 5-year relative survival rate is 90%; 5-year survival rates for patients diagnosed at the regional and distant stage are 70% and 12%, respectively [4]. The 5-year observed survival rate for colon and rectal cancer patients between 1998 and 2000 are shown in Table 1 by cancer staged from the 7th edition of the AJCC staging system (from National Cancer Institute's SEER database) [5]. The observed estimates in Table 1 may be lower than actual survival rates since it includes patients who could have died from other causes than cancer during the observed timeframe (e.g. heart disease).

Table 1 Survival rates for rectal and colon cancer by stage

5-year Observed Survival Rate		
Stage	**Colon Cancer (%)**	**Rectal Cancer (%)**
II	74	74
IIA	67	65
IIB	59	52
IIC	37	32
IIIA	73	74
IIIB	46	45
IIIC	28	33
IV	6	6

The general process for rectal cancer detection and treatment is captured in Figure 2. The process consists on first detecting and diagnosing the cancer, determining the stage of the cancer, and selecting the treatment (two or more types of treatment may be combined or used in sequence) based on the cancer stage prognosis and physician's expertise. After treatment, follow up and monitoring is recommended to assess treatment effectiveness and as a preventive measure. In practice, there are algorithms in place that suggests the treatment combination based on the cancer stage and cancer type. Patients with rectal cancer stage II and III are recommended to have neoadjuvant therapy, as presented by treatment selection

algorithm for rectal cancer patients created by the MD Anderson Cancer Center [6]. Each process component is described in detail in the next few sections.

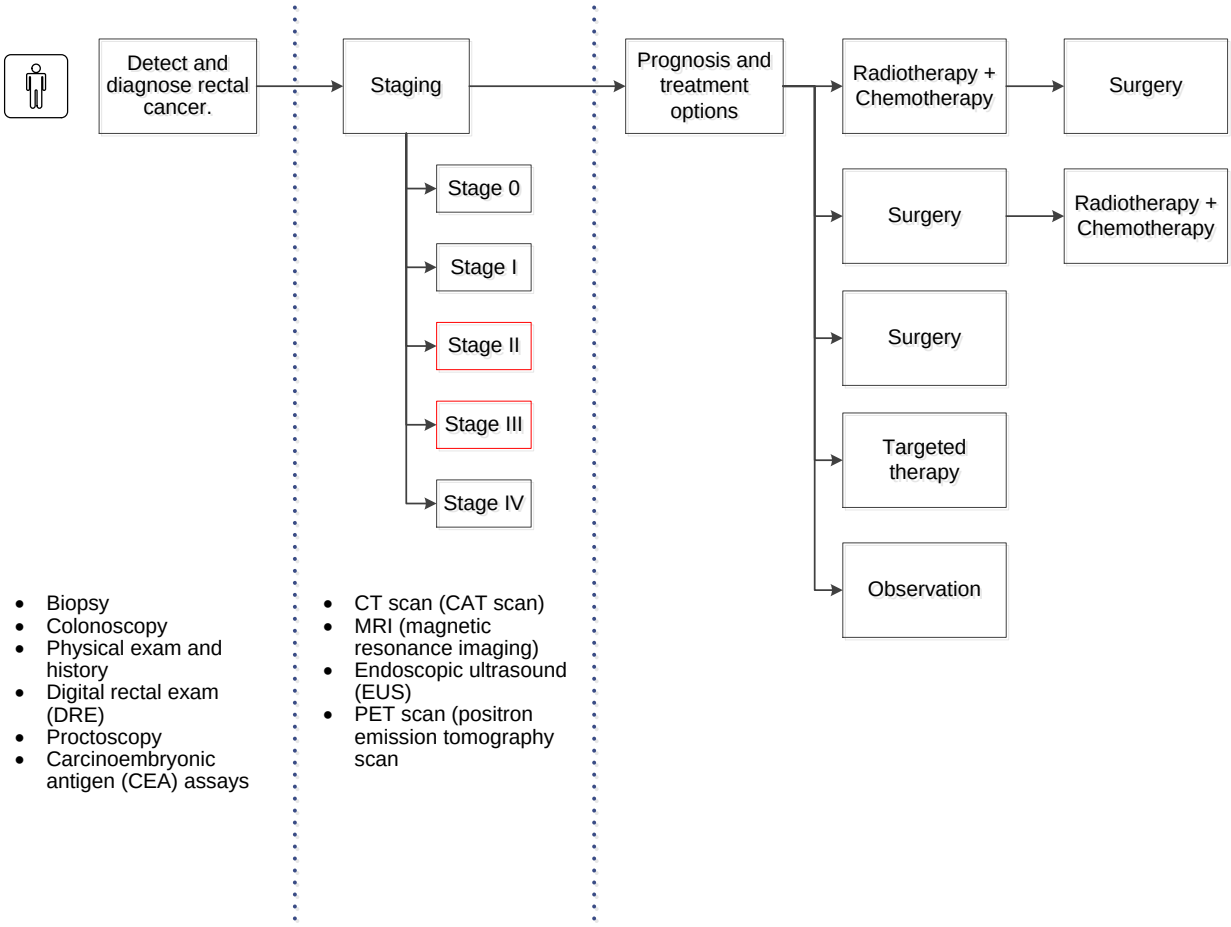


Figure 2 Rectal cancer detection and staging process

1.1.1 Rectal Cancer Diagnosis

Most people in early colon or rectal cancer stages do not experience the symptoms of the disease. Thus, screening tests are recommended to detect and diagnose the cancer before it further progresses. One or more of tests used to detect and diagnose colon and rectal cancer include [7]:

- Endoscopic tests are nonsurgical procedures to examine and remove suspicious tissue or polyps. Depending on how far up the colon is examined, three tests are performed: proctoscopy to view the rectum; sigmoidoscopy to view of the rectum and lower colon; and colonoscopy to view the entire colon
- Endoscopic ultrasound: a picture (sonogram) is obtained by bouncing high-energy sound waves (ultrasound) off internal organs
- Imaging tests infuse energy through a patient and can show abnormal body structures. Changes in energy patterns are captured to create an image or picture that is reviewed by a physician and include: computed tomography scan (CT), magnetic resonance imaging scan (MRI), and positron emission tomography scan (PET)
- Digital rectal exam
- Carcinoembryonic antigen (CEA) measures the quantity of this protein in the blood of patients who have may have colon or rectal cancer

1.1.2 Staging

Staging is the process of determining the spread and extent of the cancer tumor once it has been diagnosed. It is based on the results of the physical exam, biopsies, blood and imaging tests. The American Joint Committee on Cancer (AJCC) staging system, also known as the TNM system, is the tool most commonly staging used for colorectal cancer [2]. The TNM consists of three key elements: 'T' defines how much the tumor has grown into the wall of the intestine; 'N' defines the extent of spread to other lymph nodes; and 'M' defines whether the cancer has metastasized to other organs of the body

Once the patient's T, N and M categories have been determined, a stage grouping (from stage I to stage IV in Figure 2) is determined from the least advanced to the most advanced stage. The TNM combinations for each cancer stage are presented in Appendix A.

1.1.3 Treatment Options

There are different types of treatment for rectal cancer, some are standard practice and others are being tested in clinical trials. According to the National Cancer Institute (NCI), four types of standard treatment are used: surgery, radiation therapy (RT), chemotherapy, and targeted therapy [8]. There treatments can be performed separately or combined as shown in Figure 2. The oncologist will select the best therapy based on the type of cancer, stage and location of the tumor.

The primary treatment used in rectal cancer is surgical resection [9]. According to the NCI, local excision of clinical tumors is commonly used for selected patients in rectal cancer stage T1. For higher stages of rectal cancer, a total mesorectal excision (TME) is the treatment of choice. Since the introduction of TME for rectal cancer, reduced local recurrence rates and improved oncologic outcomes have been observed [10]. Depending on the surgeon's experience, the rate of complications, such as blood loss and anastomotic leaks, are low. Furthermore, radiotherapy before surgery appears to benefit patient outcomes even with improvements in surgical technique [10].

RT is the most commonly prescribed treatment in rectal cancer treatment. Approximately 50% of cancer patients will receive RT alone or in combination with other treatments [11]. When used before surgery, the goal is to shrink the tumor to make surgery or chemotherapy more effective. When used afterward, it is used to destroy any cancer cells that might remain after surgery [6]. There are two basic types of RT:

- External beam radiation is administered by a machine and rotates around the patient's body to deliver a high dose of radiation directly to the tumor (some of the tissue around the tumor can also be affected).

- Internal radiation, also known as brachytherapy, consists of a radiation source that is implanted in the body at the tumor site. Based on the type of the tumor, the appropriate equipment is selected for treatment.

A combination of radiation and chemotherapy before radiation (also known preoperative chemo-radiation (CRT) or neoadjuvant therapy) has become the standard of care for patients with clinically staged T3–T4 or node-positive disease based on the results of clinical trials [9]. CRT may be given before surgery to shrink the tumor, make it easier to remove the cancer, and lessen problems with bowel control after surgery. Even if all the cancer that can be seen at the time of the surgery is removed, some patients may be given radiation therapy or chemotherapy after surgery to kill any cancer cells that are left. Treatment given after the surgery to lower the risk of relapsing is called adjuvant therapy.

1.1.4 Adverse Effects of Radiation Treatment

For patients with rectal cancer stage II and III, neoadjuvant treatment with RT and 5-FU-based chemotherapy is preferred compared to adjuvant therapy in reducing local recurrence and minimizing toxicity [12]. However, there are specific challenges and adverse effects associated with the RT in rectal cancer patients. These include:

- Gastrointestinal disorders: diarrhea, bleeding, abdominal pain and obstruction due to stenosis or adhesions
- Genitourinary dysfunction: incontinence, retention, dysuria, frequency and urgency
- Sexual Dysfunction: in males, a long-term deterioration of ejaculatory and erectile function; and in females, RT was associated with vaginal dryness and diminished sexual satisfaction
- Second Cancers: risk of second cancers from organs within or adjacent to the irradiated target. The most common second cancers include gynecologic and prostate.

RT after or before surgery treatment has negative effects on toxicity and the quality of life of the patient; therefore, treatment options should be discussed with the patient.

1.2 Personalized Medicine

Personalized medicine refers to the use and implementation of the patient's unique biologic, clinical, genetic and environmental information to make decisions about their treatment or course of action [13]. Cancer therapy is implemented on a watch-and-wait basis for most patients. Although an individual's clinical information (cancer stage) is used to decide which regimen is likely to work best, only data referring to outcomes of larger groups of patients are currently considered. Under the umbrella of personalized medicine is genomic medicine.

Genomic medicine refers to "the use of information from genomes (from humans and other organisms) and their derivatives (RNA, proteins, and metabolites) to guide medical decision making" [13]. The discovery of patterns in gene expression data and examining a person's genome makes possible to make individualized risk predictions and treatment decisions. A patient predisposition to treatment and health states can now be characterized by their molecular information, and useful classifiers and prognostic models can be developed to more strategically make decisions.

There has been a significant improvement in sensitivity as DNA microarray technology continues to advance. DNA microarray and gene expression profiles data have made possible to understand and make new discoveries at the molecular level regarding human conditions and diseases, especially cancer [14]. However, a challenge facing this area of study is the complexity and amount data across multiple samples.

This research is motivated by the question of whether it is possible to determine which patients will more likely benefit from receiving RT as part of their cancer treatment. Clinical

decision-making regarding RT is still based on estimated overall level of tumor aggressiveness, but current decision models are not personalized for predicting the benefit from RT for a specific patient [15]. Torres-Roca developed and validated a system biology model of cellular radiosensitivity which lead to the discovery of novel radiation specific predictive biomarkers [16]. The clinical applications of this type of personalized predictive model have the potential to identify patients likely to benefit from certain treatment and determine a more effective treatment strategy.

1.3 Patient-Centered Decision Making

There has been an increasing trend in the way patients are moving from being a passive actor of their disease management process to actively making decisions regarding their treatment. It could now be expected that patients will at least give true informed consent to their treatment, if not actually making such treatment decisions themselves. Depending in the stage of the cancer, the decision of receiving a treatment is a matter of several factors and implications that influence the patient to accept or reject treatment. Further treatment may prolong life or relieve symptoms, but in some cases will not eradicate the disease. A trade off must be made between possible benefits and likely side effects [17].

It is still unclear to what extent patients are involved in their decision making and how they can resolve their personal uncertainty regarding their treatment options [18]. Kiesler, 2006 [19], reviewed studies regarding the involvement of patients in the decision making process, it was found that although a large proportion of patients want to be fully informed and actively participate in their treatment decisions with their physicians, a considerable proportion of patients prefer to have little to no detailed information about their condition or involvement in medical decisions. Moreover, this shared decision process is dynamic in the sense that it will vary depending on the patient preferences, time with condition, and stage among others.

This work is based on the idea that the decision making process should consider the individual patients preferences for which treatment, if any, should be selected. Different significant predictors for overall survival, quality of life, cost-effectiveness, and response to treatment include individual patient genomic profile factors, prognostic biomarkers, and socio-economical patient characteristics. This information can help the patient make informed decisions regarding their treatment, based on their individual preferences and personal situation.

1.4 Review of Literature

This review of the literature concentrates on decision models used to select viable treatments for patients with cancer. Databases in the area of engineering and medicine were used to search articles with publication date from 01/01/2000 until 05/01/2014: Compendex (engineering village), PubMed, Medline CSA, ScienceDirect, and Web of science. Keywords used were: (Cancer) AND [(Decision Model) OR (treatment selection)].

A large of proportion of articles found in cancer decision making focus in determining which prognostic factors and biomarkers are the most significant predictors in the assessment of different outputs (e.g. Survival, Recurrence rate and chances of metastasis). The information, criteria, methods and objectives used in the models to make the treatment selection decision are listed in Table 2.

The objectives and criteria used in cancer treatment selection models involve intrinsic trade-offs between survival and quality of life. Summers (2007) assessed trade-offs between quantity and quality of life particular to prostate cancer patients as well as among different side effects to determine which treatment would be optimal for a specific patient [20]. [21], [22], [23], [24], used a utility score and defined it as the relative value patients assign to potential health states. Utility values were obtained from interviews or the literature. Some of the

treatment complications considered include: sexual dysfunction, urinary symptoms bowel dysfunction, and death. Szumacher, 2005 [25], implemented a decision model based on patients preferences in regards to convenience of treatment plan, pain relief, overall quality of life, individual's chances of survival and out-of-pocket costs. Survival, chance of metastasis and risk of relapse are usually compared to quality of life measures: In [26] and [27] models are evaluated based on the probability of the cancer relapsing after an amount of time, and [20], [24], [27] assessed the chance of the cancer spreading to other organs as decision criteria. On the other hand, Another number of articles concentrated specifically on the cost effectiveness of various strategies [28], [29], [27]. Van Gerven, 2007 [30], focused on the maximization of patient benefit, while simultaneously minimizing the cost of treatment.

Among the methods utilized in the literature, different types of Markov decision analysis framework were the most used [20], [21], [22], [23], [29] and [30]. A Markov decision process extends a Markov chain by allowing actions and rewards to incorporate both choice and motivation, also the Markov property ensures that the future state is independent of the past state given the current state of a random process. In [28], [29], [27] decision tress and cost-effectiveness analysis as a strategy to select strategies. Multi-criteria optimization models were used in [31], [32] to find the best dose–volume histogram (DVH) values by varying the dose–volume constraints on each of the organs at risk (OARs). Other methods used include: neural networks and multivariate statistical analysis [25]. In most cases, patient's preferences are not considered in these models to make individual recommendations. Therefore, future analyses need to provide outcomes stratified by more specific risks and preferences.

The data used as inputs in the models include tumor anatomy factors, patients' characteristics, and cost estimates. Tumor anatomy is also considered using the TNM staging system in various studies [24], [28], [29], [30]. Gleason score and prostate-specific antigen

(PSA) are important input for prostate cancer treatment selection [21], [20], [22], [24]. Age is the most commonly patient factor considered in the models [21], [20], [22], [24], [30], [23], [28], [26], [25]. Other patient and health factors include: gender, race, treatment history, comorbidities and laboratory test results.

1.5 Problem Statement

Treatment decision making for cancer is complex. Every patient is unique with their own genetic traits, predisposition to side effects and preferences. The patient and clinician's subjective judgment plays a vital role in making sound treatment decisions. Furthermore, various patient-specific factors make it difficult to objectively and quantitatively compare various treatment decisions.

Radiation Therapy (RT) is the most commonly prescribed single agent in cancer therapeutics. Approximately, half of cancer patients receive RT as part of their treatment. There has been great improvement in the quality and effectiveness of RT delivery in the last years. Unfortunately, neoadjuvant CRT is not beneficial for all patients. The treatment response ranges from a pathologic complete response (pCR) to a resistance. It is reported that only 10 to 20 percent of patients with advanced rectal cancer show pCR to neoadjuvant CRT. Nowadays, patients with no response or minimum tumor response to neoadjuvant CRT before its initiation are not being identified [33].

We are entering in a new era of personalized, patient-specific care, and with the advent of low-cost individual genomic and proteomic analysis, we are on the path of employing patient's biologic data to systematically predict the best course of therapy [34]. Identifying patients that potentially could benefit from CRT and justifying a given treatment path will hopefully minimize side effects caused by the current treatment practices. This is the based premise for the work presented in this dissertation.

Table 2 Summary of cancer treatment selection models in the literature

Data Considered in Decision Models		
Tumor Anatomy	Gleason Grade	[21], [20], [22], [24],
	TNM or mass	[30], [28], [24], [29]
	PSA	[20], [24]
Patients characteristics	Age	[21], [20], [22], [24], [30], [23], [28], [26], [25]
	Gender	[30], [26], [25]
	Race	[26], [25]
	Treatment history	[30], [26]
	Comorbidities	[21]
	Laboratory results	[26]
Costs		[30], [23], [28], [29], [25], [27]
Decision Criteria		
Quality of life		[20], [22], [30], [23], [24], [25]
Patient Utility		[21], [22], [30], [23], [32]
Survival		[20], [28], [24], [29], [25]
Cost effectiveness		[23], [28], [29], [27]
Chance of metastasis		[20], [24], [27]
Risk of relapse		[26], [27]
Disutility		[20]
Tumor Response		[30]
Planning target volume (PTV)		[31], [32]
Methods		
Markov framework		[21], [20], [22], [30], [23], [29]
Cost-Effectiveness analysis		[23], [28], [29], [27]
Decision trees		[28], [29], [27]
Bayesian Networks		[30], [24]
Optimization modeling		[31], [32]
Multivariate analysis		[25]
Neural Networks		[26]

1.6 Global Research Objectives

The general objectives of this work are two-fold:

- Objective 1: Build and validate a prediction model based on the gene expression profiles of a sample of cell lines for the response of a patient to RT (Radiosensitivity) using their genomic information.
- Objective 2: Integrate measures of the patient's clinical information: survival, biological characteristics and anticipated adverse effects into a patient-centered prescriptive model that determines the most appropriate course of action at a given stage (II and III) for rectal cancer.

1.7 Document Organization

This dissertation is organized in four chapters (See Figure 3). Chapter 1 presents a review of the literature, defines the problem, and presents the objectives and hypotheses of this research. Chapter 2 presents a prediction model of radiosensitivity of cancer tumor cells in response to radiation therapy using gene expression profiles; in Chapter 3, a fuzzy approach for treatment selection in cancer treatment is developed considering various criteria; and Chapter 4 presents the conclusions, limitations and opportunities future research.

