South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

2018

Adaptive Interventions Treatment Modelling and Regimen Optimization Using Sequential Multiple Assignment Randomized Trials (Smart) and Q-Learning

Abiral Baniya

South Dakota State University

Recommended Citation

Baniya, Abiral, "Adaptive Interventions Treatment Modelling and Regimen Optimization Using Sequential Multiple Assignment Randomized Trials (Smart) and Q-Learning" (2018). *Electronic Theses and Dissertations*. 2473.
https://openprairie.sdstate.edu/etd/2473

ADAPTIVE INTERVENTIONS TREATMENT MODELLING AND REGIMEN OPTIMIZATION USING SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS (SMART) AND *Q*-LEARNING

BY

ABIRAL BANIYA

A thesis submitted in partial fulfillment of the requirements for

Master of Science

Major in Electrical Engineering

South Dakota State University

2018

ADAPTIVE INTERVENTIONS TREATMENT MODELLING AND REGIMEN OPTIMIZATION USING SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS (SMART) AND *Q*-LEARNING

ABIRAL BANIYA

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Electrical Engineering degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Qiquan Qiao, PhD Date
Major Advisor

Huitian Lu, PhD Date
Thesis Advisor

Steven Hietpas, PhD Date
Head, Electrical Engineering and Computer Science

Dean, Graduate School Date

ACKNOWLEGEMENTS

The research work demonstrated in the thesis was supported by the grants from Sanford Health and their weight management profile plan. Also, I would like to acknowledge the master's program of Electrical Engineering and Computer Science Department, South Dakota State University for giving me an opportunity to work on the research project.

Further, I would like to express my sincere gratitude to Dr. Qiquan Qiao, my advisor and Dr. Huitian Lu, my thesis advisor for including me in their research group and providing me with unwavering support throughout the research period. Continuous feedback from them have proved to be crucial in improving the quality of this thesis. I would also like to thank Dr. Gary Hatfield and Dr. Matthew Vukovich, my thesis committee members for their time, support and consideration on reviewing my thesis and providing me with constructive comments. Also, I would like to thank my research group in SDSU for their help and support.

Finally, last but not the least I would like to thank my family, especially my lovely parents for their consistent advice and support from overseas, without them this would not have been nearly possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

ADAPTIVE INTERVENTIONS TREATMENT MODELLING AND REGIMEN OPTIMIZATION USING SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS (SMART) AND *Q*-LEARNING

ABIRAL BANIYA

2018

Nowadays, pharmacological practices are focused on a single best treatment to treat a disease which sounds impractical as the same treatment may not work the same way for every patient. Thus, there is a need of shift towards more patient-centric rather than disease-centric approach, in which personal characteristics of a patient or biomarkers are used to determine the tailored optimal treatment. The “one size fits all” concept is contradicted by research area of personalized medicine. The Sequential Multiple Assignment Randomized Trial (SMART) is a multi-stage trials to inform the development of dynamic treatment regimens (DTR’s). In SMART, a subject is randomized through various stages of treatment where each stage corresponds to a treatment decision. These types of adaptive interventions are individualized and are repeatedly adjusted across time based on patient’s individual clinical characteristics and ongoing performance. The reinforcement learning (Q-learning), a computational algorithm for optimization of treatment regimens to maximize desired clinical outcome is used in optimizing the sequence of treatments. This statistical model contains regression analysis for function approximation of data from clinical trials. The model will predict a series of regimens across time, depending on the biomarkers of a new participant for optimizing the weight management decision rules.

Additionally, for implementing reinforcement learning algorithm, as it is one of the machine learning approach there should be a training data from which we can train the model or in other words approximate the function, *Q*-functions. Then the approximated functions of the model should be evaluated and after the evaluation they should be further tested for applying the treatment rule to future patients. Thus, in this thesis first the dataset obtained from Sanford Health is first restructured, to make it conducive for our model utilization. The restructured training data is used in regression analysis for approximating the *Q*-functions. The regression analysis gives the estimates of coefficients associated to each covariate in the regression function. The evaluation of model goodness-of-fit and fulfillment of underlying assumptions of simple linear regression are performed using regression summary table and residual diagnostic plots. As a two stage SMART design is put into practice, the *Q*-functions for these two stages are needed to be estimated through multiple regression using linear model. Now, finally after analyzing the fit adequacy the model is applied for prescribing treatment rules to future patients. The prognostic and predictive covariates of new patient is acquired and the optimal treatment rule for each treatment decision stage is assigned as the treatment that results in maximum estimated values of *Q*-functions. The estimated values of each regime were also computed using the value estimator function and regime that has the maximum estimated value was chosen as optimal treatment decision rule.

CHAPTER 1

INTRODUCTION

1.1 Background

On a daily basis decision makers or doctors play a vital role of recommending treatments to patients in any kind of pharmacological practice [1]. Managing or treating a chronic illness generally involves a sequence of treatment decisions in which factors such as response to previous treatments, severity of symptoms and medicinal side-effects are to be taken under consideration while deciding on if, when and how current treatment status needs to be altered. Previously, these decisions were made based on identifying a single best treatment for a particular disease, however, the clinical treatment design has begun to shift towards more patient- centric approach rather than disease-centric one [2]. The notion “Personalized Medicine” is based on the fact that two patients given the same treatment may well respond differently or in other words a treatment that worked for one patient may not work for the other. Further, Topol writes that “We have entered a new era of medicine, in which each person can be near fully defined at the individual level, instead of how we practice medicine at the population level [3].” Therefore, Personalized Medicine underpins the posit that rather than direct focus on disease diagnosis and treatment allocation, pharmacological practices should aim towards more personalized approach which takes into account the patient’s biomarkers or characteristics and these should dictate the treatment that will work best for an individual.