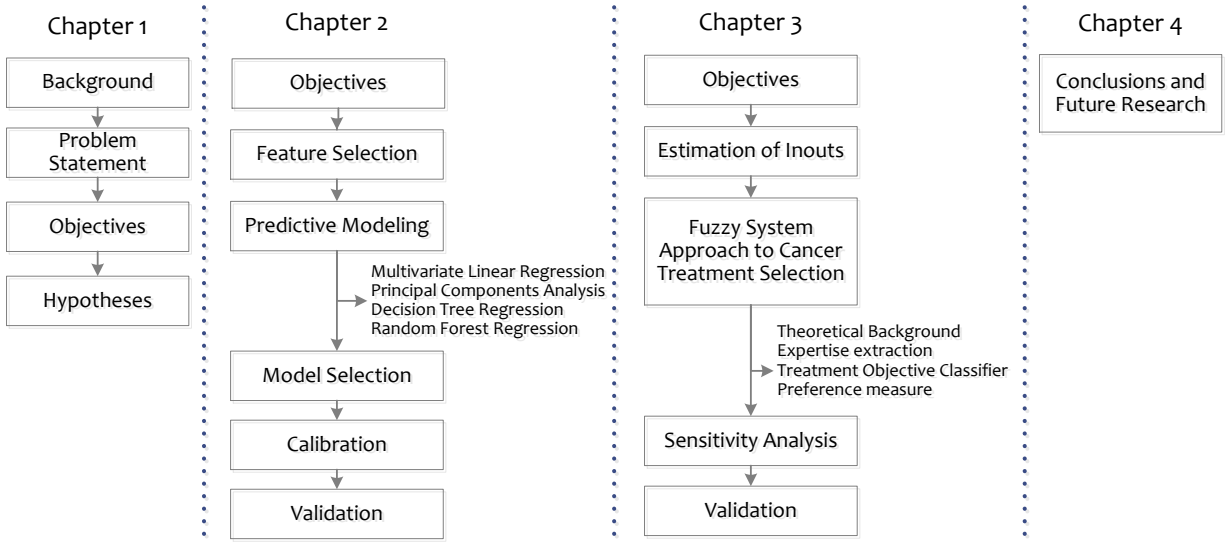


Figure 3 Dissertation organization

CHAPTER 2: PREDICTION OF RADIOSENSITIVITY OF CANCER TUMOR CELLS IN RESPONSE TO RADIATION THERAPY USING GENE EXPRESSION PROFILES

2.1 Introduction

Radiation therapy (RT) is the most commonly prescribed cancer treatment and can be effective in curing cancer. The success rates for RT are comparable with those achieved with surgery in some cancers (prostate , head and neck and cervical cancer) [35]. Over the past decades, RT effectiveness has improved by the discovery of physical approaches that optimizes the radiation dose to tumors and space normal tissues. With the introduction of microarrays and the use of gene expression to identify features in medical outcomes, identification of gene signatures and pathways activated in the response of cells to radiation can result in the development of treatment options which gene expression is controlled within the irradiated tumor (e.g. BUdR and IUdR were among the first classes of biological agents analyzed as radiosensitizers to enhance the effects of radiotherapy treatment) [36].

Decision making and treatment selection in radiation oncology is subjective and based on clinic-pathological features of a large group of patient outcomes [16]. In personalized medicine, the objective is to select the most appropriate course of treatment that fits an individual patient's needs and characteristics. Genomic medicine technological advancements has now the potential of predicting a patient predisposition to RT. Microarrays technology is one of the most widely adopted methods of genomics analyses. Microarrays experiments generate

functional data on a genome-wide scale, and can provide important data for biological interpretation of genes and their functions [37].

The complexity and dimensionality of the data generated from gene expression microarray technology requires advanced computational approaches. Machine learning and supervised learning methods provide tools to develop predictive models from available data, and it is effective when dealing with large amounts of biological data. In this dissertation, we present a methodology to organize and analyze gene expression data and test whether it results in an accurate predictive model of tumor radiosensitivity.

Machine learning refers to the type of computational techniques that are used to develop a "model" from a set of observations of a system. The term "model" assumes that there exists an approximate relationships between the parameters considered in the system. The goal is to predict a quantitative (regression) or qualitative (classification) outcome using a set of attributes or features [38]. Consequently, supervised learning refers to the subset of machine learning methods where the input–output relationship is assumed to be known.

Supervised learning is commonly used in the computational biology area ranging from gene expression data to analysis of interactions between biological subjects [38]. Some of the most commonly used supervised learning methods used in computational biology include: neural networks, support vector machine, logistic regression, multivariate linear regression, decision tree-based models and ensembles (random forest). A review of these methods is presented in the following section.

This chapter consists on the development of a personalized diagnostic tool to predict radiotherapy (RT) efficacy using the patient genomic information and estimate likelihood of response to RT of an individual patient. In the next chapter, the results of this model will be

implemented into a decision model with the objective of guiding the patient and physician decision on the selection of a cancer treatment strategy.

2.2 Review of Prediction Models in Computational Biology

A summary of the methods, relevant literature, strengths, limitations and opportunities are presented in Table 3. Methods used in prediction models for various areas of computational biology were categorized into: artificial neural networks; support vector machines; decision tree-based methods; and logistic regression.

Artificial neural networks (ANN) and support vector machines are among the most commonly used black box machine learning tools in the literature. ANN-based approaches may be applied for classification, predictive modelling and biomarker identification within data sets of high complexity [39]. More recent studies using ANN approaches in system biology include: a validated a reduced (from 70 to 9 genes) gene signature capable of accurately predicting distant metastases by Lancashire et al [40]; a model to predict Parkinson's disease using micro-array gene expression data by Sateesh Babu et al [41]; and a gene expression-based model to select 20 genes that are closely related to breast cancer recurrence by Chou et al [42].

The support vector machine (SVM) algorithm consists on a hyperplane or a set of hyperplanes in a high-dimensional space, which are then used for classification or regression [43]. Support vector machines (SVM) have a number of mathematical features that make them attractive for gene expression analysis due to its ability of dealing with large data sets with high data dimensionality, ability to identify outliers, flexibility in choosing a similarity function and sparseness of the solution [44]. According to Statnikov et al, multi-category SVM are the most effective classifiers in performing accurate cancer diagnosis using gene expression data [45]. However, most studies find that the main limitations of SVM are the lack of interpretability of

the results and estimates for the underlying probability, and the heuristic determination of the Kernel parameters.

Table 3 Summary of prediction models in computational biology

Method	Relevant Literature	Advantages	Limitations (L) Opportunities (O)
Artificial neural networks	[40]–[42], [46]–[50]	• Can process data containing non-linear relationships and interactions • Can handle noisy or incomplete data • Capable of feature selection in high dimensional data • Good predictive performance	• (L)Hard to interpret (O) Sensitivity analysis and rule extraction can be used extract knowledge • (L) Prone to over-fitting (O) re-sampling and cross-validation can be used to address this issue • (L) Multiple solutions associated with local minima
Support vector machines and kernels	[44], [45], [51]–[54]	• Can process data containing non-linear relationships and interactions • Can provide a good out-of-sample generalization • Optimality problem is convex	• (L) Large margin classifiers are known to be sensitive to the way features are scaled (O) data normalization • (L) sensitive to unbalanced data (O) assign a different misclassification cost to each class • (L) Kernel parameters are data-dependent (O) Try a linear and a non-linear kernel • (L) Prone to over-fitting (O) Local alignment kernel
Decision tree-based methods and Random forest	[55]–[64]	• Readily understandable Interpretable • Ability to rank the attributes according to their relevance in predicting the output	• (L) Classification performance of a single tree lower than other methods (O1) Classification performance could be improved by combining more than two features at each node (O2) Classification performance is improved by aggregation of predictions by ensembles • (L) Decision trees are sensitive to the training data set used and overfitting (O) Random forest use bootstrapping to estimate outcomes by aggregation of difference trees • (L) Inadequate to perform regression of continuous values (O) Tree ensembles use a large number of tree to obtained aggregated solutions and good performance
Logistic regression	[65]–[74]	• Most commonly used method in classifications problems • Often used as benchmark to compare models • Can handle nonlinear effect, interaction effect and power terms • Readily understandable Interpretable	• (L) LR can only be used to predict discrete functions • (L) Parameter estimation procedure of LR assumes an adequate number of samples for each combination of independent variables (O) Needs to make sure a large sample size and determine adequate number of samples for each combination • (L) Independent binary variable must be balanced (O) Resample the available data to obtain a balanced dataset

In models using logistic regression for classification, the outcome of interest is assumed to be binomially distributed with the logistic function $f(y) = 1/(1+\exp^{-y})$. The variable y is a measure of the contributions of the parameters $y = \beta_0+\beta_1x_1+...+\beta_nx_n$, where β_0 is a constant term and the β_1, β_2, ..., β_n are regression coefficients. Zhu and Hastie [69] present a summary of the implementation of a penalized logistic regression (PRL) model and an algorithm using univariate ranking (UR) and recursive feature elimination (RFE) to select a fewer genes than other methods. Among the extensions of logistic regression models, Shevade et al. [70] implements a sparse logistic model to suggest a gene selection method that is efficient and can be applied to identify marker genes. Finally, Chen et al [74] conducts a review on variations of logic regression: logic feature selection, Monte-Carlo logic regression, genetic programming for association studies, and modified logic regression-gene Expression Programming, and evaluates the performance of each method using genotype data.

The origin of tree-based learning methods is often credited to Hunt [75], but the method became recognized in the field of statistics by Breiman et al. [76] with the Classification And Regression Trees (CART). Since then, more decision-tree based methods have been proposed to improve the prediction accuracy by aggregating the predictions given by several decision trees for the same outcome. Although decision tree models were originally designed to address classification problems, they have been extended to handle Univariate and multivariate regression. Random forests (RF) models [77] is a randomization method that modifies the node splitting of the CART procedure as follows: at each node, K candidate variables are selected at random among all input candidate variables, an optimal candidate test is found for each of these variables, and the best test among them is eventually selected to split the node [78].

This study develops and compares a number of supervised learning methods appropriate to the structure and objectives of the models. Based on the performance of the models, a prediction model trained in tumor cell gene expression data is validated in two independent clinical outcomes datasets for patients that received pre-operative RT.

2.3 Objectives

The objective of this research study is to predict radiation sensitivity (Radiosensitivity), defined based on cellular clonogenic survival after 2 Gy (SF2) for 48 cell lines (see Table 4), and estimates as in equation (1). Since gene expression profiles are available for all cell lines, gene expression is used as the basis of the prediction model.

$$SF2 = \frac{number\ of\ colonies}{total\ number\ of\ cells\ plated\ \times platting\ efficiency} \quad (1)$$

- Hypothesis 1**:** A radiosensitivity cell-based prediction model can be validated using clinical patient data from rectal and esophagus cancer patients that received RT before surgery.
- Hypothesis 2**:** A radiosensitivity genomic-based prediction model could identify patients with rectal cancer that may benefit from RT treatment by assigning higher values of SF2 to radio-resistant patients and lower values of SF2 to radio-sensitive patients.

Radiosensitivity is defined based on cellular clonogenic survival after 2 Gy (SF2) for 48 cell lines. Since gene expression profiles are available for all cell lines, gene expression is used as the basis of the prediction model. Radiosensitivity prediction has been studied by [16], [79] where a clinically validated radiosensitivity index (RSI) has been defined to estimate radiosensitivity. The proposed approach differs from [16], [79] the response SF2 transformation process and in the gene expression selection process, using a statistically procedure versus a biological feature selection approach.

2.4 Methods and Materials

Cell lines are used to construct the prediction model and were obtained from the NCI [35]. Cells were cultured as recommended by the NCI in Roswell Park Memorial Institute medium (RPMI) 1640 supplemented with glutamine (2 mmol/L), antibiotics (penicillin/ streptomycin, 10 units/mL) and heat-inactivated fetal bovine serum (10%) at 37ºC with an atmosphere of 5% CO2.

Analyses using microarrays technology has been widely adopted for generating gene expression data on a genomic scale. Gene expression profiles were from obtained from Affymetrix U133plus chips [80] from a previously published study by Eschrich, 2009 [81].

2.4.1 Output

A transformation function (equation 2) is applied to the SF2. Originally SF ranges between 0 and 1; with the transformation functions, SF2 can range between -∞ and ∞. The objective of this transformation is to enhance the extremes values of SF2 (radio-sensitive and radio-resistant responses). The transformation follows equation 2 and represented in Figure 4:

$$T_{SF2} = \frac{1}{1 - SF2} - \frac{1}{SF2} \tag{2}$$

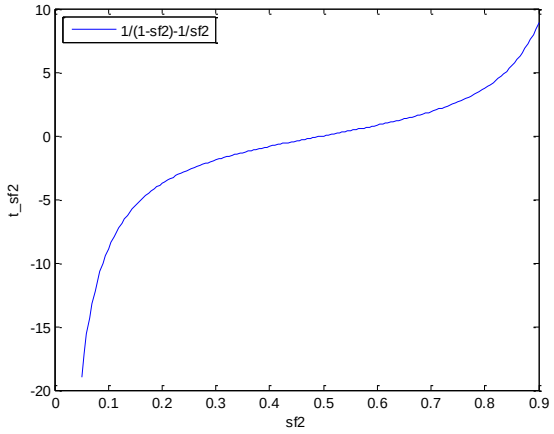


Figure 4 SF2 and transformed SF2

The survival fraction at 2 Gy (SF2) of 48 human cancer cell lines used in the regression model was obtained from Torres-Roca, 2005 [35] and are presented in Table 4.

Table 4 SF2 measured values for 48 cell lines in the database

Cell Line	Tissue of Origin	Measured SF2	Cell Line	Tissue of Origin	Measured SF2
Breast_bt549	Breast	0.632	Leuk_ccrfcem	Leukemia	0.185
Breast_hs578t	Breast	0.79	Leuk_hl60	Leukemia	0.315
Breast_mcf7	Breast	0.576	Leuk_molt4	Leukemia	0.05
Breast_mdamb231	Breast	0.82	Melan_loximvi	Melanoma	0.68
Breast_t47d	Breast	0.52	Melan_m14	Melanoma	0.42
Breast_mdamb435	Breast	0.1795	Melan_malme3m	Melanoma	0.8
Cns_sf268	CNS	0.45	Melan_skmel2	Melanoma	0.66
Cns_sf539	CNS	0.82	Melan_skmel28	Melanoma	0.74
Cns_snb19	CNS	0.43	Melan_skmel5	Melanoma	0.72
Cns_snb75	CNS	0.55	Melan_uacc257	Melanoma	0.48
Cns_u251	CNS	0.57	Melan_uacc62	Melanoma	0.52
Colon_colo205	Colon	0.69	Ovar_skov3	Ovarian	0.9
Colon_hcc-2998	Colon	0.44	Ovar_ovcar4	Ovarian	0.29
Colon_hct116	Colon	0.38	Ovar_ovcar5	Ovarian	0.408
Colon_hct15	Colon	0.4	Ovar_ovcar8	Ovarian	0.6
Colon_ht29	Colon	0.79	Ovar_ovcar3	Ovarian	0.55
Colon_km12	Colon	0.42	Prostate_du145	Prostate	0.52
Colon_sw620	Colon	0.62	Prostate_pc3	Prostate	0.484
Nsclc_a549atcc	Non-Small Cell Lung	0.61	Renal_7860	Renal	0.66
Nsclc_ekvx	Non-Small Cell Lung	0.7	Renal_a498	Renal	0.61
Nsclc_hop62	Non-Small Cell Lung	0.164	Renal_achn	Renal	0.72
Nsclc_hop92	Non-Small Cell Lung	0.43	Renal_caki1	Renal	0.37
Nsclc_ncih23	Non-Small Cell Lung	0.086	Renal_sn12c	Renal	0.62
Nsclc_h460	Non-Small Cell Lung	0.84	Renal_uo31	Renal	0.62

2.5 Feature Selection

Standard prediction models and variable reduction methods face an important challenge with the dimensionality of the data. This is the case for the area of genomic applications where the number of genes is considerably higher than the samples available to study them. In this problem, a total of m = 54,675 potential candidates (gene expression) are considered to be part of the prediction models with a total of n = 48 observations tumor cells. The most commonly used approaches, such as PCA, require for n > m. However, this problem shows m>>n. Thus, a methodology to reduce the sample size and to identify features that are statistically independent (low correlation values) is recommended. The objectives of the dimension reduction procedure presented here are to:

- Identify independent (not highly correlated) features
- Improve performance of prediction models by removing irrelevant predictors
- Improve efficiency of modeling using fewer features
- Reduce the selection of effects whose influence on dependent variable is mostly random

Our approach is an Univariate method that selects the most relevant (statistically significant) features one by one and excluding the rest, as show in [82]. This technique is computationally simple and fast to process high-dimensional datasets, and it is independent of the classification/regression models. When using this procedure, feature dependencies are ignored. Thus, a step to extract independent features has to be included (step 5 below). The procedure to select the candidate predictors include:

1. Start with 54,675 gene expressions:
2. Merge repeated gene expression by replacing with average
3. Normalize labels in datasets to create a single data file (Cell-lines have different labels in the various files)

4. Conduct response variable transformation
5. Univariate ranking: perform univariate regression with each gene versus T_SF2:
6. If (p-value >= 0.0001) then Variable is kept in the model; Otherwise, variable is excluded
7. Identify independent variables:

7.1 Estimate correlation matrix

7.2 If (correlation coefficient >= 0.9) then select gene with higher R^2 in reg for t_sf2 in cluster

7.3 Otherwise, consider this variable "independent".

8. End with the reduced data set containing 169 features (gene expressions)

The dimension reduction process presented in this study is also compared with two other feature selection methods such support vector machines. The subset of selected variables from the 54,675 gene expression probeset ID did not match previous subset selected, and selected subset was much larger with 12,399 (highly correlated) gene expression probeset IDs. Since subset of selected features was different for all methods there is no evidence to support one method over the other. The support vector machine variable selection steps used for this approach has been documented by Rakotomamonjy (2003) in the Journal of Machine Learning Research [83].

2.6 Predictive Model Development

Predictive models are developed and compared based on their performance. The experimental design of the models is presented in Figure 5. The process to build, test and validate the models has been used in the literature of supervised learning methods in computational and systems biology [38], and it can be summarized as follows:

1. Learning sample (LS) consists of 48 cell lines
2. Build model on LS using the default parameterization of the method using cross-validated: 2/3 learning sample (ls.s1), 1/3 testing sample (ls.s2)
3. Evaluate the accuracy of model on the test sample ls.s2
4. If the accuracy results are not acceptable, then play with different values of the parameter K (for random forest)
5. Select the value K* that leads to performance on S2.
6. Build selected model on LS and validate predictions on TS to get an estimate Acc_{final} of its accuracy. There are two TS datasets and will be described in the validation section.

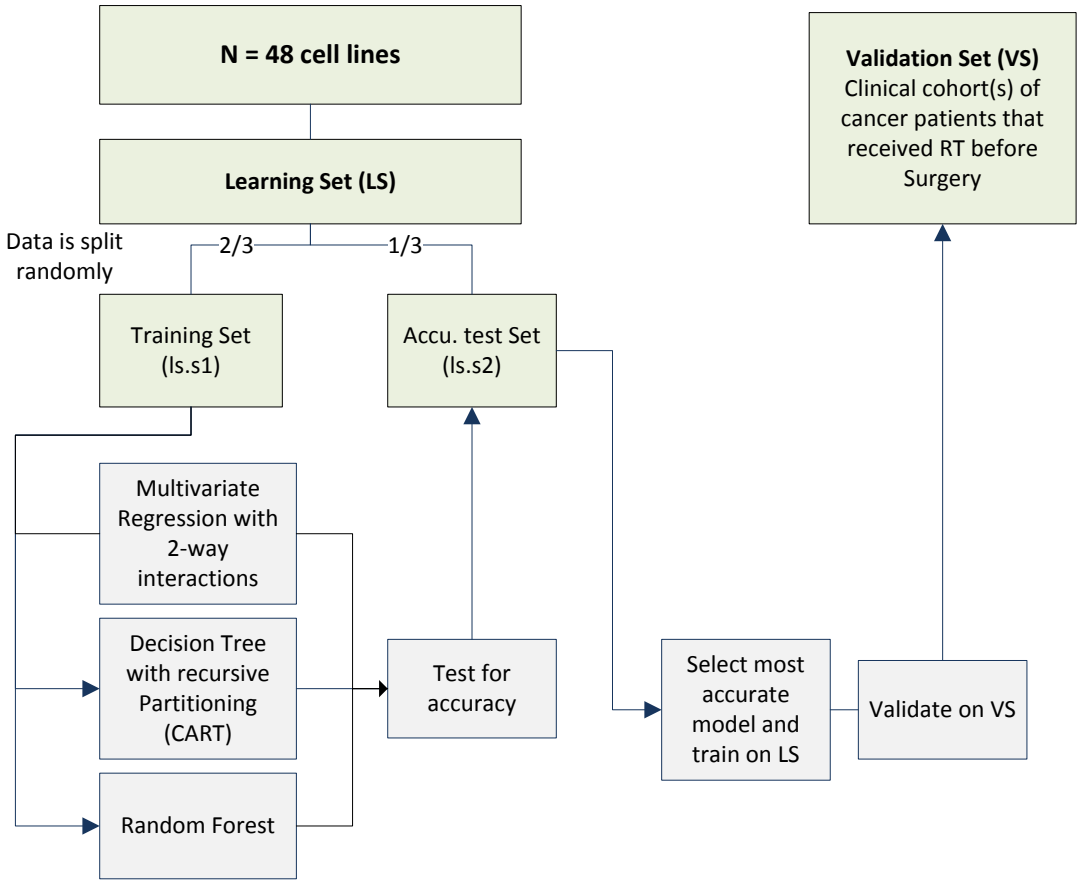


Figure 5 Experimental design

In the selection of a prediction model, there is tradeoff between simplicity and wholeness. Simpler models can be more understandable, computationally tractable. On the other hand, more complex models tend to fit the data better and to capture more information from available data. Two simple models (a Multivariate regression model and a decision tree model) and a more complex model (random forest) are created and compared to select the most appropriate model in the prediction of radiation sensitivity.