As the personalized medical decision-making process is sequential and involves careful assessment of patients’ individual characteristics and their ongoing performance so that the nature of outcomes can be improved over time. Dynamic

Treatment Regime (DTR) also known as adaptive interventions [4] or adaptive treatment strategies [5] follows a sequential decision-making protocol comprising of series of treatment decisions that take the current patient's health information and their past treatment history as inputs and outputs the time and procedure for treatment alteration. Hence, the Adaptive treatment strategies (ATSs) are a vastly expanding area of clinical research and is preferred as more formal means of implementing personalized medicine. The strategy discussed in ATS is the allocation of treatment at each sequential treatment decision point that depends upon patient's history of covariates and past treatments. In a simple scenario, we can consider a single decision point rule where a patient is prescribed drug A if he is overweight otherwise prescribed drug B. This scenario can occur in sequence at each lengthy follow-up visit of the patient where treatment allocation decision is undertaken each time. The follow-up visits can be considered as number of stages in the DTR setting where in each stage the treatments are tailored to alterations in patient's characteristics and their response to previous treatments.

The ATS or DTR strategy involves multistage treatment decisions, thus we need to design a multistage and sequential clinical trial for obtaining a high-quality observational training dataset. Hence, a Sequential Multiple Randomized Trials (SMART) design randomizes the treatments based on individual patient's biomarkers and clinical history at each stage of sequential decision-making process. This design supports adaptive interventions that adapts to the system dynamics in a multi-stage trial through a sequence of decision rules which dictate the intervention path in order to maximize the long-term primary outcomes [6]. Using this SMART design, researchers can collect or construct high quality training data which can identify treatment allocation strategies that will eventually optimize patient's health status. The concept of adaptive interventions mainly consists of two main components:

- Individualized interventions based on patient's characteristics and needs.
- Interventions are time varying as they repeatedly adapt over time responding to participants ongoing performance and varying needs.

Thus, we need an experimental setup with a sequence of decision rules called DTR where patients covariates and treatment history are taken as input and the recommended treatment decision rule is the output of the system. In this experimental setup, the goal is to figure out the optimal sequential decision rule described as one that maximizes the desirable clinical outcome. Approximate dynamic programming and *Q*-learning a generalization reinforcement learning and regression analysis technique with function approximation for obtaining an optimal decision sequence in clinical interventions and services, are very popular as the nature of clinical decision making is sequential. This reinforcement learning method called *Q*-learning is particularly more appealing as it is easy to implement and perhaps more importantly can be understood by non-statistical personals. The algorithm involves learning an optimal regimen from patient's data generated using clinical reinforcement trial [7]. *Q*-learning involves approximating the *Q*-functions defined by time indexed parameters of patients' biomarkers that is obtained by regression analysis at each intervention stage. The regression based approximation of *Q*-function is implemented using linear model. In this model, the inputs are the training data generated from SMART design and the outputs are the approximated functions for potential final outcome. Finally, the optimal treatment policy or the potential final outcome is defined by the treatment sequence that maximizes these Q-functions.

1.2 Literature Review

Reinforcement learning was introduced in pharmacological practice by Pineu et al. (2007) to represent the concept of adaptive interventions by applying hypothetical SMART study on alcohol dependence [8]. Pineau considered using reinforcement learning for data analysis of studies that involved patients randomized to multiple clinical trials, sequentially or more precisely SMART designs. Similarly, in same year, Murphy et al. (2007, Neuropsychopharmacology) suggested that Q-learning can be an important breakthrough for designing ATS and constructing decision rules in chronic psychiatric disorders [9].

Further, Ma et al. in years of 2015 and 2016 published two different works on establishing Personalized Treatment Rules in the field of oncology [10, 11]. First, they provided an overview on statistical methods to establish optimal treatment rules for individualized medicine and also discussed examples in different medical context, oncology being the emphasis. Various statistical inference methods for identifying Individualized Treatment Rules (ITR) such as Multiple Regression for Randomized Clinical Trial Data, Survival Analysis and methods for observational data and high- dimensional biomarkers were introduced. Also, they discussed some advanced methods of inference such as Robust Inference and data mining using machine learning and the performances of these methods were evaluated for ITR. Secondly, Ma et al. implemented Bayesian Predictive Framework for integrating high-dimensional set of genomic features data with clinical responses and treatment histories of patients. However, unsupervised clustering with microevolutionary process was used which was very complicated as personalized medicine in field of oncology may have some limitation because of the fact that biomarkers or characteristics obtained from small set of sample or panel genes is never adequate to describe heterogeneity inherent to the diseases.

Between years 2011-14 Q- learning, a reinforcement learning algorithm, has been a popular method for determining optimal treatment regimen operating data generated from clinical trials assignment. In 2011, Zhao et al. implemented Q-learning for learning an optimal regimen using training data generated from clinical trials assigned to patients with Non-Small Cell Lung Cancer [7]. The combination of “clinical reinforcement trial” assignment for obtaining training dataset and support vector regression for Q-function approximation were incorporated to estimate optimal regimens that are individualized to patients’ subpopulation. Although the simulation studies depicted small estimation bias while using sample size $N \geq 100$, several challenges were faced in determining appropriate sample size and learning generalization error for clinical reinforcement trial design.

In 2012, Shani and Moodie et al. performed two different experiments that involved adaptive interventions, clinical trial assignments and treatment regimen optimization. First, Shani et al. introduced Q-learning and the use of Q to indicate the *quality* of given or chosen treatment [12]. They implemented Q-learning which is a regression based function approximation method, with linear estimates to prescribe adaptive interventions for children with ADHD and the training data was obtained from ADHD SMART study (Center for Children and Families, SUNY at Buffalo, William E. Pelham as PI). The operation of this learning algorithm was illustrated for using in data from SMART design with different settings such as SMART design with no embedded tailoring variables, re-randomization depending upon intermediate outcomes, re-randomization depending upon an intermediate outcome and prior treatment. The advantages such as inclusion of both direct and indirect effects during intervention stages, control for optimal-second stage intervention while operating effects of first-stage interventions and reduced potential bias of *Q*-learning over other regression based approach were also discussed.