2.6.1 Multivariate Regression with 2-way Interactions

Linear regression is a method used in building models from data for which dependencies can be closely approximated [84] and predicting the value of a response (y) from a set of predictors (x_i). Let $x_1, x_2, \ldots, x_{169}$ be a set of 169 predictors believed to be associated with the transformed response T_SF2. The linear regression model for the j^{th} has the form given in (3):

$$T\_SF2_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_{169} x_{169} + \epsilon_j \quad (3)$$

The matrix notation is $\hat{y} = X\beta$ where ϵ is a random error with $E(\epsilon_j) = 0,\ Var(\epsilon_j) = \sigma^2$, $Cov(\epsilon_j, \epsilon_k) = 0\ \ \forall j \neq k$, and $\beta_i, i = 0,1,..,169$ are the regression coefficients. The approach to estimate the vector $\beta's$ in this study is the least square estimation: The value of β that minimizes the sum of square residuals $(Y - X\beta)'(Y - Z\beta)$ and the decomposition is given by (4):

$$\sum_{j=1}^{n} (y_j - \bar{y})^2 = \sum_j (\hat{y} - \bar{y})^2 + \sum_j \hat{\epsilon}^2 \quad (4)$$

The goodness of fit (GOF) of the model is measured by the proportion of the variability that the model can explain given by R^2. The formulation and motivation of the use of R^2 and other performance measures of GOR have been extensively addressed in the literature [85].

The creation of the multivariate regression model allowed for 2-way interactions to be considered as predictors in the regression model. The steps to build the models are as follows: (1) the model was coded using proc glmselect in SAS 9.3. (2) The selection process consisted

on a stepwise forward selection (effects already in the model do not necessarily stay as the fit is iteratively tested considering all candidate variables at every step). The decision criteria used considers the optimal value of the Akaike information criterion (AIC) and the adjusted R^2 to access the trade-off between the goodness of fit of the model and the penalization number of predictors in the system (overfitting). The AIC value is given by $AIC = 2k - 2ln(L)$, where k is the number of parameters and L is the value of the likelihood function.

The value of the adjusted R^2 is also presented in Figure 6. It can be observed that the value for the adjusted R^2 does not considerably improve after step 7; therefore the total number of interaction effects in the model is eight. A summary of the selection process and significant predictors' interactions, parameter estimates and performance measures (AIC and adjusted R^2) can be found in Table 5.

Table 5 Multivariate regression model selection

Step	Interaction of effects (gene expression)		Parameter estimate	Number of effects in model	adjusted R^2	AIC
0	intercept	1	58.207248	1	0	184.8924
1	222868_s	1554636_a	-1.976624	2	0.6657	133.5468
2	226367_a	244039_x_	-1.916222	3	0.7498	120.9651
3	208923_a	1557248_a	-0.187086	4	0.7967	112.4197
4	243559_a	1564276_a	1.555853	5	0.8443	101.1404
5	236687_a	1564128_a	-2.664955	6	0.8766	91.5949
6	215703_a	1557062_a	0.833148	7	0.897	84.6667
7	202252_a	238735_at	-0.132294	8	0.9112	79.3727*

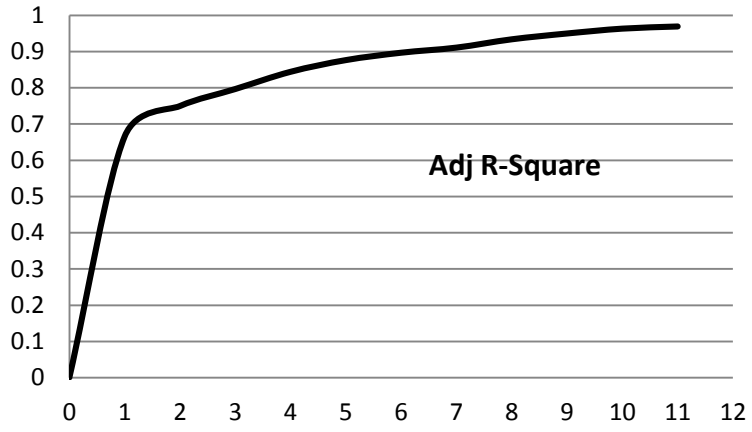


Figure 6 Model performance in terms of adjusted R-square

2.6.2 Classification and Regression Trees

The description of the decision tree methodology is included in this manuscript since it is the basis of the random forest methodology (a set of trees). A decision tree induction is a method of data analysis that maps the dependency relationships in the data [84], and it is sometimes subsumed by the category of cluster analyses. The goal with CART is to build a regression tree and predict radiosensitivity (SF2) based on the gene expression profiles available using recursive partitioning or rpart in R [86]. The following steps are followed to build the tree in rpart:

The Splitting criteria, as proven by Breiman et al [87], of a node A into two sons A_R and A_L is given by (5):

$$P(A_L)r(A_L) + P(A_R)r(A_R) \leq P(A)r(A) \quad (5)$$

where: P(A) is the probability of A for future observations, and r(A) is the risk of A. However, rpart considers measures of impurity or diversity for the note splitting criteria.

Let f be the impurity function defined by (6):

$$I(A) = \sum_{i=1}^{C} f(p_{iA}) \tag{6}$$

where p_{iA} is the proportion of the elements in A that belong to class i. Therefore, if I(A) = 0 when A is pure, f must be concave with f(0) = f(1) = 0. the split with the maximal impurity reduction (the Gini or information index) is used.

The measure of impurity can be implemented using the generalized Gini index or alert priors. This model was implemented in rpart software package in R 2.15.1, and only altered priors is available [86] for the analysis. The model building process also estimates a measure of importance for the predictors in the decision tress based on the sum of the goodness of split or adjustment agreement. This is very useful when two variables are similar and one must be selected to enter the models.

Cross-validation can be performed in decision trees using recurve partitioning. The data is divided into n groups. The model is trained in all groups except for one, the predicted class is computed, and it summed over all groups for each parameter estimate. The chosen tree will be the one with the complexity parameter with the smallest risk, computed in the full dataset.

Finally, decision trees can be built to address classification or regression problems. For regression problems, as is the case for the problem considered in this research, the splitting criterion used to decide the best split for the predictor candidates is estimated by $SS_T - (SS_L + SS_R)$, where SS_T is the sum of squares for the node, and SSR, SSL are the sums of squares for the right and left son. The decision tree model seeks to split the node in order to maximize the between groups SS in the ANOVA method. The prediction error for a new observation is estimated by $(y_{new} - y_{avg})$.

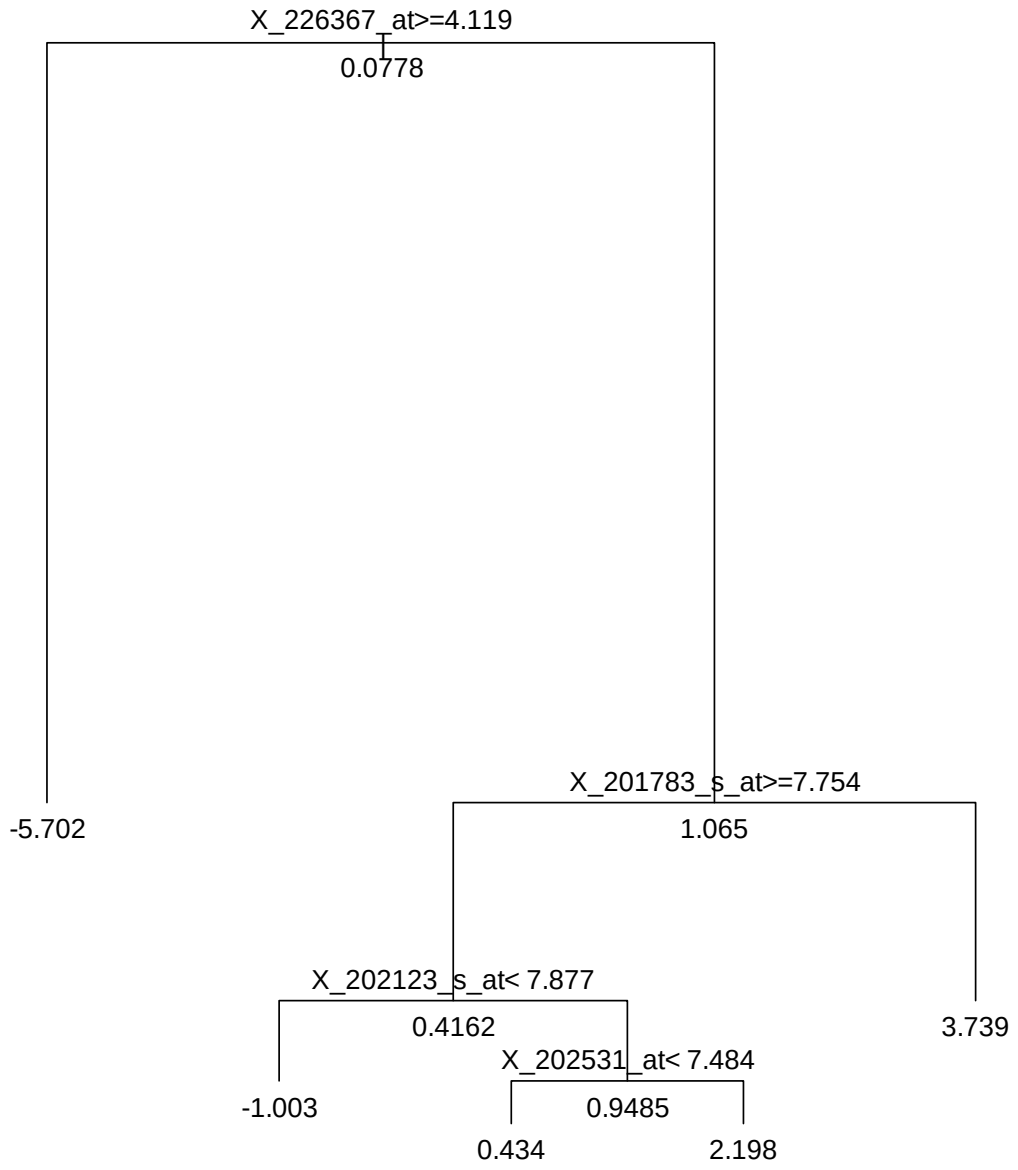


Figure 7 Decision tree prediction model

2.6.3 Random Forest

Supervised learning provides techniques to learn predictive models only from observations of a system and is therefore well suited to deal with the highly experimental nature of biological knowledge [78].

Breiman's Random Forests algorithm builds each tree from a bootstrap sample like Bagging but modifies the node splitting procedure as follows: at each test node, K attributes are

selected at random among all input attributes, an optimal candidate test is found for each of these attributes, and the best test among them is eventually selected to split the node [88].

The prediction model for radiosensitivity was built using the randomforest package in R [89]. The selected predictors (gene expression profiles), ranked in the order the variable reduced prediction error, are presented in Figure 8. The algorithm used to build the prediction model is summarized in Figure 9.

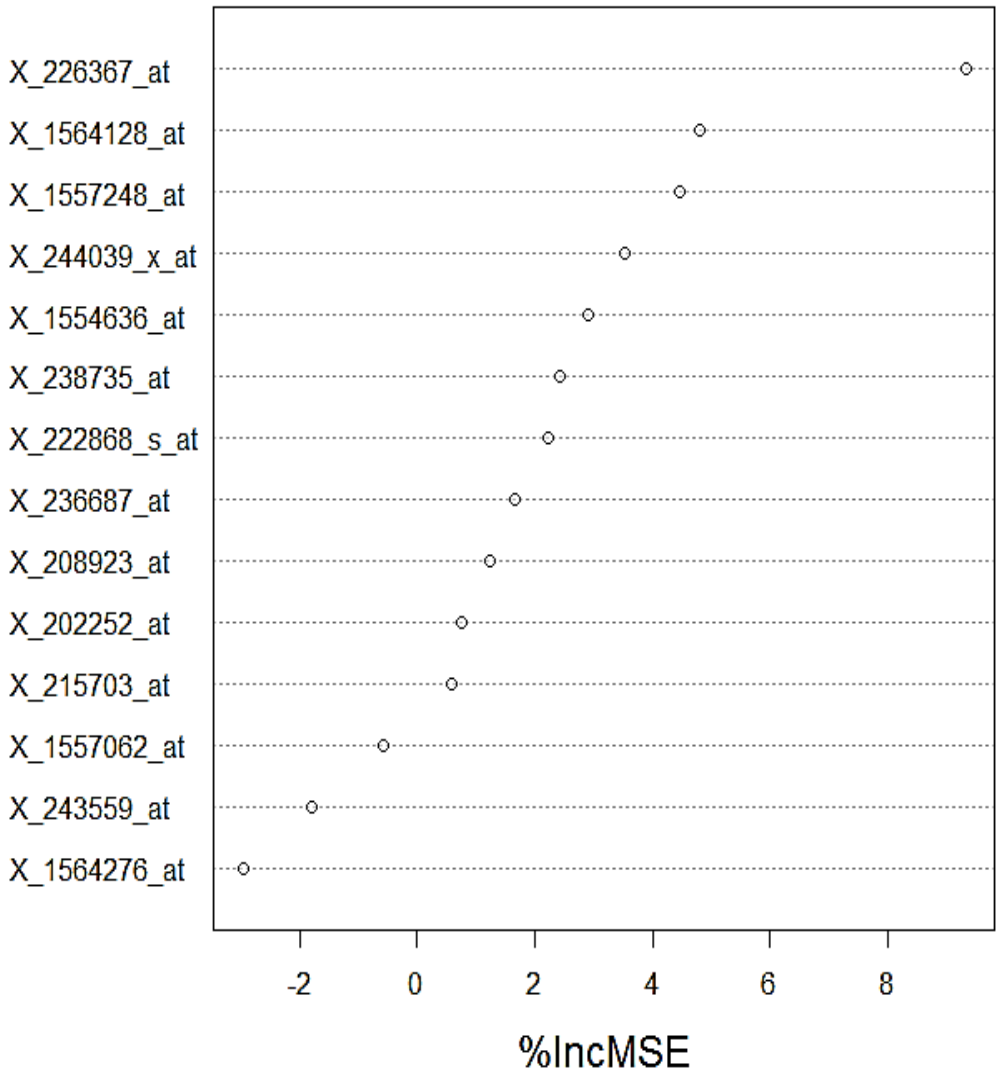


Figure 8 Variable importance based on entropy reduction

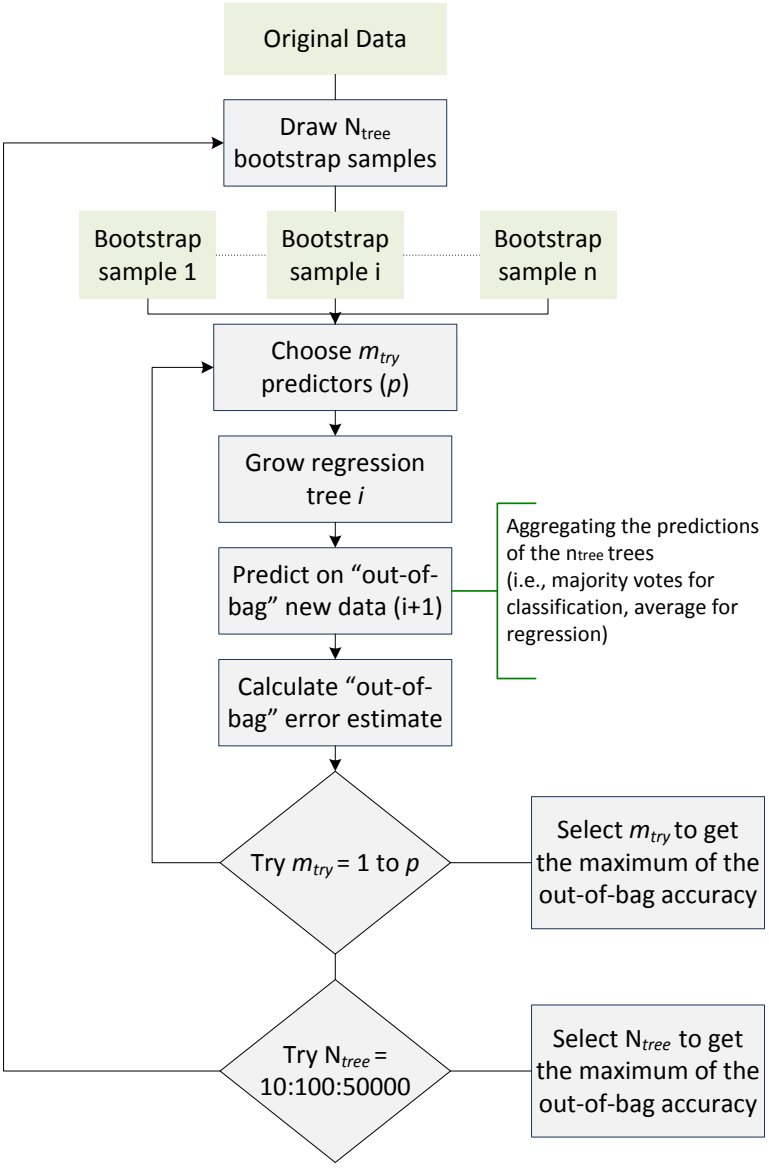


Figure 9 Random forest algorithm

2.7 Validation

The predictive models were validated in three independent datasets: a dataset of 20 patients with rectal cancer that received neoadjuvant treatment, and a dataset of 12 esophageal cancer patients that received neoadjuvant treatment. Clinical Outcomes are classified into responder(R) and non-responder (NR).

2.7.1 Rectal Cancer Dataset

The sample size consisted of 20 patients with rectal cancer. The results of the tests are: test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0185 using the random forest model and 0.003144 using regression model (See Figure 10 and Figure 11).

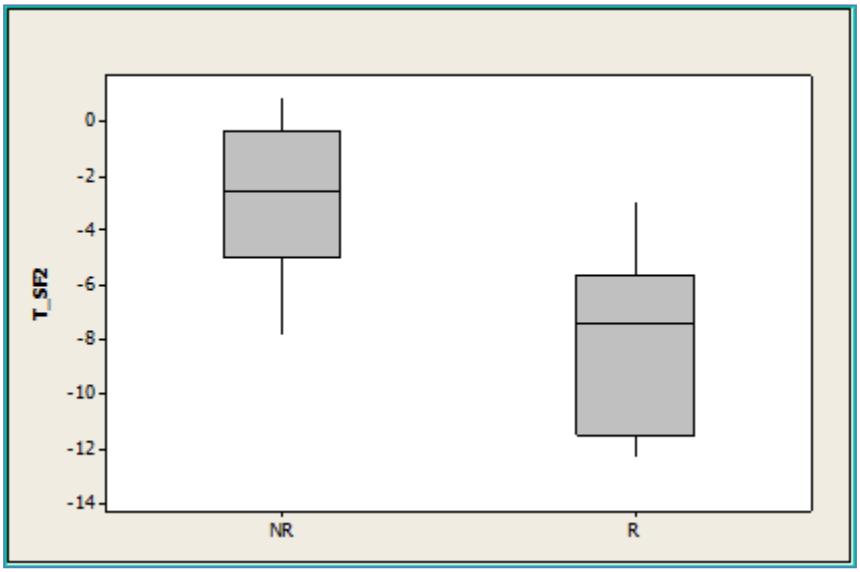


Figure 10 Multivariate regression prediction results on the rectal cancer dataset

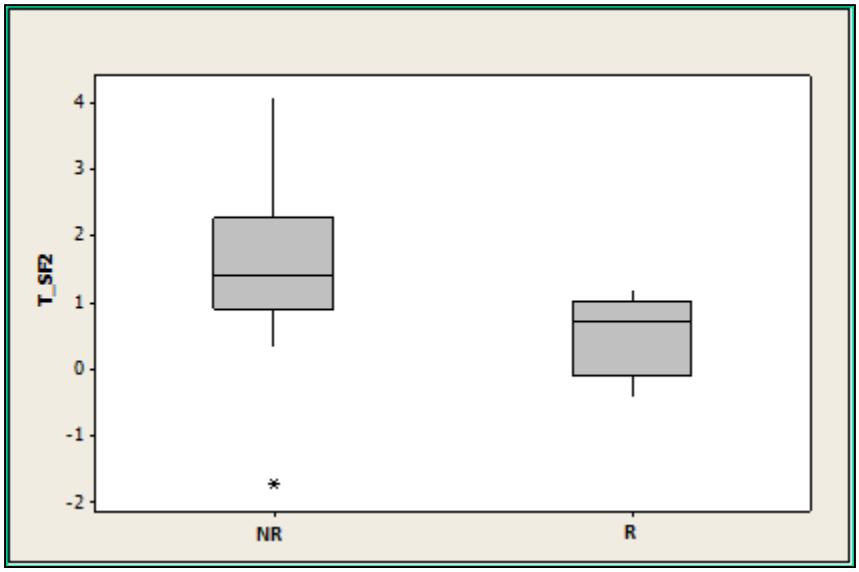


Figure 11 Random forest prediction results on the rectal cancer dataset

2.7.2 Esophageal Cancer Dataset

The sample size consisted of 12 patients with esophageal cancer. Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.026 using the random and 0.032 using regression model (See Figure 12 and Figure 13).

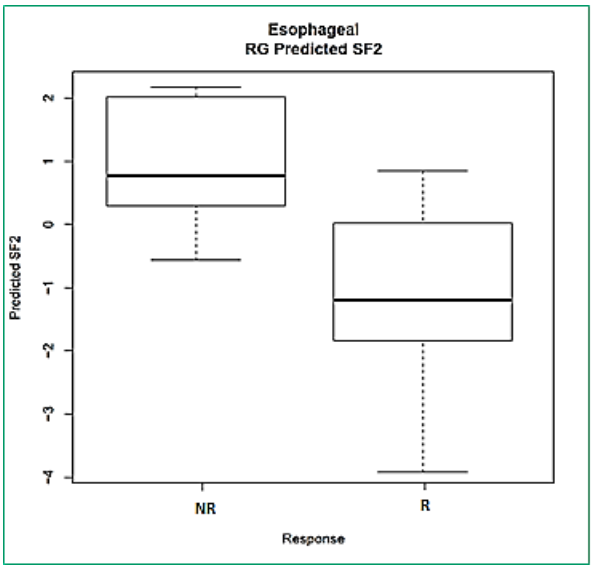


Figure 12 Multivariate regression prediction results on the esophageal cancer dataset

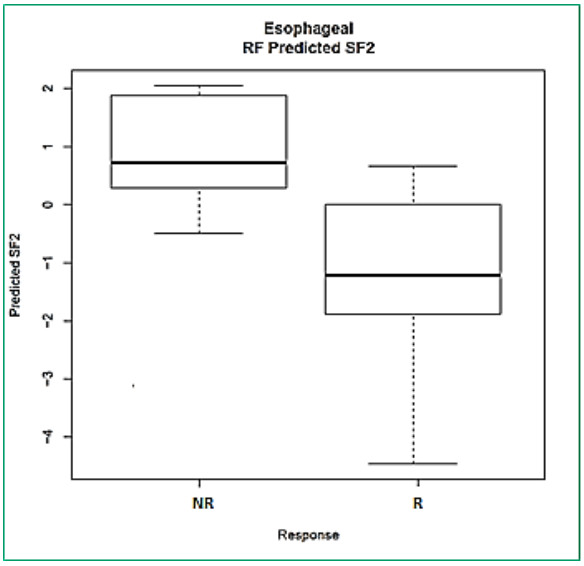


Figure 13 Random forest prediction results on the esophageal cancer dataset

2.8 Discussion

In this study, the microarray gene expression data processing and prediction model is built following four modeling parameters:

1. Response variable transformation: SF2 for 48 cancer cell lines was transformed using a mathematical function to augment the lower and upper extremes (related to Radiosensitive and Radioresistant cell lines) of the radiosensitivity/radioresistance spectrum
2. Dimensionality reduction: candidate gene expression probesets were selected using a univariate regression analysis with statistical significance ($p <= 0.001$)
3. Model building: Breiman's Random Forest algorithm [77] which is an ensemble of decision trees, was trained using the learning sample of the 48 human cancer cell lines to predict the transformed SF2
4. Model calibration: statistically significant differences ($p < 0.05$) were found between the median of the training set of the cell lines and the validation set of patients. We estimated the calibration parameters based on the calculated difference in medians.

This study provides clinical support for a practical and novel assay to predict tumor radiosensitivity. Due to the difference in experimental measurement in DNA microarray gene expression values among different cohorts, calibration methods should be created to standardize validation across different sites. Further testing of this technology in larger clinical populations is supported.

CHAPTER 3: A FUZZY APPROACH FOR TREATMENT SELECTION IN CANCER TREATMENT

The objective of this research is to develop a decision support model that can determine the most appropriate treatment strategy by combining clinical expertise and individual patient preferences concerning the treatment options available and desired clinical outcomes. The model based design and decision-making consists of a multiple-input/multiple-output (MIMO) fuzzy logic controller (FLC). For this work, we extract the knowledge from historical data, specific to colorectal cancer patients receiving radiation therapy and/or surgery from 2004 to 2010. The fuzzy system presented follows the theoretical structure presented by [90], and gets expanded by using data driven expertise acquisition and inclusion of a preference measure. The decision model and treatment strategy is evaluated using the following criteria: survival, adverse events, and efficacy. Efficacy is measured in terms of patient's response to radiation therapy or radiosensitivity (the prediction model for radiosensitivity was discussed in Chapter 2). Finally, several patient decision options are presented and compared using sensitivity analysis to present scenarios based on their individual characteristics and preferences.

This chapter is organized as follows. We first introduce basic FLC related concepts to be used throughout this dissertation involving the definitions of a fuzzy sets, fuzzy input, fuzzy output variables and fuzzy state space. Then, a review of the literature is presented that focuses on fuzzy decision support models (FDSM) is presented, followed by the research objectives and hypotheses. The FDSM approach and results are also included. Finally, the chapter closes with the sensitivity analysis, conclusions and future research.

3.1 Concepts in Fuzzy Logic

Classical sets are referred to as crisp sets in fuzzy set theory to differentiate them from fuzzy sets. A crisp set C of the universe of discourse, or domain D, can be represented by using its characteristic function μ_C:

The function $\mu_C: D \rightarrow [0,1]$ is a characteristic function of the set C if and only if for all d in (7):

$$\mu_C(d) = \begin{cases} 1 \; if \; d \in C \\ 0 \; if \; d \notin C \end{cases} \tag{7}$$

Therefore, for crisp sets every element of d either $\mathrm{d} \in \mathrm{C}$ or $\mathrm{d} \notin \mathrm{C}$. It is not the same for fuzzy sets. Given a fuzzy set F, it is not necessary that $\mathrm{d} \in \mathrm{F}$, or $\mathrm{d} \notin \mathrm{F}$. We can generalize this function to a membership function that assigns every $\mathrm{d} \in \mathrm{C}$ a value from the unit interval [0,1] instead from the two element set {0,1}.

The membership function μ_F of a fuzzy set F is a function defined as $\mu_F: D \rightarrow [0,1]$. Every element $d \in D$ has a membership degree $\mu_F(d) \in [0,1]$. Thus, the fuzzy set F is completely determined by (8):

$$F = \{(d, \mu_F(d)) \mid d \in D\} \tag{8}$$

where D and F are continuous domains, and μ_F is a continuous membership function. Figure 14 (a) and (b) shows the characteristic function of a crisp set and the membership function of a fuzzy set respectively. Support of F denoted as supp(F) refers to the elements of D that have degrees of membership to F.

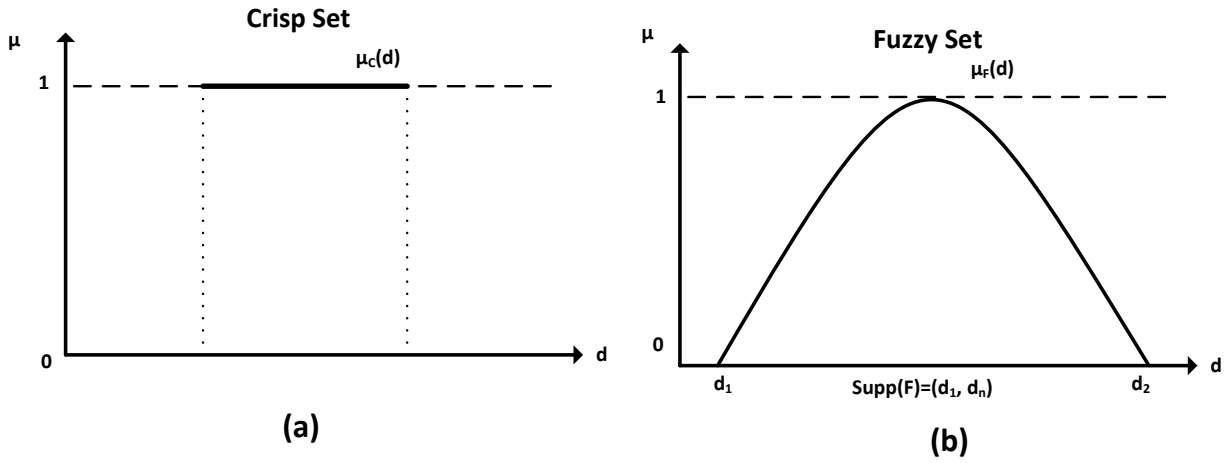


Figure 14 The characteristic function of a crisp set (a) and the membership function of a fuzzy set (b)

Throughout this document, only fuzzy sets with convex membership functions are considered. A fuzzy set F is convex if and only if:

$$\forall x, y \in X \forall \lambda \in [0,1]: \mu_A(\lambda \cdot x + (1-\lambda) \cdot y) \geq \min(\mu_A(x), \mu_A(y))$$

3.1.1 Fuzzy Inputs and Outputs

The FLC described here uses inputs and output variables whose states variables are $x_1, x_2, \cdots, x_n$. Let X be a given closed interval of real numbers, a state variable x_i with values in the fuzzy sets are fuzzy state variables, and the set of these fuzzy values are called *term-set*. The values x_i are denoted as TX_i, and the $\mathrm{j}-\mathrm{th}$ value of the $\mathrm{i}-\mathrm{th}$ fuzzy state is denoted as LX_{ij}. Each LX_{ij} is defined by the membership function in (9):

$$\mathrm{LX_{ij}} = \int_X \mu_X(x)/x \tag{9}$$

where $\mu_X(x)/x$ is the degree of membership of the crisp value x_i^* of x_i to the fuzzy value LX_{ij} of x_i (Figure 15).

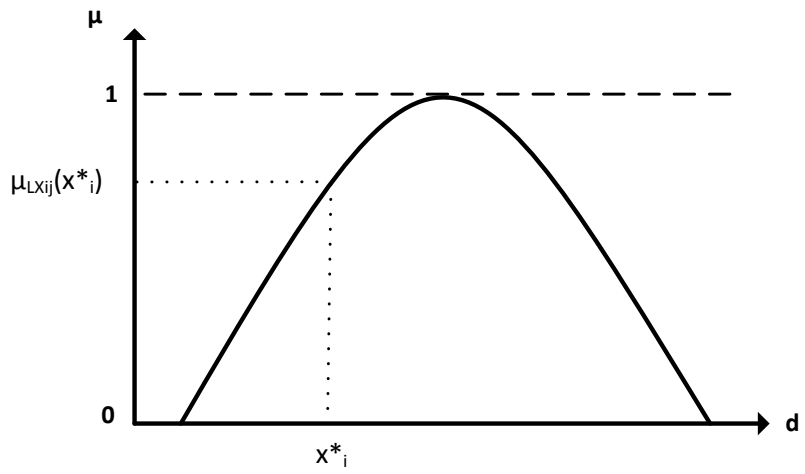


Figure 15 Degree of membership of the crisp value to the fuzzy value of the fuzzy state variable

We refer to the fuzzy values LX_{ij-1} and LX_{ij+1} as the left and right neighbor of the fuzzy value LX_{ij} respectively. Also, it is required that each fuzzy value shares a certain degree of membership with its left and right neighbors:

1. $\mathrm{supp}(\mathrm{LX_{ij-1}}) \cap \mathrm{supp}(\mathrm{LX_{ij}}) \neq \emptyset$
2. $\mathrm{supp}(\mathrm{LX_{ij}}) \cap \mathrm{supp}(\mathrm{LX_{ij+1}}) \neq \emptyset$
3. $\mu_{\mathrm{LX_{ij-1}}}(x) + \mu_{\mathrm{LX_{ij}}}(x) = 1$
4. $\mu_{\mathrm{LX_{ij}}}(x) + \mu_{\mathrm{LX_{ij+1}}}(x) = 1$

3.1.2 The Fuzzy State Space

Given a fuzzy state vector $x = (x_1, x_2, \cdots, x_n)^T$, each x_i takes some fuzzy value $\mathbf{LX_i} \in TX_i$. Therefore, a random fuzzy state vector can be written as $\mathbf{LX} = (LX_1, LX_2, \ldots, LX_n)^T$. Each fuzzy state variable takes its fuzzy values amongst the elements of a finite term-set; therefore, there is a finite number of different fuzzy state vectors, denoted as LX^i (for I = 1,2,…, M). The center of a fuzzy region, $LX^i = (LX_1^i, LX_2^i, \ldots, LX_n^i)^T$ defined by the crisp state vector $x^i = (x_1^i, x_2^i, \cdots, x_n^i)^T \in X^n$, where x_k^i are crisp values such that $\mu_{\mathrm{LX_{ij}}}(x_1^i) = 1, \mu_{\mathrm{LX_{ij}}}(x_2^i) = 1, \cdots, \mu_{\mathrm{LX_{ij}}}(x_n^i) = 1.$

The general form of a model is given as $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$, where **f** is a $n \times 1$ state vector and **u** is the $n \times 1$ input vector, and let $\mathbf{u} = \mathbf{g}(\mathbf{x})$ be the control law. Then, we can estimate the closed loop system as $\dot{\mathbf{x}} = \mathbf{f}\big(\mathbf{x}, \mathbf{g}(\mathbf{x})\big)$.

3.2 Review of Related Literature

Fuzzy logic has proven to be a reliable method for approximate reasoning [84] since it possess an easy user-interface and incorporates linguistic variables. In addition, fuzzy logic-based models can be used for non-linear, imprecise, complex systems by implementing human experience, knowledge, and practice as a set of inference rules. However, fuzzy logic also presents some challenges when dealing with decision making within probabilistic uncertainty, and the automatic inclusion of fuzzy rules.

Fuzzy set theory effectively handles the deterministic uncertainty and subjective information of clinical decision-making. Other decision-making approaches include neural networks, utility theory, statistical pattern matching, decision trees, rule-based systems, and model-based schemes. Fuzzy set theory has been successfully used alone or combined with neural networks and expert systems to solve challenging biomedical problems in practice.

In machine learning, knowledge acquisition from examples (clinical patient data) is the most common practical approach [91]. Other hybrid techniques have been used in the literature dealing with multiple objectives and/or criteria. Some of these hybrid techniques are expanded in the following paragraphs.

Fuzzy neural networks are popular due to relative ease of application. However, they generally lack insight into the decision making process and similar levels of comprehensibility. Fuzzy multi-objective decision-making (FMODM) where a fuzzy Pareto optimal solution set is provided as a final solution [92], [93], is limited in that only some specific membership functions (i.e. triangular distribution form) are used to deal with fuzzy parameters and fuzzy

goals. Also, the values of objective functions are only described by crisp values, which is sometimes not appropriate in practice since it would be preferred to deal with a range of values for the objective function.

Fuzzy decision trees or fuzzy rule-based systems have the objective to induce decision procedures with discriminative, descriptive, or taxonomic bias for classification of other samples [91]. This follows the comprehensibility principle [94] which recommends that decision procedures use language and mechanisms suitable for human interpretation and understanding. Lastly, a fuzzy discrete event system approach to determining optimal treatment regimens was developed by Ying et al, 2006 [90], in which an optimal treatment is selected after finding the maximum possible agreement among a number of experts (physicians). However, this expert system requires detailed knowledge from a group of experts and cannot be generalized.

Although fuzzy inference systems have been applied in many engineering fields, they have not been extensively applied for medical decision modeling. Models like Bayesian Decision Theory/models are appropriate for groups of patients but are complicated in application to individual patient factors.

The proposed method in this research tackles limitations currently found in the fuzzy-based models. They include (1) Decision flexibility: compared to current fuzzy rule-based models, the decision process can be dynamic, allowing the decision maker to change priorities for the rule after learning about the set of options; thus, the patient's preferences can be represented as priorities in the expert system. (2) Uncertainty: referring to the imprecision inherent in human judgments, probabilistic characteristics may be incorporated in some parameters of the decision model.

3.3 Objectives

- Objective 1: Develop an expert decision knowledge-based system that effectively depicts patient preferences and evaluate rectal cancer treatment options
- Objective 2: Integrate patient-centered measures into a decision model that considers multiple criteria

3.4 Hypotheses

- Decision procedures implemented in the model can use language and mechanisms suitable for human interpretation and understanding
- The physician and the patient can jointly use these models to compare alternative medical interventions and make a decision on choosing the most appropriate intervention for the patient.
- The decision model is capable of incorporating weights to prioritize conflictive objectives for the treatment outcomes. The decision framework allows decision makers to modify priorities for the various criteria/objectives considered to make the selection of treatments.

3.5 Fuzzy Inference System Approach

The FDSM developed in this work is applicable for the diagnostic phase in which the current or future state of a person's health status is inexact and treatment options are generally subject to predetermined clinical pathways and medical expertise. Subjective human decision making (physician's or patient's) play a significant role in defining the status of state. The status mostly likely is not crisp and neither is the transition from one state to another [90]. An appropriate representation of the inherent subjectivity and uncertainty in fields like medicine and treatment selection is provided by the fuzzy inference system theory.

The decision inputs used for the decision model are acquired using retrospective data analysis. The description of the data used and the process of rule mining are presented in section 3.5.4. The model results present the treatment of choice for stage II and stage III (no metastasis) rectal cancer patients that have not received treatment before. Three treatment regimens are considered as alternatives for the patient:

1. Surgery alone (S)
2. Radiation and Surgery, either neoadjuvant and adjuvant (RS)
3. Observation/No treatment (NT)

Decision making in cancer treatment is generally performed by the physician who will recommend a treatment based on his/her expertise. Considerations include weighting several factors to increase patient chances of survival while minimizing potential adverse effects. The essential elements of an effective cancer treatment regimen include:

1. Minimizing treatment toxicity and adverse effects -- this is measured in terms of toxicity of the treatment.
2. Selecting a treatment that can cure or eliminate the cancer tumor -- this is measured in terms of the 5 yr. Survival rate of the patient.
3. Selecting a treatment sufficiently intense increase chances of survival and reduces rate of recurrence -- this is measured in terms of radiosensitivity.