Furthermore, Moodie et al. implemented *Q*-learning and mentioned that it is a popular method for estimating DTRs [13]. This reinforcement learning method was used to examine the effects of breastfeeding on verbal cognitive ability and growth of infant and it was based on observational data from Promotion of Breastfeeding Intervention Trial. First, the authors discussed the use of *Q*-learning according to different settings such as with multiple regression models, non-regular settings and with observational dataset. Secondly, they discussed upon the simulation study for comparison of performances using the *Q*-learning with certain adjustments such as inclusion of (1) covariates as linear terms in *Q*-function, (2) propensity score (PS) in *Q*-function (3) including quintiles of PS as covariates in the *Q*-function and (4) Inverse Probability of treatment weights (IPTW). Finally, a case study was presented on The Probit Study that analyzed the breastfeeding and vocabulary test results. In this case study hospitals and other polyclinics that were affiliated with Republic of Belarus were randomized to breastfeeding promotion intervention model presented by WHO/UNICEF. The intervention or decision rule suggested 98% infants, at age 6.5 months who were breastfed until 3 months scored maximum in vocabulary and only 33% scored maximum who were breastfed until 6 months. Finally, it was concluded that *Q*-learning is an appealing method for constructing DTR and recommended that the covariates in the model for *Q*-functions should be directly included during function approximation process.

Whereas if the relationships between input cofounders and outcome are not properly understood than it is necessary to consider splines of polynomial functions to ensure adequate model fit.

Conditional mean and variance modelling for smooth transformation of data before applying non-smooth and nonmonotone operation of *Q*-function approximation using regression analysis method was introduced by Laber et al. in 2014, for adequate fitting and well interpretable model [14]. In *Q*-learning the value that maximized the second-stage *Q*-function or the optimal value is assigned as potential optimal outcome for first-stage regression, this process is replaced by two ordinary mean-variance function modeling described in Interactive model building technique called *IQ*- learning. The model was implemented in Monte Carlo simulations and Sequenced Treatment Alternatives to Relieve Depression (STAR*D,2004) study that involved a sequentially randomized study of major depressive disorder [15]. Although the process of defining contrast and main effect functions for assigning optimal second stage outcome seems appealing, the modelling of conditional distribution of these estimated contrast functions is complicated and it may result as inadequate model subsequently.

Lastly, Schulte et al. in 2014 implemented *Q* and *A* learning methods to estimate the optimal treatment regimens and the contrast between these two statistical estimation methods was also discussed [16]. *A*-learning posits that the entire *Q*-function is not needed to be defined for optimal regime estimation, however, this statistical framework only requires the regression model representing treatment contrasts and probability of a particular treatment being assigned to a patient at each intervention given the patient information at these points. Further, the simulation study was performed for one decision and two decision points and also applied to STAR*D study which involves four stages with each stage consisting 12 weeks of treatment period. This study suggested that although *A*-learning is more robust to model misspecification than *Q*-learning but its performance degrades when there are more than two treatment options at each stage and the decision rules for defining optimal treatment regime is very complex. It is also mentioned that *Q*-learning is more practical and more familiar to data analysts as it consists preset of standard model diagnostic tools.

Therefore, in summary the above literature discussed different statistical estimation methods such as IPTW, PS in *Q*-functions, *A*-learning including the reinforcement learning approach of *Q*-learning and implemented them in medical research such as ADHD studies, breastfeeding case-study, STAR*D study for estimating optimal treatment regime. However, no such research or data analysis has been done on weight management treatment plans and although there are some limitations on operating *Q*-learning it seems more practical and adequate for defining ATS. Similarly, the SMART design for obtaining training data is a promising way for using this optimization algorithm as it defines sequential decision making and randomizes treatment at each decision points so we can observe the treatment that results maximum potential outcome or in other words optimal treatment decision.

1.3 Motivation and Objectives

There is a need of an accurate mathematical model for trial assignment and optimization of adaptive treatment strategies or personalized treatment regimens for weight management plans in Sanford Profile.

The main objective of this research was to design a Sequential Multiple Randomized Trial (SMART) for trial assignment to operate adaptive interventions and restructure the Sanford Profile weight management dataset per this design and eventually use this dataset for implementing *Q*-learning, a reinforcement learning algorithm that involves function approximation using regression analysis to optimize the sequence of decision rules for personalized treatment. Thus, in order to achieve this objective, the following tasks were performed:

1) Restructuring of Sanford Health Data on weight management according to SMART design with two stages or decision points and two treatment options at each of these points for preparing a training dataset with 210 observations.
2) Implementing *Q*-learning algorithm which involves regression analysis for function approximation where input are the training data and output is the approximated function for potential final outcome.
3) The regression summary at each stage is obtained which gives the estimated coefficients of each independent predictor variables in approximated regression function. It also provided the values of Residual Standard Error which is the standard deviation of the residuals or error giving how close the fit of regression line is to the points.
4) Graphical analysis was performed using regression diagnostics plots for each stage regression to further analyze the fit adequacy and check underlying assumption of applied regression model.
5) The optimization of adaptive decision rule was presented according to maximum values of *Q*-functions and treatment resulting maximum *Q*-value was assigned as optimal intervention rule.

6) The regime value was estimated using weighted average of the outcomes observed from patients in trial and the regime with maximum estimated value was assigned as optimal treatment regime.