A prediction model for radiation sensitivity has been developed by Torres-Roca [15], [35] using a gene expression classifier. Since this gene classifier is not currently being used in practice, and no data is available, we will estimate the impact of this factor using a sensitivity analysis. The theoretical framework adopted for the fuzzy logic implemented here was developed by Dr. Hao Ying at al. [90]. We have expanded this methodology by the inclusion of patient preference in the decision making process and a data driven expertise acquisition

method using predictive models and clinical trials results. This study does not use the estimated clinical features intrinsic to historical treatment regimens; instead, it personalizes theses estimates using patient's characteristics. The overall methodology of this decision framework is summarized in Figure 16.

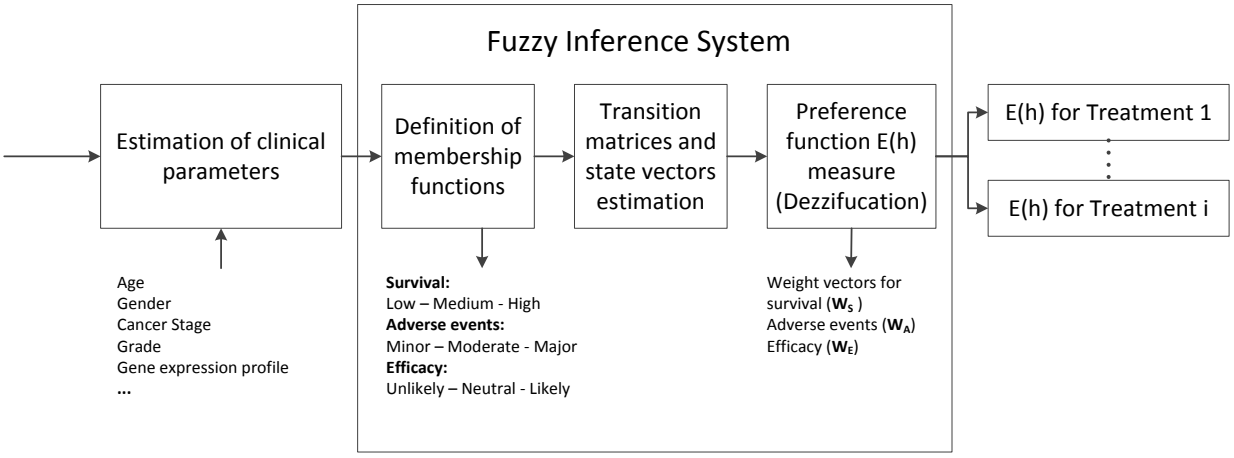


Figure 16 Fuzzy inference system approach

3.5.1 State Transitions Matrices

The treatment selection as presented in Table 6 is made considering three criteria: cause-specific survival rate (survivability), adverse events and efficacy.

The three clinical parameters have been chosen for treatment selection (other parameters can be considered, but are not in the scope of this work), and each parameter has three levels. Three fuzzy state vectors, denoted as q_1, q_2, and q_3 represent the state of survivability, adverse effects, and efficacy respectively. Survivability state vector q_1 has four components: initial, low, medium, and high, and it is a 1 × 4 vector with the initial state being represented by [1 0 0 0]. Adverse events state vector q_2 has four components first, second, and third grade, and it is a 1 × 4 vector with the initial state being represented by [1 0 0 0]. Lastly, efficacy state vector q_3 has four components unlikely, neutral, and likely, and it is 1 × 4 vector with the initial state being represented by [1 0 0 0]. Therefore, if q_3 = [0 0 0.2 0.8], it means

that the treatment is in a state with membership of 0.2 for neutral efficacy and 0.8 for likely efficacy. The model has 27 possible combinations (3x3x3=27) and 9 transition matrices for the 3 regimens.

Table 6 Decision model elements and membership functions

Decision Criteria	Category	Membership Function
Cause-specific Survival rate	Low	$\begin{cases} e^{-\frac{1}{2}\left(\frac{x-55}{5}\right)^2}, & x > 55 \\ 1, & x \leq 55 \end{cases}$
	Medium	$e^{-\frac{1}{2}\left(\frac{x-55}{6}\right)^2}, \quad -\infty < x < \infty$
	High	$\begin{cases} 1, & x > 85 \\ e^{-\frac{1}{2}\left(\frac{x-85}{5}\right)^2}, & x \leq 85 \end{cases}$
Adverse events	1st grade	$\begin{cases} e^{-\frac{1}{2}\left(\frac{x-20}{5}\right)^2}, & x > 20 \\ 1, & x \leq 20 \end{cases}$
	2nd grade	$e^{-\frac{1}{2}\left(\frac{x-30}{6}\right)^2}, \quad -\infty < x < \infty$
	3rd grade	$\begin{cases} 1, & x > 45 \\ e^{-\frac{1}{2}\left(\frac{x-45}{5}\right)^2}, & x \leq 45 \end{cases}$
Efficacy	unlikely	$\begin{cases} e^{-\frac{1}{2}\left(\frac{x-45}{5}\right)^2}, & x > 45 \\ 1, & x \leq 45 \end{cases}$
	Neutral	$e^{-\frac{1}{2}\left(\frac{x-65}{6}\right)^2}, \quad -\infty < x < \infty$
	Likely	$\begin{cases} 1, & x > 85 \\ e^{-\frac{1}{2}\left(\frac{x-85}{5}\right)^2}, & x \leq 85 \end{cases}$

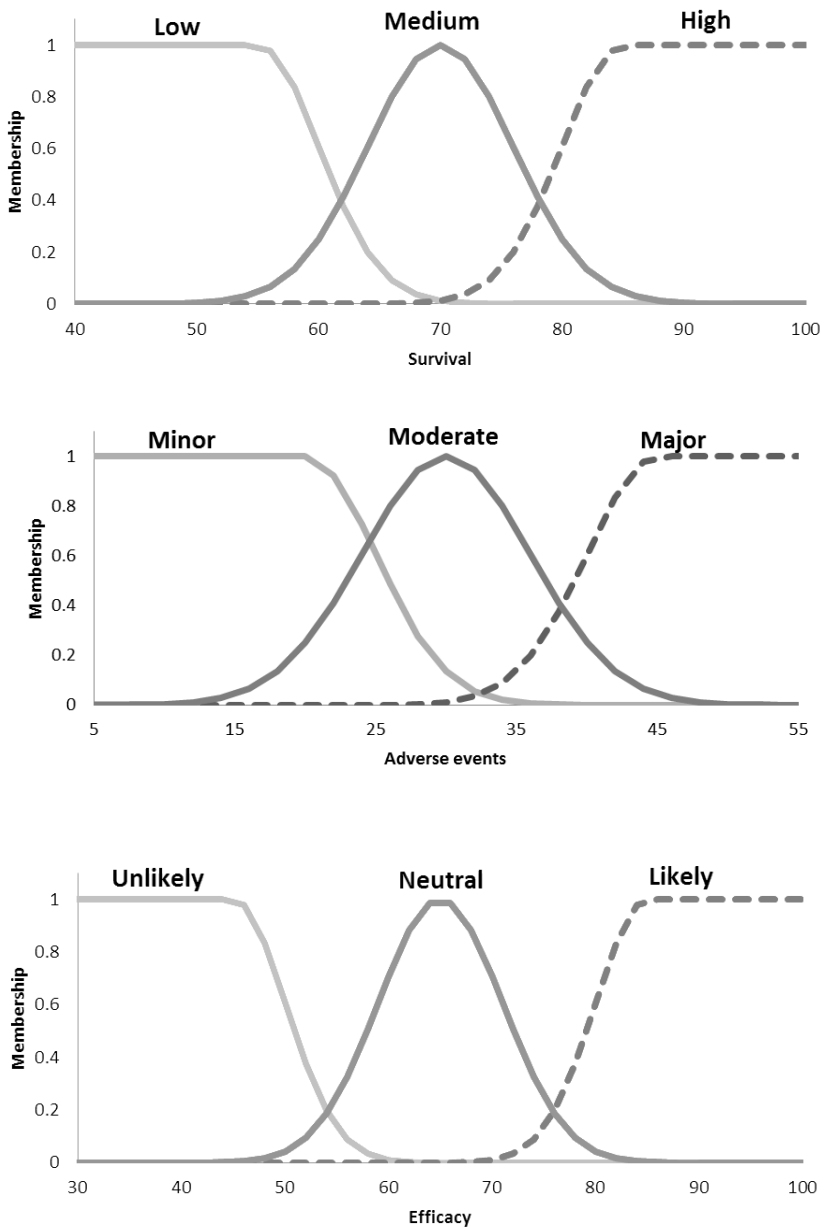


Figure 17 Membership functions in terms of survival, adverse events and efficacy

3.5.2 Membership Functions

Semi-Gaussian functions are used to produce gradual changes of membership (Table 6) and have been empirically defined based on the parameters of the data used (SEER databases) and clinical trials for survival, adverse effects (toxicity) and efficacy. These constraints are

tunable to the discretion of the analyst/decision maker. The memberships functions levels and constrains are shown in Figure 17.

3.5.3 Input Data

The Surveillance, Epidemiology, and End Results (SEER) database was used for our model survival calculations and analysis. According to the National Cancer Institute: "The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. The SEER Program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data." A signed research data agreement was approved to access these data and is included in appendix C. The variables available in this database are included in appendix D.

The data processing steps to obtain the patient cohort used in the analysis is presented in Figure 18. The demographic, tumor/cancer stage statistics and treatment options are presented in table 7, 8 and 9 respectively.

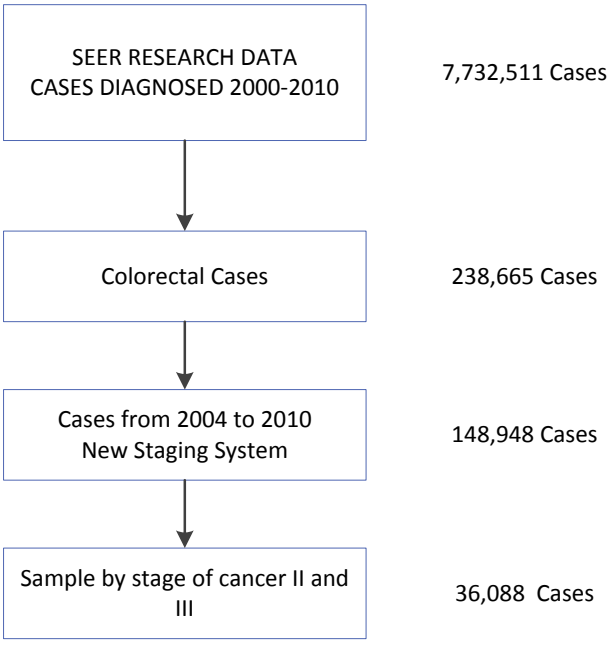


Figure 18 Pre-modeling and knowledge extraction data processing steps

Table 7 Patient cohort descriptive statistics

Patient characteristics	n	%
Gender		
Male	18442	51.1
Female	17646	48.9
Race		
White	30198	83.68
Black	4173	11.56
Other (American Indian/AK Native, Asian/Pacific Islander)	1595	4.42
Unknown	122	0.34
Marital Status		
Married (including common law)	19588	54.28
Single (never married)	4504	12.48
Widowed	7110	19.7
Divorced	3151	8.73
Unknown	1346	3.73
Separated	389	1.08
Age		
10-14	1	0
15-19	17	0.05
20-24	47	0.13
25-29	112	0.31
30-34	262	0.73
35-39	515	1.43
40-44	977	2.71
45-49	1890	5.24
50-54	2905	8.05
55-59	3451	9.56
60-64	3991	11.06
65-69	4433	12.28
70-74	4553	12.62
75-79	4648	12.88
80-84	4230	11.72
85+	4056	11.24

Table 8 Cancer and tumor stage statistics

Cancer State	n	%
Derived AJCC Stage Group, 6th ed (2004+)		
Stage IIA	15346	42.52
Stage IIB	2588	7.17
Stage IIIA	2247	6.23
Stage IIIB	9560	26.49
Stage IIIC	6284	17.41
Stage III NOS	63	0.17
Grade		
I	2409	6.68
II	24603	68.18
III	6917	19.17
IV	815	2.26
Cell type not determined	1344	3.72

Table 9 Treatment options

Treatment	n	%
Procedure		
No treatment (not recommended or refused)	551	1.53
Surgery	29872	82.78
Radiation and Surgery (neoadjuvant or adjuvant)	5220	14.47
Radiation alone	445	1.23
Reporting Source		
Hospital inpatient	35432	98.18
Radiation Treatment Centers or Medical Oncology Centers	120	0.33
Laboratory Only (hospital-affiliated or independent)	211	0.58
Physician's Office/Private Medical Practitioner (LMD)	176	0.49
Nursing/Convalescent Home/Hospice	1	0
Other hospital outpatient units/surgery centers	148	0.41
Insurance type		
Insured	30540	84.63
Any Medicaid	3434	9.52
Uninsured	1355	3.75
Insurance status unknown	759	2.1

Cause-specific survival (DTH_CLASS variable in SEER database) was used as the dependent variable for the development of a logistic regression model. This variable designates if the person died of cancer for cause-specific survival. DTH_CLASS = 1 if alive or dead due to other causes, and DTH_CLASS = 0 if dead due to colorectal cancer. The dataset was split into an 80% training data and 20% validation data. A stepwise process was performed to select the variables in the model (0.05 significant level for entry and exit at each step). All final variables are significant in the model as presented in Table 10. Performance was measured in terms of the area under the curve (auc) for training and validation data (auc= 0.741 in the training data and 0.71 in the validation data). The values for the parameter estimates are included in the appendix E.

Table 10 Logistic regression chi-square values for selected variables

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Treatment type	2	199.6288	<.0001
Marital Status	5	48.2899	<.0001
Race	3	39.7533	<.0001
Gender	1	13.3585	0.0003
Age	15	643.4756	<.0001
AJCC Stage	5	854.2814	<.0001
Grade	4	152.7802	<.0001
Insurance	3	48.8115	<.0001

Results from the logistic regression suggest that higher chances of survival are associated with treatments where radiation and surgery are combined (either neoadjuvant or adjuvant), compared to surgery alone. However, patients receiving no treatment were 4.76 times more likely of no surviving than receiving surgery alone. Women were 1.16 more likely to

survive than men. Patients in cancer stages 2B, 3B and 3C has the least chances of survival compared to patient in earlier stages. Finally, cancer tumor with grades 3 and 4 were 0.6 times more likely to be associated with higher survival rates. Odds ration estimates are presented in Table 11.

Table 11 Odds ratio estimates for logistic regression

Odds Ratio Estimates	
Effect	**Point Estimate**
Treatment: Observation vs Surgery	0.211
Treatment: Radiation/surgery vs Surgery	1.05
Gender: F vs M	1.16
Stage: D_AJCC_S 2B vs 2A	0.333
Stage: D_AJCC_S 3A vs 2A	0.934
Stage: D_AJCC_S 3B vs 2A	0.425
Stage: D_AJCC_S 3C vs 2A	0.233
Stage: D_AJCC_S 3N vs 2A	0.483
Grade: 2 vs 1	0.997
Grade: 3 vs 1	0.597
Grade: 4 vs 1	0.589
Grade: N vs 1	0.69
Insurance: Medicaid vs Insured	0.677
Insurance: Uninsured vs Insured	0.843
Insurance: Unknown vs Insured	1.301

In this study adverse events are measured by the Toxicity. According to the NCI, Toxicity grade ranges from one to five, where 1 = Mild, with no or mild symptoms; no interventions required; 2 = Moderate side-effects; 3 = Severe but not life-threatening; limitation of patient's ability to care for him/herself; 4 = Life Threatening or Disabling side-

effects; 5 = Death related to adverse event. For example, the acute morbidity criteria used to grade toxicity from radiation therapy in the gastro-intestinal area is included in Table 12.

Table 12 Criteria used to grade toxicity from radiation therapy. Content from the RTOG Acute Radiation Morbidity Scoring Criteria.

	1	2	3	4	5
Lower gastro-intestinal.	No change	• Increased frequency or change in quality of bowel habits not requiring medication • Rectal discomfort not requiring analgesics	• Diarrhea requiring parasympatholytic drugs • Mucous discharge not necessitating sanitary pads • Rectal or abdominal pain requiring analgesics	• Diarrhea requiring parenteral support • Severe mucous or blood discharge necessitating sanitary bags • Abdominal distention	• Acute or subacute obstruction, fistula or perforation • GI bleeding requiring transfusion • Abdominal pain or tenesmus requiring tube decompression or bowel diversion

Given any given patient, whose clinical characteristics and predicted outcomes per treatment have been estimated based on their individual characteristics, the transition matrices are calculated using the membership functions defined. For example, consider the case where the estimated/predicted clinical outcomes of one patient are given as shown in Table 13. The transition matrices are calculated by determining the degree of membership of the clinical parameters for survivability, adverse events and efficacy. For the example, the transition matrices for survivability and the three treatment options are given in Table 14.

The probabilities of transferring from the initial state (no previous cancer treatment) to the medium and high survivability state are 0.839 and 0.161 respectively (Table 14). This study does not consider patients that were previously treated for cancer, otherwise the low, medium and high row would contain non-zero probabilities. The other six transition probabilities are estimated in the same way and can be found in appendix F.

The state factors are calculated for the initial state after the first round of surgery: [max(0,0,0,0) max(0.839,0,0,0) max(0.161,0,0,0)] = [0 0.839 0.161]. Thus, the patient state after surgery is in 0.839 in a "medium survivability" state and 0.0161 in a "high survivability" state. The state vectors for all treatment options and clinical parameters are given in Table 15.