1.4 Organization of the Thesis

Chapter 1 provides the introduction on the subject and also background about personalized medicine. Numerous literature that describes various models on personalized healthcare are also described in this chapter. Also, the need of mathematical model and optimization is explained in this chapter. Similarly, Chapter 2 defines the theory behind the model and also provides the definition and scopes of personalized medicine. In this chapter different framework such as SMART design, reinforcement learning and statistical inference which are implemented in the model are also described.

Chapter 3 defines the overall methodology for development of the model which is further employed to prescribe personalized treatment rule. In this chapter methods of acquiring training data, mathematical framework for model, optimization assumptions, residual analysis and sampling for model validation are described. Chapter 4 shows the results obtained by applying model to define treatment rule for future patients. In this chapter the results of data restructuring, regression analysis, residual diagnostic plots and regime value estimates can be observed. Lastly, Chapter 5 describes the summary of the model and conclusions from the model employment. Also, in this chapter the future work is listed so that useful modifications and enhancement to the present model can be applied.

CHAPTER 2

THEORY

2.1 Personalized Medicine

2.1.1 Definition and scope

Personalized Medicine is a medical term that highlights the methodical use of individual patient's information for optimizing that patient's health. This pharmacological paradigm is motivated by the fact that patients usually respond differently to a treatment when primary outcome and side effects are compared among a group of patients. The heterogeneity in treatment response among a group of patients when a treatment is assigned to them has caused the ideological transition of researchers from the notion of one-size-fits all to more logical method of personalized medicine. Benefits of personalized medicine include improved compliance or adherence to prescribed treatment which will result in enhanced patient care and reduces the overall cost of healthcare. The phrase personalized medicine is not only popular in medical community or among physicians but is also making its mark among many quantitative researchers or statisticians. The reason behind this growth of interest is due to the methodological challenges involved in constructing the treatment rules in personalized medicine as they are evidence-based, and data driven. Thus, there is a broader scope and unprecedented surge of interest among statisticians, engineers, computer scientists and other quantitative researchers in this field of research which are leading to many efficient methodological developments.

Dynamic treatment regimen, an important aspect of personalized medicine defines a set of treatment rules at each treatment decision time and these treatment rules are tailored to an individual patient according to the patient's biomarkers, history, characteristics and response to previous treatments. These decision rules prescribe the treatment the physician should follow or treatment decision she/he should take at each decision points and the characteristics that influence the treatment decisions can be demographics, case history, genetic information and other medical parameters of an individual patient.

The concept of personalized medicine was described and appreciated in medicine since 1960s and soon the publication followed on the Medline interface in 1999 [17]. Thus, tailoring treatments based on an individual patient's biomarkers has become a focus area for researchers in area of personalized medicine. Figure 2.1 shows the evolution of personalized medicine from year 2000 to 2015 and various stages of progress within these years [18]. From years 2000-2005 numerous projects were undertaken for profiling personal genome with the aim providing personal genome information to general public at a low cost. The era of personalized medicine began in 21st century medicine history which mainly focuses on pharmacogenetics, molecular diagnostics and empirical treatments. Similarly, years 2005-2010 witnessed an evolution in modern medicine by introduction of bioinformatics, genetic screening, pharmacoproteomics and pharmacogenomics. Furthermore, years 2010-2015 have seen substantial amount of development in field of personalized medicine as in these years the concept of presymptomatic treatment, integrated healthcare, automated systems and rational therapies came into practice.

Therefore, when integrating the pieces on a drawing board, the evolutionary process of transfer from conventional medicine to personalized medicine is inevitable and modern technologies in field of medicine has made medical professionals who are trained in prebiotechnology era to retire and move towards use of these newer technologies that involves genomic knowledge, molecular medicine, pharmacogenetics and pharmacogenomics. Also, there is a need to bridge the gap between these two instances of medicine or in other words between conventional and personalized medicine. For this purpose, Genomics and Personalized Medicine Act [19] was passed in 2006 by the US government.

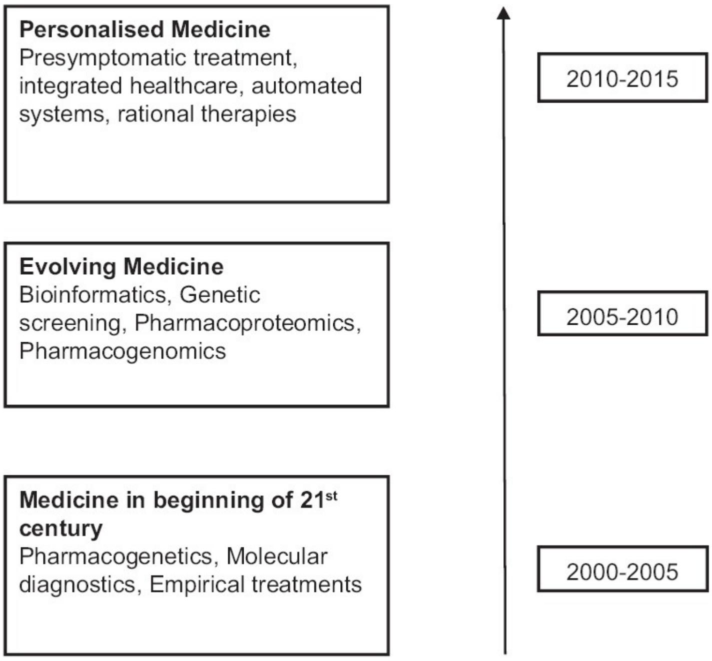

Personalised Medicine
Presymptomatic treatment, integrated healthcare, automated systems, rational therapies
Evolving Medicine
Bioinformatics, Genetic screening, Pharmacoproteomics, Pharmacogenomics
Medicine in beginning of 21st century
Pharmacogenetics, Molecular diagnostics, Empirical treatments
2010-2015
2005-2010
2000-2005

Figure 2.1. Evolution of Personalised medicine [18]

2.1.2 Medical Decision-Making Process

As mentioned in earlier chapters, the decision-making process is vital in pharmacological practice as these decision rule is critical to the patient’s well-being in long run. Although, decision makers try their best and take decisions as per their experience for improvement in patient health, these decisions may not comply and may provide altered results depending on the varying patient’s characteristics and biomarkers. This is where personalized medicine comes in useful as the personalized treatments march towards realizing a set of decision rules that governs the decision-making process or in other words informs a physician what to do in each decision-making stage where each decision solely depends upon the patient’s characteristics such as demographics, case history, genetic information, etc. For developing these decision rules generally, the notions from decision theory such as *utility* are taken into consideration and the decisions are undertaken based on these notion's values.

Decision-theoretic approach have been taken into consideration since long time in medical and health care decision-making. Parmigiani on 2002, asserted that the decision theory ideas contribute in structuring and formally defining the goal and assists in gathering, organizing and integrating the quantitative information that are required for medical decision-making process [20]. Further, Parmigiani describes the Bayesian approach for medical decision-making. However, here we consider different approach and introduce single-stage and multi-stages decision problems in context of personalized medicine and describe them mathematically.

2.1.2.1 Single-stage Decision

To understand the general idea of how decision theory, contributes to the notion of personalized medicine, consider a single-stage decision problem where the clinician should prescribe a single optimal treatment for an individual patient. When this patient comes for a regular clinic visit the clinician will be able to observe certain characteristics of the patient such as biomarker, results of some diagnostic test or results from previous treatment. We consider these variables as history of the patient and denote them by *o*, based on the values of *o* the decision-maker has to decide for example whether to prescribe treatment *a* or *a'*. Thus, this setting is asking for a decision rule which can be for example, "administer treatment *a* to the patient if his individual characteristic *o* is lower than some threshold value, prescribe treatment *a'* otherwise". Hence, decision rule is nothing but mapping of current state of patient that is described by available information prior to treatment, into the space of possible treatment decisions that a clinician can prescribe.

The decision-making process involves statistical evaluation of the *utility* of decision undertaken and the state at which this decision is made. The *utility* function, *u(o,a)* describes the utility of prescribing treatment *a* at state *o*. Wald (1949), described the foundations of general theory of statistical decision functions and derived that the statistical decision problems can be expressed in form of opportunity *loss (or regret)* function denoted as: $L(o, a) = \underbrace{sup}_{a} u(0, a) - u(o, a),$ where the supremum is taken over all possible treatment decisions at point a [21]. After describing the loss function *L(o,a)* the goal now is to search for the treatment decision that minimizes this loss function at state *o* and the decision that results in minimum loss function is equivalent to optimal treatment decision. As the optimality of treatment decision depends upon the state *o* which differ from one individual to other, thus this type of decision-making is personalized. An alternative to loss function is formulating the problem directly in terms of utility itself but the twist here is that the treatment decision which maximizes the utility is chosen as optimal one for given state *o*. There are various ways for defining utility function depending on the problem definition, one way is to assign it the conditional expectation of primary outcome (Y) at the given state, i.e. $u(o, a) = E_a(Y|o)$. The expectation value is calculated according to the probability distribution at treatment decision *a*.

Another method of describing optimal treatment is derived from different econometrics and bio-medical literatures is known as *welfare contrast* [22-27]. The welfare contrast gives the difference between the utilities of two different treatment decisions, in our case, treatments a and *a'*. It is represented mathematically as, $g(o, a, a') = u(o, a) - u(o, a')$ where, $g(o, a, a')$ gives the welfare contrast value and $u(o, a), u(o, a')$ are the utilities corresponding to treatment decisions a and *a'*, respectively. In this case, we should note that a denotes the optimal decision so the value of $g(o, a, a')$ is the regret of administering treatment decision *a'*. Thus, the treatment that results in minimum regret value is considered as optimal treatment decision at point o. The welfare contrast is also known as *blip* value in case of multi-stage decision problems which is described next [28].

2.1.2.2 Multi-stage Decision

As we move ahead from single-stage to multi-stage treatment decisions we need to consider the effects of decisions made at each stage as the decision taken at one stage can affect those made on later stages. Also, in multistage scenario instead of only considering which treatment to choose among the treatment choices we need to be conscious about which treatment to follow after a treatment is prescribed. In this context, a Dynamic Treatment Regime (DTR) is administered to an individual patient and can be described as a set of decision rules, where one treatment decision is made at each intervention stage. These treatment rules adapt according to the state or characteristics of patients which is time varying and depends on previous treatments or patient's history. For assigning decision rule at each stage the system takes the patient's individual characteristics such as biomarkers, response to previous treatments and other demographics as input, and outputs the optimal recommended treatment for that individual which may be drug dosage, timing of treatment, treatment type, etc. DTRs are also known as treatment strategies [5, 29-31], adaptive treatment strategies [32, 33] or treatment policy [34-36]and can be understood as the system that supports a decision maker for making clinical or treatment decisions in medical scenario.

The next goal in multi-stage decision is the optimization of these DTR's that involves first definition of the optimization criteria and then use of some optimization algorithm to obtain maximum utility. The optimization criteria are defined by the maximization of utility functions which can be quantiles such as median or other characteristics of outcome distribution. The primary goals in this multi-stage decision scenario can be listed as follows:

- Comparison of utility between two or more treatment rules in every decision-making stage.
- Optimization of DTR or identifying optimal treatment decision in each decision stages by comparing utility values of each treatments and assigning the one with maximum utility value as optimal.