Table 13 Example of predicted patient clinical parameters

		Survivability	**Adverse Events**	**Efficacy**
Option 1	Surgery alone	75	30	60
Option 2	Radiation and Surgery	85	40	70
Option 3	Observation	50	10	10

Table 14 Survival transition matrices

	initial	low	med	high	
Surgery alone	0	0.000	0.839	0.161	initial
	0	0	0	0	low
	0	0	0	0	med
	0	0	0	0	high
	initial	low	med	high	
Radiation and Surgery	0	0.000	0.042	0.958	initial
	0	0	0	0	low
	0	0	0	0	med
	0	0	0	0	high
	initial	low	med	high	
Observation	0	0.994	0.004	0.000	initial
	0	0	0	0	low
	0	0	0	0	med
	0	0	0	0	high

Table 15 State vectors for all treatment options and clinical parameters

Survivability				
	initial	low	med	high
Surgery alone	0.0000000	0.0003983	0.8389319	0.1606699
Radiation and Surgery	0.0000000	0.0000000	0.0420877	0.9579123
Observation	0.0000000	0.9936665	0.0037032	0.0000000
Adverse events				
	initial	Minor	Moderate	Major
Surgery alone	0.0000000	0.1180479	0.8722622	0.0096900
Radiation and Surgery	0.0000000	0.0003918	0.2912250	0.7083832
Observation	0.0000000	0.9722278	0.0277722	0.0000000
Efficacy				
	initial	unlikely	neutral	likely
Surgery alone	0.0000000	0.0154773	0.9845175	0.0000052
Radiation and Surgery	0.0000000	0.0000052	0.9845175	0.0154773
Observation	0.0000000	1.0000000	0.0000000	0.0000000

3.5.4 Measure of Preference

A measure of preference needs to be created for the patient or the physician to compare and select from different regimens. The function, E(h) in equation (10), is defined as the weighted average of the new state vectors:

$$E(h) = \alpha \cdot W_S + \beta \cdot W_A + \gamma \cdot W_E \tag{10}$$

where W_S, W_A and W_E are the weight vectors for survival, adverse effects and treatment efficacy. The decision maker will assign a weight factor to each clinical parameter, and based on their clinical profile, the treatment with the highest $E(h)$ will be selected. Consider the preference scenarios on Table 16 for the case given in Table 13. For instance, given the state vectors from Table 15 and preference weights in Table 16, $E(h)$ for scenario 4 and surgery as treatment option is calculated as follows:

$$E(h) = [\,0\ 0\ 0.839\ 0.161][0\ 0.2\ 0.6\ 0.1]^T + [0\ 0.118\ 0.872\ 0.010\,][0\ 1\ 0\ 0\,]^T$$

$$+ [\,0\ 0.015\ 0.985\ 0][0\ 0.2\ 0.6\ 0.1]^T = 0.279$$

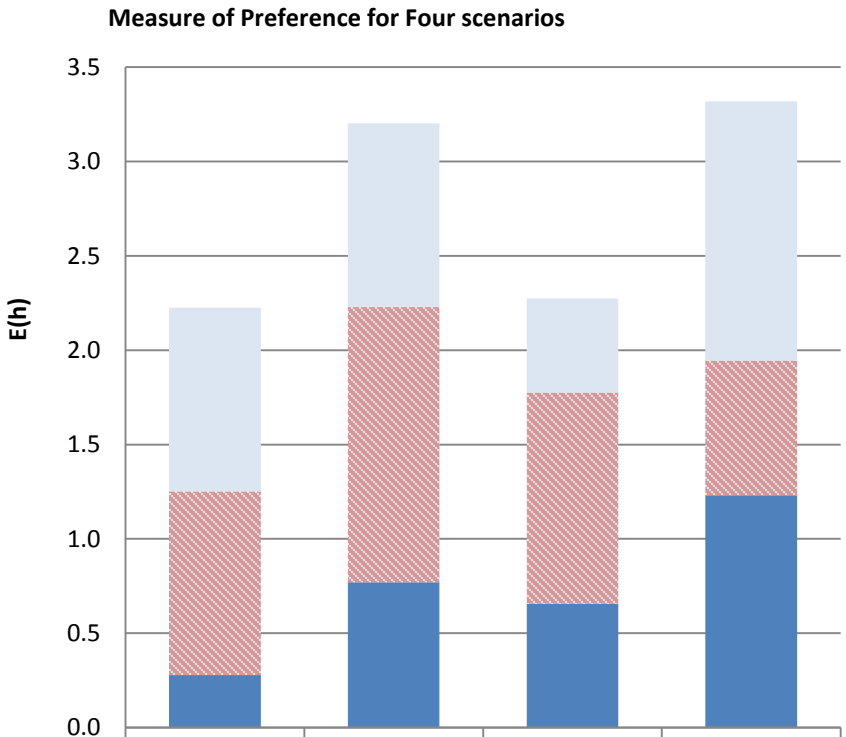
Table 16 Simulation of various preference scenarios

Scenario	W_S	W_A	W_E	Decision preference and description
1	[0 0 0 1]	[0 1 0 0]	[0 0 0 1]	Ideal scenario: all weight assigned to high survival, having minor adverse effects, and likely treatment efficacy.
2	[0 0 0 1]	[0 1 0 0]	[0 0 0.5 0.5]	All weight assigned to high survival, having minor adverse effects, and equal preference on neutral and likely treatment efficacy.
3	[0 0 0 1]	[0 0.5 0.5 0]	[0 0 0 1]	All weight assigned to high survival, having same preference on minor or moderate adverse effects, but likely treatment efficacy.
4	[0 0.25 0.6 0.15]	[0 1 0 0]	[0 0.25 0.6 0.15]	Determining treatment decision for a patient preferring to have the least adverse effects, and medium to high survival, and neutral to likely treatment efficacy.

The preference measure is similarly calculated for all treatment options and results are presented in Figure 19.

The results obtained in Figure 19 are dependent on the clinical parameters of the patient used as example (Table 13), given a different patient with the same preferences levels, these results will differ depending on the predisposition to treatment success of each patient.

In scenario 1, the two preferred treatment options are radiation plus surgery or observation (no treatment). This is a result of the patient's 100% preference of choosing a treatment with minor adverse effects, therefore a high level of preference to not perform any treatment. However, radiation and surgery have a slightly higher preference level since the patient also chose to have a treatment with high chances of survival and likely treatment efficacy.



	1	2	3	4
Observation	0.972	0.972	0.500	1.373
Radiation and Surgery	0.974	1.458	1.119	0.714
Surgery alone	0.279	0.771	0.656	1.231

Figure 19 Results of simulation of various preference profiles

In scenario 2 and 3, radiation and surgery are selected as the preferred treatment options. This is the result of having a high preference of a treatment that has high survival chances, minor to moderate adverse effects and neutral to likely efficacy.

In scenario 4, observation was chosen as the treatment choice for a patient that strongly prefers a treatment with minor adverse effects, moderate survival chances and treatment efficacy.

Other scenarios and treatment selections based on minor adverse effects and maximum survival are presented in Figure 20 and Figure 21.

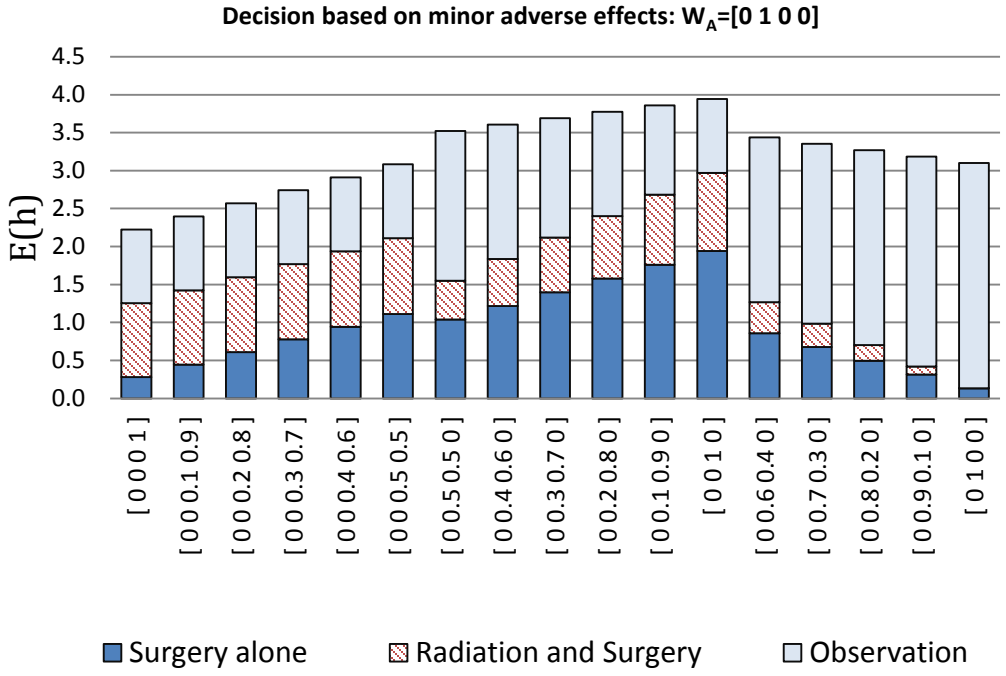


Figure 20 Sensitivity analysis based for survival

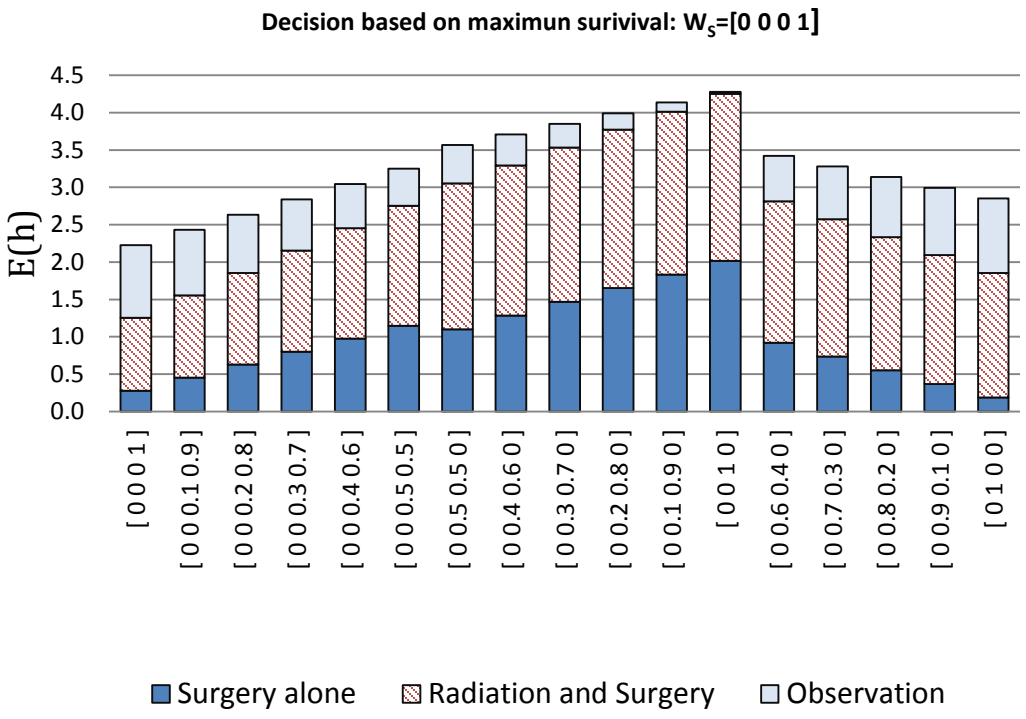
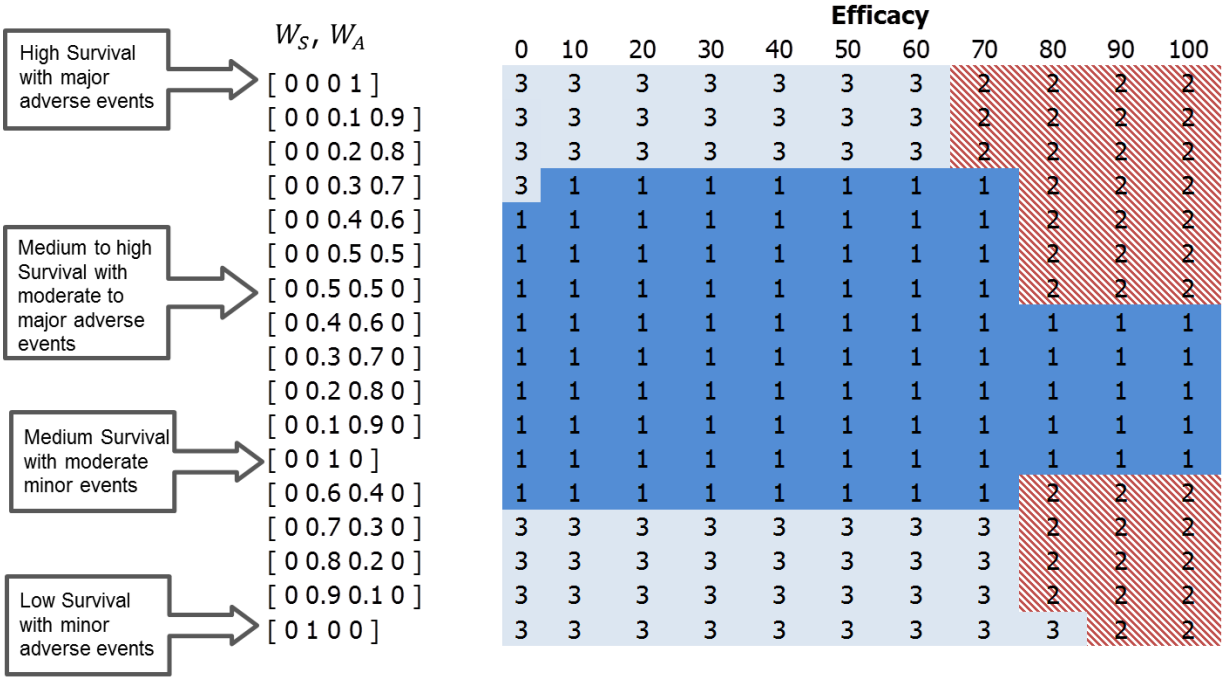


Figure 21 Sensitivity analysis based for efficacy

3.5.5 Sensitivity Analysis for Radiosensitivity

One of the main objectives of this work is to assess the inclusion of treatment efficacy in terms of radiosensitivity (chapter 2). For this assessment, a sensitivity analysis was conducted to evaluate the treatment selection change when this criterion is included in the decision making process. As treatment efficacy can have values from 0 to 100. A value of 0 represents a patient that is completely radio-resistant (therefore resistant to radiation treatment), and a value of 100 a patient is completely radiosensitive (therefore sensitive to radiation treatment).

For the example in this chapter, the analysis consisted on determining the treatment selected when the treatment efficacy (radiosensitivity) increased from 0 to 100. Results are presented in Figure 22 (1: surgery; 2: radiation and surgery; and 3: observation).



W_S, W_A	Efficacy 0	10	20	30	40	50	60	70	80	90	100
[0 0 0 1]	3	3	3	3	3	3	3	2	2	2	2
[0 0 0.1 0.9]	3	3	3	3	3	3	3	2	2	2	2
[0 0 0.2 0.8]	3	3	3	3	3	3	3	2	2	2	2
[0 0 0.3 0.7]	3	1	1	1	1	1	1	1	2	2	2
[0 0 0.4 0.6]	1	1	1	1	1	1	1	1	2	2	2
[0 0 0.5 0.5]	1	1	1	1	1	1	1	1	2	2	2
[0 0.5 0.5 0]	1	1	1	1	1	1	1	1	2	2	2
[0 0.4 0.6 0]	1	1	1	1	1	1	1	1	1	1	1
[0 0.3 0.7 0]	1	1	1	1	1	1	1	1	1	1	1
[0 0.2 0.8 0]	1	1	1	1	1	1	1	1	1	1	1
[0 0.1 0.9 0]	1	1	1	1	1	1	1	1	1	1	1
[0 0 1 0]	1	1	1	1	1	1	1	1	1	1	1
[0 0.6 0.4 0]	1	1	1	1	1	1	1	1	2	2	2
[0 0.7 0.3 0]	3	3	3	3	3	3	3	3	2	2	2
[0 0.8 0.2 0]	3	3	3	3	3	3	3	3	2	2	2
[0 0.9 0.1 0]	3	3	3	3	3	3	3	3	2	2	2
[0 1 0 0]	3	3	3	3	3	3	3	3	3	2	2

Figure 22 Sensitivity analysis for various treatment efficacy levels

3.6 Discussion

This study contributes to the arena of patient centered decision-making, using knowledge extracted from available data to guide patient and physician treatment selection. Although the expertise used in this model is acquired from current cancer practices in the United States and historical data over 6 years, the decision models need to be updated to reflect current values for cause-specific survival rate and toxicity. Knowledge is extracted from clinical trials results for toxicity for various treatments, and from the SEER data (2004-2010) from the National Cancer Institute to predict the patient cause-specific survival rate. A limitation for this study is the inclusion/updating the expertise. The updating of decision rules and other parameters can represent a complete redesign of the entire system.

Decision making is normally physician-dependent, and this study explores the inclusion of patient preferences in the decision making process. This can be of value moving forward the initiatives to make the patient the center of all process improvement and practices in the healthcare environment (PCORI mission).

At the time when this dissertation was written, radiosensitivity still has not been implemented in current medical practices, therefore, the sensitivity analysis performed to evaluate the impact of this clinical parameter is innovative and valuable. Current efforts are being made to make radiation sensitivity (radiosensitivity) to be part of current oncology practices. This would be of great benefit to better guide and customize treatment selection to each patient's individual characteristics. As it was presented in Figure 22, the selection of treatment was significantly influenced by this information. Based on the case given, radiation combined with surgery was only the preferred treatment choice when the value for radiation sensitivity was over 70%.

CHAPTER 4: CONCLUSIONS AND FUTURE RESEARCH

This research is relevant to the continuously evolving area of personalized medicine, specifically by:

- Developing decision models that allow patients to assess alternative options for treatment and make informed decisions based on their preferences and characteristics
- Advancing fundamental understanding of cancer biology and clinical oncology that can promote the prevention, detection, and treatment of cancer diseases

Genomic patient data although existent, it is rarely used in clinical settings for real-life patient care [34]. However, given the research interest and on-going research growth associated to this area, it is necessary to develop decisions models that consider individual genomic information and that are ready for adoption and transition once this information becomes readily available.

Based on the current limitations that found applicable to fuzzy decision frameworks, this research can potentially be of transformative nature. Specifically the intellectual merit can be summarized as follows:

1. Interdisciplinary research by integrating concepts from the fields of artificial intelligence, medicine, biology, biostatistics, economics, and mathematical programming to develop a decision aid approach whose solution are beyond the scope of a single area of research practice, and with an expected high practical value
2. Solution approach that is specific to modeling doctor's expertise and human preferences in the evaluation of alternatives

3. The model integrated criteria is relevant to cancer treatment selection, but it can be applicable to other scenarios where conflictive objectives are being considered
4. Comprehensibility principle: the decision model allows the use of language and mechanisms suitable for human interpretation and understanding. The fuzzy component allows us to capture concepts with graduated characteristics

4.1 Conclusions

This study provides clinical support for a practical and novel assay to predict tumor radiosensitivity. Due to the difference in experimental measurement in DNA microarray gene expression values among different cohorts, calibration methods should be created to standardize validation across different sites. Further testing of this technology in larger clinical populations is supported.

The proposed method in this research approaches limitations currently found in the fuzzy-based models: (1) Decision flexibility: compared to current fuzzy rule-based models, the decision process for this approach can be dynamic, allowing the decision maker to change priorities for the rule and be presented with a set of options; thus, the patient's preferences can be represented as priorities in the expert system. (2) Uncertainty: referring to the imprecision inherent in human judgments, uncertainty may be incorporated in some parameters of decision model.

4.2 Future Research

The predictive models developed in this research are predicting radiosensitivity with acceptable performance. These models are also capable to discriminating between responders and non-responders when the models were validated against clinical data. However, a random forests model is considered a black box machine learning algorithm. This work will continue on exploring methods that can help us understand the patterns in the algorithm of how genes

interact with each other. This can be achieved by the use of sensitivity analysis that will "unmask" the functions that associate the input variables with the response variables (which is the case for random forest, support vector machines and neural networks).

The fuzzy decision framework presented in this research only considers patients that have not received cancer treatment before, but it can expanded to patients in other treatment states (e.g. after radiation therapy or patients with recurring cancers) where decision-making is more complex. Also, the decision model would benefit of further research in the clinical parameters used as input in the models; especially for adverse effects (as an indicator of quality of life). Finally, the decision models should also include confidence intervals for the clinical parameters to account for the uncertainty of the clinical estimates used.

REFERENCES

[1] National Cancer Institute, "Colon and Rectal Cancer," 2012. [Online]. Available: http://www.cancer.gov/cancertopics/types/colon-and-rectal.

[2] American Cancer Society, "What is Colorectal Cancer?," 2013. [Online]. Available: http://www.cancer.org/acs/groups/cid/documents/webcontent/003096-pdf.pdf. [Accessed: 10-Jun-2013].

[3] B. a Kohler, E. Ward, B. J. McCarthy, M. J. Schymura, L. a G. Ries, C. Eheman, A. Jemal, R. N. Anderson, U. a Ajani, and B. K. Edwards, "Annual report to the nation on the status of cancer, 1975-2007, featuring tumors of the brain and other nervous system.," *J. Natl. Cancer Inst.*, vol. 103, no. 9, pp. 714–36, May 2011.

[4] American Cancer Society, "Colorectal Cancer: Facts & Figures 2011-2013," *Atlanta Am. Cancer Soc.*, pp. 2006–2008, 2011.

[5] Surveillance Epidemiology and End Results (SEER) Program, "SEER Data, 1973-2009," 2009. [Online]. Available: http://seer.cancer.gov/data/.

[6] MD Anderson Cancer Center, "Cancer Treatment: Radiation Therapy," 2014. [Online]. Available: http://www.mdanderson.org/patient-and-cancer-information/cancer-information/cancer-topics/cancer-treatment/radiation/index.html.

[7] MD Anderson Cancer Center, "Rectal Cancer Diagnosis," 2014. [Online]. Available: http://www.mdanderson.org/patient-and-cancer-information/cancer-information/cancer-types/rectal-cancer/diagnosis/index.html.