Thus, the key in estimating optimal regimen is defining the utility functions where the process is data-driven and an extension to single-stage decision problems described earlier. To achieve these goals different utility functions were considered in various literatures in past such as multiple stage-specific regret (loss) functions [37], stage-specific blip functions (welfare contrasts) in structural nested mean models framework [28]. Along with it other method known as *Q*-learning will be implemented for further analysis in this thesis. *Q*-learning uses conditional expectation of primary outcome or potential outcome as the utility function and the potential outcome framework is discussed next.

2.1.3 Framework on Potential Outcomes

The potential outcome framework is used to quantify the result of assigning a treatment in different stages of a dynamic treatment rule. Hence, comparing this outcome value we will be able to estimate the utility values and build an optimized decision rule. The framework was introduced by Neyman [38] for analyzing statistical problems in agricultural experiments where time-dependent randomized trials were considered. This presented framework was further extended by Rubin [39] and Robins [40] in time-dependent randomized trials and observational data in the context of dynamic treatment regimen. Thus, we can define potential outcomes as the set of all outcomes that is obtained when a treatment or a sequence of treatments is administered to an individual patient.

Now, consider a two stage DTR setting where A_1 denotes the treatment at first stage and similarly A_2 denotes that at second stage. Next, we need to consider the baseline information which describes the characteristics of an individual before the treatment at stage 1 is prescribed and it is denoted by X_1. Further, X_2 denotes the additional information such as result of treatment at stage 1 and other biomarkers. Let Y denote the final outcome after treatment at stage 2 is given and it is also our outcome of interest. The observed data trajectory of an individual patient would be (X_1, A_1, X_2, A_2, Y), where we can define potential outcome prior to second stage treatment as $X_2*(a_1)$, if treatment $A_1=a_1$ and that at end of second stage treatment as $Y*(a_1, a_2)$ for treatment sequence of $A_1=a_1$ and $A_2=a_2$.

Furthermore, in this framework of potential outcome following three assumptions are important for estimating effects of dynamic treatment regimens:

- Stable Unit Treatment Value Assumption (SUTVA) [40] states that there should not be any disturbance in treatment between individuals or one patient's potential outcome should not be interfered by treatment allocated to another patient. This assumption provides the stability or consistency in a way where potential outcome will be equal to observed outcome or in other words it maintains the connectivity between potential and observed outcomes. It can also be expressed mathematically as $X_2*(a_1) = X_2$ and $Y*(a_1, a_2) = Y$.
- Next, the assumption of sequential ignorability which is also known as no unmeasured confounding or conditional exchangeability) defines that, depending upon the time-dependent covariates and treatment history up to time t_j, assigning treatment at stage A_j can be made independent of potential

outcomes of the individual. If *j= 1,2,* and for regime (*a₁, a₂*), $A_1 \perp [X_2^*(a_1), Y^*(a_1, a_2)]|H_1$ and $A_1 \perp Y^*(a_1, a_2)|H_2$. The assumption always holds for the process of sequential randomization which is usually performed in the experimental setting of SMART design but must be evaluated according to the problem or observational dataset in hand.

- At last, we need to consider the assumption of positivity, which defines the feasibility of a set of regimes for which treatment history with positive probability of observation should also have positive probability of the treatment results following the decision rule up to time t_j with defined covariate or treatment history. If this assumption is violated, we need to reconsider the treatment regimens as violation of it will make us unable to estimate the effects of DTRs.

Hence, the goal of DTR is to treat a patient with optimal treatment depending upon the characteristics or evidence provided prior to treatment assignment. The Bellman's principle of optimality states that, "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision"[41]. Thus, we can use the theory of dynamic programming where by knowing the functional distribution of the potential outcomes (e.g. $X_2*(a_1)$ and $Y*(a_1, a_2)$) we can estimate the optimal decision resulting maximum average outcome. For implementation of above discussed processes and also making sure that all three assumptions are not violated an experimental design is needed such as SMART which is explained next [42].

2.2 Sequential Multiple Assignment Randomized Trials (SMART) designs

2.2.1 Definition and applications

As observational data are usually high dimensional, and they also tend to violate sequential ignorability assumption of DTR estimation we need to move towards more practical and experimental setting. For this purpose, numerous literature on clinical trial design employing experimental data are present [5, 29-32, 42, 43] that defines construction of sequential multiple assignment randomized trial (SMART) where a patient is randomized more than one time through all possible treatment options at each stage and the treatment resulting maximum utility is selected as optimal treatment that defines DTR. The SMART design offers randomization of treatment options based on the individual patient’s biomarkers and clinical history. The design also supports adaptive interventions which adapts to the system dynamics in a multi-stage trial through a sequence of decision rules that dictates the intervention path in order to maximize long-term primary outcomes [6].

The main difference between Randomized Control Trial (RCT) and SMART design is that the first one makes comparison between two or more treatments, whereas the later compares the treatment regimens and constructs two or more decision rules. The SMART design carries out the trial assignment process and within these assigned trials the physician recognizes an optimal one, that maximizes the patients well-being parameter is assigned to the patient. Hence, the SMART design enables an agent to figure out the best treatment at some treatment stage or decision point, the optimal treatment sequence depending upon response to previous treatments and intermediate outcomes, best timing and modes of treatment delivery, and the process of individualizing sequence of treatments according to biological, diagnostic and other patient information.

Generally, SMART design consists of two stage randomizations, where in first stage patients are randomized to either of two or more treatments and it is followed by periods of patients visit to clinic. The randomization process at second stage depends upon response to previous treatment and patient characteristics over that time period. So, in some SMART design programs a patient may or may not be randomized in the second stage depending upon the response from first stage treatment. Thus, these different types of randomization process differentiate one design from other and types of SMART designs are discussed in next topic.

2.2.2 Types of SMART designs

There are commonly three ways in which SMART designs are constructed and it can be more clearly shown in tree-diagram as in Figures 2.2, 2.3 and 2.4. SMARTs with two stages and two or three treatment decisions per stage are the ideal ones as the trial assignment in these kinds are more feasible and less time consuming. However, designs like this can contain more than two stages and more than two or three treatments for each stage. There is no compulsion that treatment in each stage should be unique for example, in Figure 2.2, treatments C and D can be same as treatments G and H, or E and F can be same as I and J. Similarly, it applies to SMART designs shown in Figure 2.3 and 2.4. Further, the treatment options at first stage and that in second stage can be same e.g., in Figure 2.2, treatments E or F can be same as treatment B and I or J can be same as A.

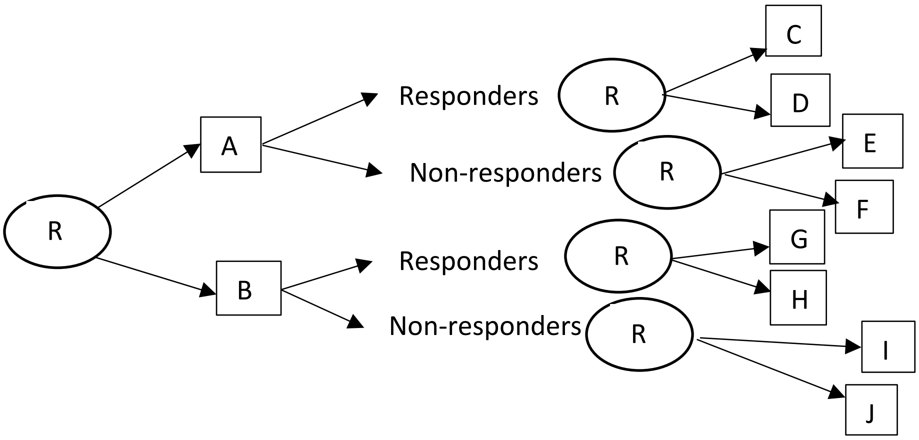


Figure 2.2. SMART design with two treatment options at each decision points where both responders and non-responders are re-randomized to available treatment options depending upon an individual's status [42].

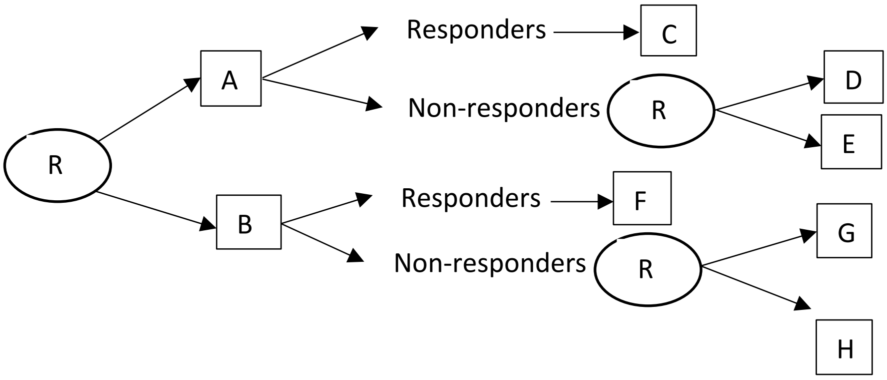


Figure 2.3. SMART where only non-responders are re-randomized [42].

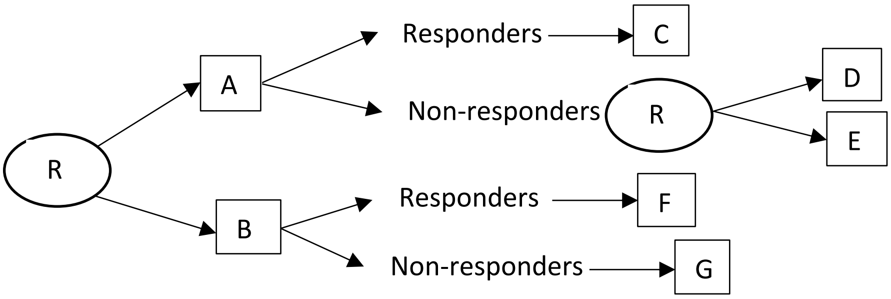


Figure 2.4. SMART where re-randomization depends on both responder status and initial treatment [42].

All above figures represent different types of SMART designs where there is distinction in process of re-randomization represented by letter R. Figure 2.2 shows the design where all patients are re-randomized to available treatments depending upon their response to previous treatments. This type of design was used for trial assignment in alcohol dependency [44] who do not respond to Naltrexone, a placebo treatment for alcoholics. If we observe this design closely we can find that there are eight ways of assigning dynamic treatment regimens embedded within this design.

Similarly, Figure 2.3 shows most general type of SMART design where re-randomization of treatments depends upon the response status of the target group. Thus, in this type of design the responders are continued to a treatment without randomization process whereas, only non-responders are randomized as our goal here is to access best second-stage treatment option for these non-responding groups. These kinds of design are mainly used for trial assignment in areas of ADHD [45], acute myelogenous leukemia [46, 47], small-cell lung cancer [48], neuroblastoma [49, 50], diffuse large cell lymphoma [51], multiple myeloma [52], and metastatic malignant melanoma [53]. There are six embedded dynamic treatment regimes in this type of SMART design.

Lastly, Figure 2.4 shows the next possible type of SMART design in which the non-responders to a particular treatment that was assigned in first stage will only be randomized in second stage of treatment decision. This type of design was used for trial assignment in treatment for nonverbal children who were 5-8 years old with autism spectrum disorders [54] under the project called the Adaptive CCNIA Developmental and Augmented Intervention Study. Therefore, these are the main three types of SMART design popular in medical literatures and implemented in trial assignment for personalized medicine. Further, in next topic we discuss about the design framework of SMART for Sanford profile project.