[8] National Cancer Institute, "Rectal Cancer Treatment (PDQ®)," 2012. [Online]. Available: http://www.cancer.gov/cancertopics/pdq/treatment/rectal/Patient/page4. [Accessed: 12-Apr-2012].

[9] National Institutes of Health, "Rectal Cancer Treatment: Treatment Option Overview," 2012. [Online]. Available: www.cancer.gov/cancertopics/pdq/treatment/rectal/Patient/page4. [Accessed: 10-Jun-2013].

[10] D. B. Stewart and D. W. Dietz, "Total Mesorectal Excision : What Are We Doing ?," *Clin. Colon Rectal Surg.*, vol. 20, no. 3, pp. 190–202, 2007.

[11] American Cancer Society, "Cancer Treatment and Survivorship Facts & Figures 2012-2013," Atlanta, 2012.

[12] I. Garajová, S. Di Girolamo, F. de Rosa, J. Corbelli, V. Agostini, G. Biasco, and G. Brandi, "Neoadjuvant treatment in rectal cancer: actual status.," *Chemother. Res. Pract.*, vol. 2011, p. 839742, Jan. 2011.

[13] G. S. Ginsburg and H. F. Willard, "Genomic and personalized medicine: foundations and applications.," *Transl. Res.*, vol. 154, no. 6, pp. 277–87, Dec. 2009.

[14] S. Eschrich and T. J. Yeatman, "DNA microarrays and data analysis: an overview.," *Surgery*, vol. 136, no. 3, pp. 500–3, Sep. 2004.

[15] J. F. Torres-Roca and C. W. Stevens, "Predicting response to clinical radiotherapy: past, present, and future directions.," *Cancer Control*, vol. 15, no. 2, pp. 151–6, Apr. 2008.

[16] J. F. Torres-roca, "A molecular assay of tumor radiosensitivity: a roadmap towards biology-based personalized radiation therapy," *Per. Med.*, vol. 9, no. 5, pp. 547–557, 2012.

[17] C. M. Gaston and G. Mitchell, "Information giving and decision-making in patients with advanced cancer: a systematic review.," *Soc. Sci. Med.*, vol. 61, no. 10, pp. 2252–64, Nov. 2005.

[18] D. Stacey, L. Paquet, and R. Samant, "Exploring cancer treatment decision-making by patients: a descriptive study.," *Curr. Oncol.*, vol. 17, no. 4, pp. 85–93, Aug. 2010.

[19] D. J. Kiesler and S. M. Auerbach, "Optimal matches of patient preferences for information, decision-making and interpersonal behavior: evidence, models and interventions.," *Patient Educ. Couns.*, vol. 61, no. 3, pp. 319–41, Jun. 2006.

[20] B. D. Sommers, C. J. Beard, A. V D'Amico, D. Dahl, I. Kaplan, J. P. Richie, and R. J. Zeckhauser, "Decision analysis using individual patient preferences to determine optimal treatment for localized prostate cancer.," *Cancer*, vol. 110, no. 10, pp. 2210–7, Nov. 2007.

[21] M. W. Kattan, M. E. Cowen, and B. J. Miles, "A Decision Analysis for Treatment of Clinically Localized Prostate Cancer," *J. Gen. Intern. Med.*, vol. 12, no. 5, pp. 299–305, 1997.

[22] V. Bhatnagar, S. Stewart, W. Bonney, and R. Kaplan, "Treatment options for localized prostate cancer: quality-adjusted life years and the effects of lead-time," *Urology*, vol. 63, no. 1, pp. 103–109, Jan. 2004.

[23] A. Konski, W. Speier, A. Hanlon, J. R. Beck, and A. Pollack, "Is proton beam therapy cost effective in the treatment of adenocarcinoma of the prostate?," *J. Clin. Oncol.*, vol. 25, no. 24, pp. 3603–8, Aug. 2007.

[24] W. P. Smith, J. Doctor, I. J. Kalet, and M. H. Phillips, "A decision aid for intensity-modulated radiation- therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model," *Artif. Intell. Med.*, vol. 46, no. 1, pp. 119–130, 2009.

[25] E. Szumacher, H. Llewellyn-Thomas, E. Franssen, E. Chow, G. DeBoer, C. Danjoux, C. Hayter, E. Barnes, and L. Andersson, "Treatment of bone metastases with palliative radiotherapy: patients' treatment preferences.," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 61, no. 5, pp. 1473–81, May 2005.

[26] C. E. Pedreira, L. Macrini, M. G. Land, and E. S. Costa, "New decision support tool for treatment intensity choice in childhood acute lymphoblastic leukemia.," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 3, pp. 284–90, May 2009.

[27] M. Morelle, E. Haslé, I. Treilleux, J.-P. Michot, T. Bachelot, F. Penault-Llorca, and M.-O. Carrère, "Cost-effectiveness analysis of strategies for HER2 testing of breast cancer patients in France.," *Int. J. Technol. Assess. Health Care*, vol. 22, no. 3, pp. 396–401, Jan. 2006.

[28] D. Marshall, K. N. Simpson, C. C. Earle, and C. W. Chu, "Economic decision analysis model of screening for lung cancer.," *Eur. J. Cancer*, vol. 37, no. 14, pp. 1759–67, Sep. 2001.

[29] R. K. Khandker, J. D. Dulski, J. B. Kilpatrick, R. P. Ellis, J. B. Mitchell, and W. B. Baine, "A decision model and cost-effectiveness analysis of colorectal cancer screening and surveillance guidelines for average-risk adults.," *Int. J. Technol. Assess. Health Care*, vol. 16, no. 3, pp. 799–810, Jan. 2000.

[30] M. a J. van Gerven, F. J. Díez, B. G. Taal, and P. J. F. Lucas, "Selecting treatment strategies with dynamic limited-memory influence diagrams.," *Artif. Intell. Med.*, vol. 40, no. 3, pp. 171–86, Jul. 2007.

[31] R. R. Meyer, H. H. Zhang, L. Goadrich, D. P. Nazareth, L. Shi, and W. D. D'Souza, "A multiplan treatment-planning framework: a paradigm shift for intensity-modulated radiotherapy.," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 68, no. 4, pp. 1178–89, Jul. 2007.

[32] T. Hong, D. Craft, F. Carlsson, and T. Bortfeld, "Multicriteria Optimization in IMRT Treatment Planning for Locally Advanced Cancer of the Pancreatic Head," *Int J Radiat Oncol Biol Phys*, vol. 72, no. 4, pp. 1208–1214, 2008.

[33] I. Garajová, S. Di Girolamo, F. de Rosa, J. Corbelli, V. Agostini, G. Biasco, and G. Brandi, "Neoadjuvant treatment in rectal cancer: actual status.," *Chemother. Res. Pract.*, vol. 2011, p. 839742, Jan. 2011.

[34] S. Ely, "Personalized medicine: individualized care of cancer patients.," *Transl. Res.*, vol. 154, no. 6, pp. 303–8, Dec. 2009.

[35] J. F. Torres-Roca, S. Eschrich, H. Zhao, G. Bloom, J. Sung, S. McCarthy, A. B. Cantor, A. Scuto, C. Li, S. Zhang, R. Jove, and T. Yeatman, "Prediction of radiation sensitivity using a gene expression classifier.," *Cancer Res.*, vol. 65, no. 16, pp. 7169–76, Aug. 2005.

[36] D. Kufe and R. Weichselbaum, "Radiation therapy: activation for gene transcription and the development of genetic radiotherapy-therapeutic strategies in oncology.," *Cancer Biol. Ther.*, vol. 2, no. 4, pp. 326–9, 2003.

[37] T. Aittokallio, M. Kurki, and O. Nevalainen, "Computational Strategies for Analyzing data," *J. Bioinform. Comput. Biol.*, vol. 1, no. 3, pp. 541–586, 2003.

[38] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology Supplementary material," no. c, pp. 1–10, 2009.

[39] L. J. Lancashire, C. Lemetre, and G. R. Ball, "An introduction to artificial neural networks in bioinformatics - application to complex microarray and mass spectrometry datasets in cancer studies," *Brief. Bioinform.*, vol. 10, no. 3, pp. 315–29, May 2009.

[40] L. J. Lancashire, D. G. Powe, J. S. Reis-Filho, E. Rakha, C. Lemetre, B. Weigelt, T. M. Abdel-Fatah, a R. Green, R. Mukta, R. Blamey, E. C. Paish, R. C. Rees, I. O. Ellis, and G. R. Ball, "A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks.," *Breast Cancer Res. Treat.*, vol. 120, no. 1, pp. 83–93, Feb. 2010.

[41] G. Sateesh Babu and S. Suresh, "Parkinson's disease prediction using gene expression – A projection based learning meta-cognitive neural classifier approach," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1519–1529, Apr. 2013.

[42] H.-L. Chou, C.-T. Yao, S.-L. Su, C.-Y. Lee, K.-Y. Hu, H.-J. Terng, Y.-W. Shih, Y.-T. Chang, Y.-F. Lu, C.-W. Chang, M. L. Wahlqvist, T. Wetter, and C.-M. Chu, "Gene expression profiling of breast cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees.," *BMC Bioinformatics*, vol. 14, no. 1, p. 100, Mar. 2013.

[43] A.-M. Lahesmaa-Korpinen, *Computational approaches in high-throughput proteomics data analysis*, no. 169. 2012, pp. 3–18.

[44] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 1, pp. 262–7, Jan. 2000.

[45] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.," *Bioinformatics*, vol. 21, no. 5, pp. 631–43, Mar. 2005.

[46] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.," *Nat. Med.*, vol. 7, no. 6, pp. 673–9, Jun. 2001.

[47] N. R. Pal, K. Aguan, A. Sharma, and S. Amari, "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering.," *BMC Bioinformatics*, vol. 8, p. 5, Jan. 2007.

[48] M. C. O'Neill and L. Song, "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect.," *BMC Bioinformatics*, vol. 4, p. 13, Apr. 2003.

[49] J. S. Wei, B. T. Greer, F. Westermann, S. M. Steinberg, C. Son, Q. Chen, C. C. Whiteford, S. Bilke, A. L. Krasnoselsky, N. Cenacchi, D. Catchpoole, F. Berthold, M. Schwab, and J. Khan, "Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma.," *Cancer Res.*, vol. 64, no. 19, pp. 6883–91, Oct. 2004.

[50] a. Narayanan, E. C. Keedwell, J. Gamalielsson, and S. Tatineni, "Single-layer artificial neural networks for gene expression analysis," *Neurocomputing*, vol. 61, pp. 217–240, Oct. 2004.

[51] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology.," *PLoS Comput. Biol.*, vol. 4, no. 10, p. e1000173, Oct. 2008.

[52] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data.," *IEEE Trans. Nanobioscience*, vol. 4, no. 3, pp. 228–34, Sep. 2005.

[53] V. Bevilacqua, P. Pannarale, M. Abbrescia, C. Cava, A. Paradiso, and S. Tommasi, "Comparison of data-merging methods with SVM attribute selection and classification in breast cancer gene expression.," *BMC Bioinformatics*, vol. 13 Suppl 7, no. Suppl 7, p. S9, Jan. 2012.

[54] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines.," *BMC Syst. Biol.*, vol. 5, no. 1, p. 161, Jan. 2011.

[55] M. Hassan and R. Kotagiri, "A new approach to enhance the performance of decision tree for classifying gene expression data.," *BMC Proc.*, vol. 7, no. Suppl 7, p. S3, Dec. 2013.

[56] G. Dong and Q. Han, "Mining Accurate Shared Decision Trees from Microarray Gene Expression Data for Different Cancers."

[57] G. R. Varadhachary, Y. Spector, J. L. Abbruzzese, S. Rosenwald, H. Wang, R. Aharonov, H. R. Carlson, D. Cohen, S. Karanth, J. Macinskas, R. Lenzi, A. Chajut, T. B. Edmonston, and M. N. Raber, "Prospective gene signature study using microRNA to identify the tissue of origin in patients with carcinoma of unknown primary.," *Clin. Cancer Res.*, vol. 17, no. 12, pp. 4063–70, Jun. 2011.

[58] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Dzeroski, "Predicting gene function using hierarchical multi-label decision tree ensembles.," *BMC Bioinformatics*, vol. 11, p. 2, Jan. 2010.

[59] M. E. Ross, X. Zhou, G. Song, S. A. Shurtleff, K. Girtman, W. K. Williams, H. Liu, R. Mahfouz, S. C. Raimondi, N. Lenny, A. Patel, and J. R. Downing, "Classification of pediatric acute lymphoblastic leukemia by gene expression profiling," *Blood*, vol. 102, no. 8, pp. 2951–2959, 2003.

[60] S. Salzberg, A. L. Delcher, H. Fasman, and J. Henderson, "A Decision Tree System for Finding Genes in DNA," *J. Comput. Biol.*, vol. 5, no. 4, pp. 667–80, 1998.

[61] C. R. Williams-DeVane, D. M. Reif, E. C. Hubal, P. R. Bushel, E. E. Hudgens, J. E. Gallagher, and S. W. Edwards, "Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes.," *BMC Syst. Biol.*, vol. 7, no. 1, p. 119, Jan. 2013.

[62] J. S. Barnholtz-Sloan, X. Guan, C. Zeigler-Johnson, N. J. Meropol, and T. R. Rebbeck, "Decision tree-based modeling of androgen pathway genes and prostate cancer risk.," *Cancer Epidemiol. Biomarkers Prev.*, vol. 20, no. 6, pp. 1146–55, Jun. 2011.

[63] D. Che, Q. Liu, K. Rasheed, and X. Tao, *Software Tools and Algorithms for Biological Systems*, vol. 696. New York, NY: Springer New York, 2011, pp. 191–199.

[64] G. Stiglic, S. Kocbek, I. Pernek, and P. Kokol, "Comprehensive decision tree models in bioinformatics.," *PLoS One*, vol. 7, no. 3, p. e33812, Jan. 2012.

[65] G. J. Mann, G. M. Pupo, A. E. Campain, C. D. Carter, S.-J. Schramm, S. Pianova, S. K. Gerega, C. De Silva, K. Lai, J. S. Wilmott, M. Synnott, P. Hersey, R. F. Kefford, J. F. Thompson, Y. H. Yang, and R. a Scolyer, "BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma.," *J. Invest. Dermatol.*, vol. 133, no. 2, pp. 509–17, Feb. 2013.

[66] A. Natarajan, G. G. Yardimci, N. C. Sheffield, G. E. Crawford, and U. Ohler, "Predicting cell-type-specific gene expression from regions of open chromatin.," *Genome Res.*, vol. 22, no. 9, pp. 1711–22, Sep. 2012.

[67] S. C. Smith, A. S. Baras, D. Ph, G. Dancik, Y. Ru, K. Ding, C. A. Moskaluk, J. Lehmann, M. Stöckle, A. Hartmann, and K. Jae, "molecular nodal staging of bladder cancer," vol. 12, no. 2, pp. 137–143, 2013.

[68] A. Schaefer, M. Jung, H.-J. Mollenkopf, I. Wagner, C. Stephan, F. Jentzmik, K. Miller, M. Lein, G. Kristiansen, and K. Jung, "Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma.," *Int. J. Cancer*, vol. 126, no. 5, pp. 1166–76, Mar. 2010.

[69] J. Zhu, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, no. 3, pp. 427–443, Jul. 2004.

[70] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, Nov. 2003.

[71] M. J. Hassett, S. M. Silver, M. E. Hughes, D. W. Blayney, S. B. Edge, J. G. Herman, C. a Hudis, P. K. Marcom, J. E. Pettinga, D. Share, R. Theriault, Y.-N. Wong, J. L. Vandergrift, J. C. Niland, and J. C. Weeks, "Adoption of gene expression profile testing and association with use of chemotherapy among women with breast cancer.," *J. Clin. Oncol.*, vol. 30, no. 18, pp. 2218–26, Jun. 2012.

[72] M. a Cobleigh, B. Tabesh, P. Bitterman, J. Baker, M. Cronin, M.-L. Liu, R. Borchik, J.-M. Mosquera, M. G. Walker, and S. Shak, "Tumor gene expression and prognosis in breast cancer patients with 10 or more positive lymph nodes.," *Clin. Cancer Res.*, vol. 11, no. 24 Pt 1, pp. 8623–31, Dec. 2005.

[73] a L. Richards, L. Jones, V. Moskvina, G. Kirov, P. V Gejman, D. F. Levinson, a R. Sanders, S. Purcell, P. M. Visscher, N. Craddock, M. J. Owen, P. Holmans, and M. C. O'Donovan, "Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain.," *Mol. Psychiatry*, vol. 17, no. 2, pp. 193–201, Feb. 2012.

[74] C. C.-M. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. Macrossan, "Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 6, pp. 1580–91, 2011.

[75] E. B. Hunt, *Concept learning, an information processing problem*. New York: Wiley, 1962.

[76] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. California: Wadsworth International, 1984.

[77] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[78] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology.," *Mol. Biosyst.*, vol. 5, no. 12, pp. 1593–605, Dec. 2009.

[79] J. F. Torres-Roca and C. W. Stevens, "Predicting response to clinical radiotherapy: past, present, and future directions.," *Cancer Control*, vol. 15, no. 2, pp. 151–6, May 2008.

[80] U. T. Shankavaram, S. Varma, D. Kane, M. Sunshine, K. K. Chary, W. C. Reinhold, Y. Pommier, and J. N. Weinstein, "CellMiner: a relational database and query tool for the NCI-60 cancer cell lines.," *BMC Genomics*, vol. 10, p. 277, Jan. 2009.

[81] S. Eschrich, H. Zhang, H. Zhao, D. Boulware, J.-H. Lee, G. Bloom, and J. F. Torres-Roca, "Systems biology modeling of the radiation sensitivity network: a biomarker discovery platform.," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 75, no. 2, pp. 497–505, Oct. 2009.

[82] B. Sackler, "Feature selection methods for classification of gene expression profiles," no. April, 2008.

[83] A. Rakotomamonjy, "Variable Selection Using SVM-based Criteria," *J. ofMachine Learn. Res.*, vol. 3, pp. 1357–1370, 2003.

[84] E. Cox, *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*, 1st ed. San Francisco: Morgan Kaufmann Publisher, 2005, pp. 67–205.

[85] N. Draper and H. Smith, *Applied regression analysis*, 2nd ed. New York, NY, 1981, p. Ch 6.

[86] T. Therneau, B. Atkinson, and B. Ripley, "Package ' rpart ,'" *R News*, 2014.

[87] T. M. Therneau and E. J. Atkinson, "An Introduction to Recursive Partitioning Using the RPART Routines," 2014.

[88] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology Supplementary material," no. c, pp. 1–10, 2009.

[89] A. Liaw and M. Wiener, "Breiman and Cutler's random forests for classification and regression." R, 2014.

[90] H. Ying, F. Lin, R. D. MacArthur, J. a Cohn, D. C. Barth-Jones, H. Ye, and L. R. Crane, "A fuzzy discrete event system approach to determining optimal HIV/AIDS treatment regimens.," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 4, pp. 663–76, Oct. 2006.

[91] C. Z. Janikow, "Fuzzy decision trees: issues and methods.," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 28, no. 1, pp. 1–14, Jan. 1998.

[92] J. Lu and G. Zhang, "FUZZY MULTI-OBJECTIVE DECISION-," *Fuzzy Multi-Criteria Decis. Mak.*, vol. 16, pp. 483–522, 2008.

[93] M. Inuiguchi and J. Ramã, "Possibilistic linear programming : a brief review of fuzzy mathematical programming and a comparison with stochastic programming in portfolio selection problem," vol. 111, pp. 3–28, 2000.

[94] R. S. Michalski, "Understanding the Nature of Learning," in *Machine Learning: An Artificial Intelligence Approach*, Morgan Kaufmann, 1986, pp. 3–26.

[95] T. Whelan, C. Sawka, M. Levine, A. Gafni, L. Reyno, A. Willan, J. Julian, S. Dent, H. Abu-Zahra, E. Chouinard, R. Tozer, K. Pritchard, and I. Bodendorfer, "Helping patients make informed choices: a randomized trial of a decision aid for adjuvant chemotherapy in lymph node-negative breast cancer.," *J. Natl. Cancer Inst.*, vol. 95, no. 8, pp. 581–7, Apr. 2003.

APPENDICES

Appendix A Rectal Cancer Detection and Staging

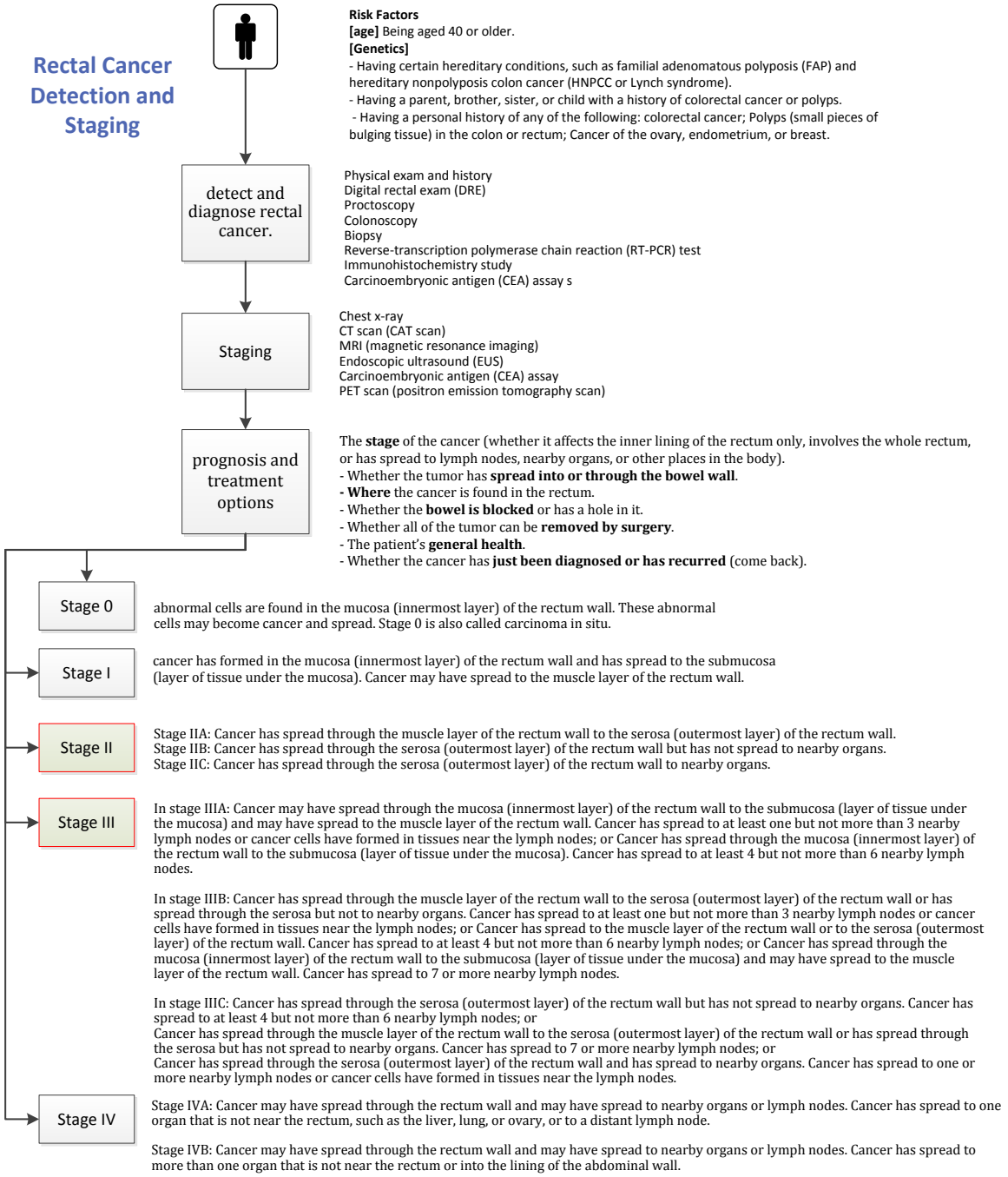
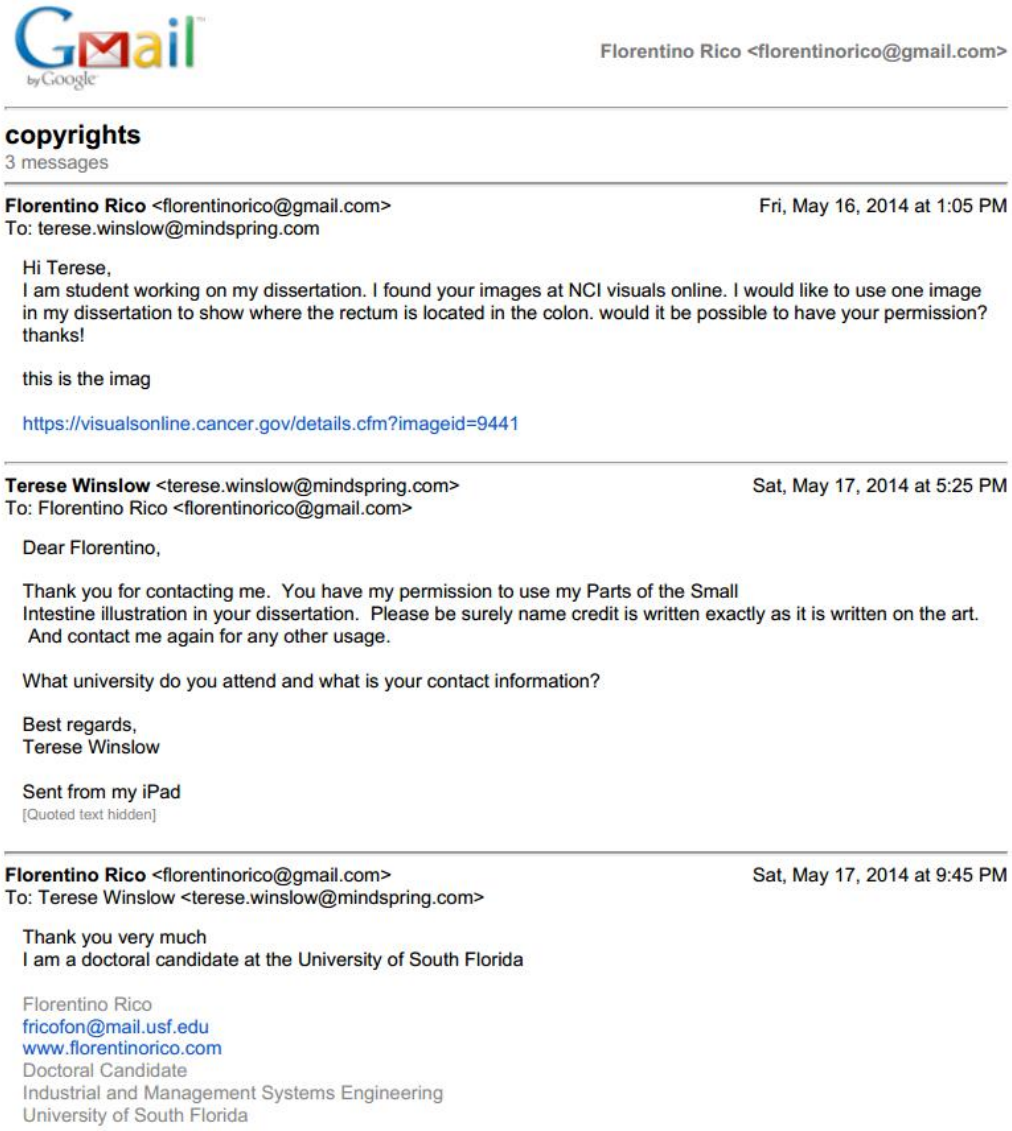


Figure A.1 Rectal cancer detection and staging

Appendix B Figure Permission

Permission to use figure 1 was provided by Terese Winslow.

terese.winslow@mindspring.com to use in this dissertation on May 17/2014



Gmail by Google

Florentino Rico <florentinorico@gmail.com>

copyrights

3 messages

Florentino Rico <florentinorico@gmail.com> Fri, May 16, 2014 at 1:05 PM
To: terese.winslow@mindspring.com

Hi Terese,
I am student working on my dissertation. I found your images at NCI visuals online. I would like to use one image in my dissertation to show where the rectum is located in the colon. would it be possible to have your permission? thanks!

this is the imag

https://visualsonline.cancer.gov/details.cfm?imageid=9441

Terese Winslow <terese.winslow@mindspring.com> Sat, May 17, 2014 at 5:25 PM
To: Florentino Rico <florentinorico@gmail.com>

Dear Florentino,

Thank you for contacting me. You have my permission to use my Parts of the Small Intestine illustration in your dissertation. Please be surely name credit is written exactly as it is written on the art. And contact me again for any other usage.

What university do you attend and what is your contact information?

Best regards,
Terese Winslow

Sent from my iPad
[Quoted text hidden]

Florentino Rico <florentinorico@gmail.com> Sat, May 17, 2014 at 9:45 PM
To: Terese Winslow <terese.winslow@mindspring.com>

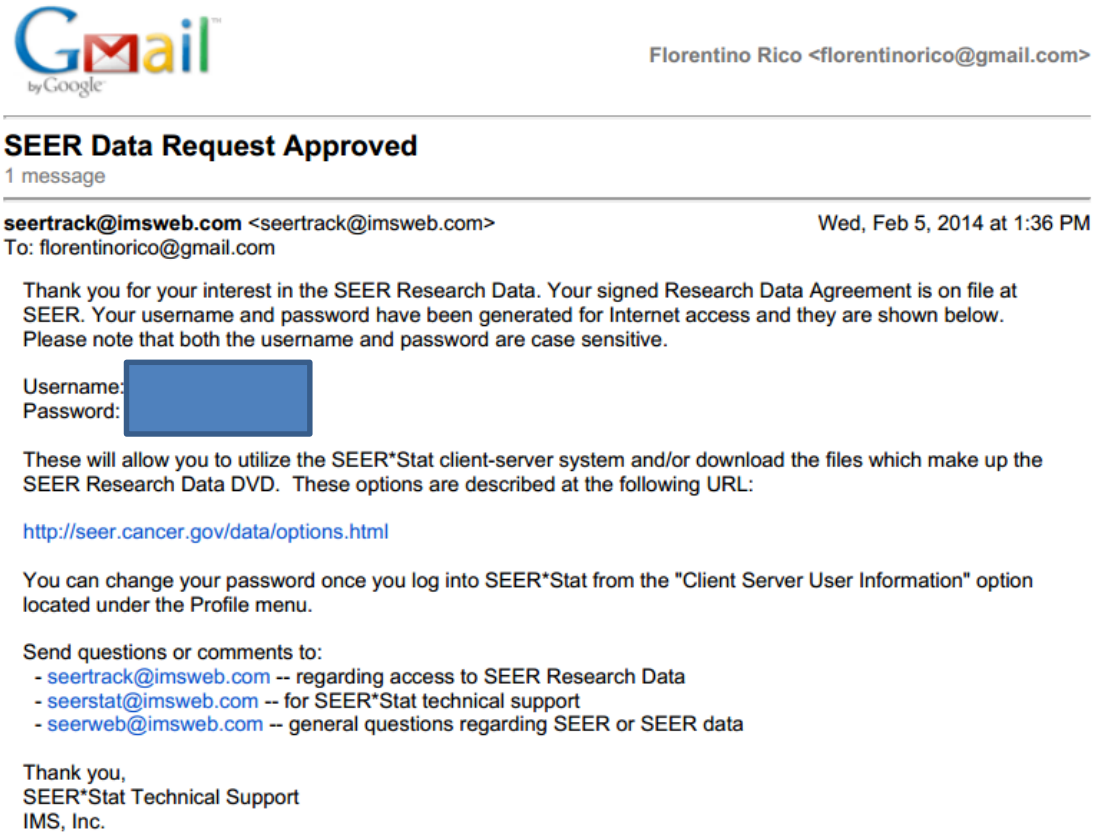
Thank you very much
I am a doctoral candidate at the University of South Florida

Florentino Rico
fricofon@mail.usf.edu
www.florentinorico.com
Doctoral Candidate
Industrial and Management Systems Engineering
University of South Florida

I have created all other figures for this manuscript, and no other material is being used or published elsewhere.

Appendix C SEER Data Use Agreement

The data use agreement was approved on Feb 5, 2014.



Florentino Rico <florentinorico@gmail.com>

SEER Data Request Approved

1 message

seertrack@imsweb.com <seertrack@imsweb.com> Wed, Feb 5, 2014 at 1:36 PM
To: florentinorico@gmail.com

Thank you for your interest in the SEER Research Data. Your signed Research Data Agreement is on file at SEER. Your username and password have been generated for Internet access and they are shown below. Please note that both the username and password are case sensitive.

Username: [redacted]
Password:

These will allow you to utilize the SEER*Stat client-server system and/or download the files which make up the SEER Research Data DVD. These options are described at the following URL:

http://seer.cancer.gov/data/options.html

You can change your password once you log into SEER*Stat from the "Client Server User Information" option located under the Profile menu.

Send questions or comments to:
- seertrack@imsweb.com -- regarding access to SEER Research Data
- seerstat@imsweb.com -- for SEER*Stat technical support
- seerweb@imsweb.com -- general questions regarding SEER or SEER data

Thank you,
SEER*Stat Technical Support
IMS, Inc.

Appendix D SEER Database Variables Used

Table D.1 SEER database variables used

VARIABLE NAME	VARIABLE DESCRIPTION
CS_SSF25	CS Site-Specific Factor 25
D_AJCC_T	Derived AJCC T
D_AJCC_N	Derived AJCC N
D_AJCC_M	Derived AJCC M
D_AJCC_S	Derived AJCC Stage Group
CS0204SCHEMA	CS Schema v0204
CASENUM	Patient ID number
REG	Registry ID
MAR_STAT	Marital Status at DX
RACE	Race/Ethnicity
NHIA	NHIA Derived Hispanic Origin
SEX	Sex
AGE_DX	Age at diagnosis
SEQ_NUM	Sequence Number--Central
DATE_mo	Month of diagnosis
DATE_yr	Year of diagnosis
SITEO2V	Primary site
HISTO2V	Histology (92-00) ICD-O-2
BEHO2V	Behavior (92-00) ICD-O-2
HISTO3V	Histologic Type ICD-O-3
BEHO3V	Behavior Code ICD-O-3
GRADE	Grade
REPT_SRC	Type of Reporting Source
REC_NO	SEER Record number
AGE_REC	Age Recode <1 Year olds
HISTREC	Histology Recode--Broad Groupings
BRAINREC	Histology Recode--Brain Groupings
NUMPRIMS	Number of primaries
SRV_TIME_MON	Survival months
SRV_TIME_MON_FLAG	Survival months flag
INSREC_PUB	Insurance Recode (2007+)
RAC_RECA	Race recode (White, Black, Other)
RAC_RECY	Race recode (W, B, AI, API
NHIAREC	Origin Recode NHIA(Hispanic,Non-Hisp)
CS_SIZE	CS Tumor size (from 2004)
CS_EXT	CS Extension
CS_NODE	CS Lymph Nodes
CS_METS	CS Mets at DX
SURGPRIM	RX Summ--Surg Prim Site
SURGNODE	RX Summ--Reg LN Examined
NO_SURG	Reason for no surgery
RADIATN	RX Summ--Radiation
RAD_SURG	RX Summ--Surg/Rad Seq
SCOPE	RX Summ--Scope Reg LN Sur
SURGOTH	RX Summ--Surg Oth Reg/Dis

Appendix E Parameter Estimates for the Logistic Regression

Table E.1 Parameter estimates for the logistic regression

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Ch-sqChiSq
Intercept		1	2.3291	10.1799	0.0523	0.819
TREATment	**No**	1	-1.0536	0.0749	197.765	<.0001
TREATment	**RS**	1	0.551	0.0505	118.921	<.0001
MAR_STAT	**D**	1	-0.1059	0.0661	2.5677	0.1091
MAR_STAT	**E**	1	0.2827	0.1703	2.7575	0.0968
MAR_STAT	**M**	1	0.1253	0.0478	6.8685	0.0088
MAR_STAT	**S**	1	-0.1908	0.059	10.4572	0.0012
MAR_STAT	**W**	1	-0.1511	0.0544	7.6988	0.0055
RAC_RECA	**B**	1	-0.4475	0.1102	16.5006	<.0001
RAC_RECA	**O**	1	0.0137	0.1226	0.0125	0.9111
RAC_RECA	**U**	1	0.5526	0.3014	3.3617	0.0667
SEX	**F**	1	0.0742	0.0203	13.3585	0.0003
AGE_REC	**3**	1	6.0071	152.7	0.0015	0.9686
AGE_REC	**4**	1	0.5897	10.2268	0.0033	0.954
AGE_REC	**5**	1	0.2508	10.1917	0.0006	0.9804
AGE_REC	**6**	1	-0.0172	10.1853	0	0.9987
AGE_REC	**7**	1	0.0268	10.1821	0	0.9979
AGE_REC	**8**	1	-0.2733	10.1804	0.0007	0.9786
AGE_REC	**9**	1	0.0385	10.1801	0	0.997
AGE_REC	**10**	1	-0.1542	10.1796	0.0002	0.9879
AGE_REC	**11**	1	0.0351	10.1795	0	0.9972
AGE_REC	**12**	1	-0.3029	10.1794	0.0009	0.9763
AGE_REC	**13**	1	-0.5177	10.1794	0.0026	0.9594
AGE_REC	**14**	1	-0.5622	10.1794	0.003	0.956
AGE_REC	**15**	1	-0.8548	10.1794	0.0071	0.9331
AGE_REC	**16**	1	-1.1174	10.1794	0.012	0.9126
AGE_REC	**17**	1	-1.3787	10.1794	0.0183	0.8923
D_AJCC_S	**2A**	1	0.7012	0.0727	93.1459	<.0001
D_AJCC_S	**2B**	1	-0.3976	0.0806	24.34	<.0001
D_AJCC_S	**3A**	1	0.6333	0.1062	35.5295	<.0001
D_AJCC_S	**3B**	1	-0.1534	0.0714	4.6129	0.0317
D_AJCC_S	**3C**	1	-0.7567	0.0725	109.074	<.0001
GRADE	**1**	1	0.2838	0.072	15.5262	<.0001
GRADE	**2**	1	0.2804	0.0387	52.5588	<.0001
GRADE	**3**	1	-0.2316	0.0441	27.6239	<.0001
GRADE	**4**	1	-0.2451	0.0889	7.6087	0.0058
INSREC_PUB	**I**	1	0.0743	0.0503	2.1852	0.1393
INSREC_PUB	**M**	1	-0.3154	0.0605	27.1385	<.0001
INSREC_PUB	**U**	1	-0.096	0.0853	1.2668	0.2604

Appendix F Transition Probabilities for Adverse Effects and Efficacy

Table F.1 Transition probabilities for adverse effects and efficacy

Adverse Effects	
Surgery alone	30
Radiation and Surgery	40
Observation	10

Surgery alone

initial	Minor	Moderate	Major	
0	0.118	0.872	0.010	initial
0	0	0	0	first
0	0	0	0	second
0	0	0	0	third

Radiation and Surgery

initial	Minor	Moderate	Major	
0	0.000	0.291	0.708	initial
0	0	0	0	first
0	0	0	0	second
0	0	0	0	third

Observation

initial	Minor	Moderate	Major	
0	0.972	0.028	0.000	initial
0	0	0	0	first
0	0	0	0	second
0	0	0	0	third

Treatment Efficacy	
Surgery alone	60
Radiation and Surgery	70
Observation	10

Surgery alone

initial	low	med	high	
0	0.015	0.985	0.000	initial
0	0	0	0	unlikely
0	0	0	0	neutral
0	0	0	0	likely

Radiation and Surgery

initial	unlikely	neutral	likely	
0	0.000	0.985	0.015	initial
0	0	0	0	unlikely
0	0	0	0	neutral
0	0	0	0	likely

Observation

initial	unlikely	neutral	likely	
0	1.000	0.000	0.000	initial
0	0	0	0	unlikely
0	0	0	0	neutral
0	0	0	0	likely