2.2.3 Design Framework

As we are now familiar to different types of SMART designs and that they are employed for trial assignment in estimation of optimal DTR. However, in Sanford project the observational dataset is restructured according to SMART design as shown in Figure 2.5, where, each subject or patient is randomized to treatment at each decision points among the available treatment options.

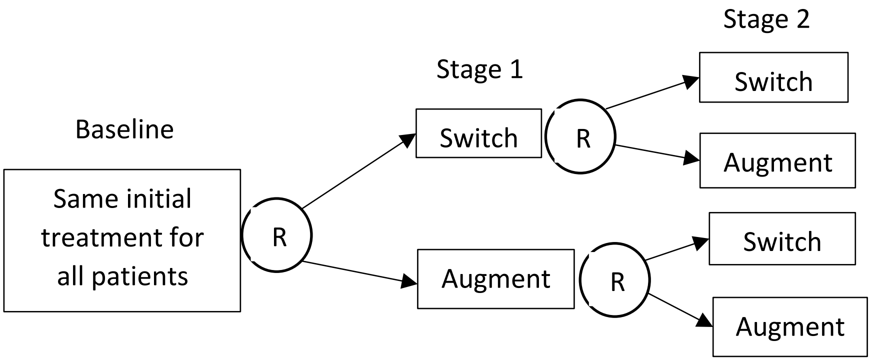


Figure 2.5. SMART design with two randomized stages and two treatment options at each stage. Patients are randomized to treatments from left to right to one of the two treatment options.

In the SMART design shown in Figure 2.5, first all the patients receive same initial treatment called baseline treatment that can be any kind of standard care. Then, after some time period the patients are driven forward to stage 1 in the design where they are randomized to one of the two treatment categories namely “switch” or “augment” current treatment. Again, after another period of time, patients in stage 1 are re-randomized in stage 2 to again either of the two treatments, “switch” or ‘augment”, the current treatment(s) from stage 1. Many different variations exist in designing of SMART, for example, the number of treatments at each stage can be more than two and also there can be more than two stages. However, here we employ a two-stage SMART with randomized binary treatments at each stage for dataset restructuring and making this conducive for applying *Q*-learning, an optimization process for estimation of optimal decision rule that will be discussed on later chapters. As the patients are randomized to binary treatments these intervention options at each stage are coded either -1 or 1. Thus, this type of adaptive intervention setting consists of four decision rules embedded in total.

2.3 Reinforcement learning and Q-learning

2.3.1 History and Definition

Machine learning, a branch of artificial intelligence has become a popular field of research for statisticians and data analyst over last few decades. The field of machine learning that involves stochastic sequential decision process is referred to as reinforcement learning (RL) in the realm of computer science. If we go back in history then we will be aware that the term "reinforcement" was coined from learning behavior of animals in experiments involving animal psychology where it points out the relation between occurrence of event and the response, so there is greater probability that the same response will occur again if the same situation is given. Let's consider x_t as state, a_t as action and r_t as reward from the action being taken in an environment where time t is discrete then the process of reinforcement learning involves:

- Trying a sequence of actions (a_t).
- Recording the consequences or rewards (r_t) of these actions.
- Statistically estimating the relationship between actions (a_t) and their consequences (r_t).
- Finally, selecting the action that produces most favorable consequence.

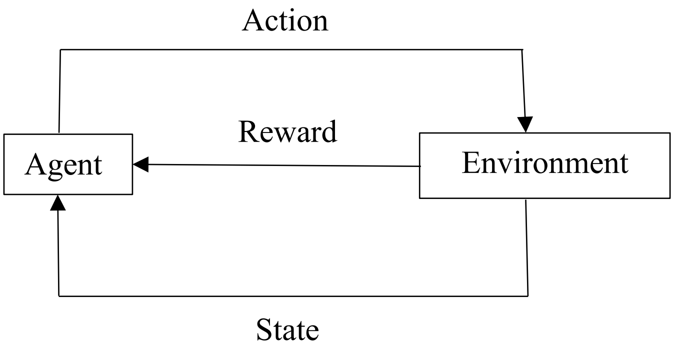


Figure 2.6. Block diagram showing basic processes involved in reinforcement learning.

Thus, as shown in Figure 2.6, reinforcement learning quantizes the interaction between a learning agent and the environment it wants to learn about [55]. In this process, first an agent (physician) observes the status of states and put forward or takes an action (treatment decision) from a set of possible actions. Then, the environment (patient) responds to these actions by observing or outputting a reward (patient's well-being) and makes a transition to new state.

Additionally, from computer science perspective various complication or computational issues may arise when there is an interaction between learning agent and the environment it wants to learn from and in this case reinforcement learning is most promising field to address these issues [55]. Although, most of the optimal control theory and adaptive design requires some model that defines the physical state of the system, reinforcement learning or more specifically Q-learning needs no such model as it is a model-free method that can be used for obtaining personalized therapies. RL is mainly popular in areas of machine learning, operations research, control theory and game theory [56], however, there its popularity has grown also in statistical and biomedical communities that uses RL for optimization of DTR's [57].

Dating back to history, first methods for solving multi-stage decision problems are dynamic programming (DP) algorithms which was introduced by Bellman in 1957 [41]. However, these classical DP techniques has some limitation while they are implied in field of RL. These limitations can be summarized in two points: First, these algorithms require a complete model of system dynamics which is we need to have full knowledge of learning environment and multivariate distribution of data in statistical terms. However, it is very complicated and impractical to have this knowledge in areas of bio-medical and healthcare. Second, DP algorithms are proven to be computationally expensive process and for high-dimensional medical data they may face another problem called “curse of dimensionality”. Anyway, DP is important in a sense that it gave the theoretical foundation for RL processes. Similarly, major breakthrough occurred in the field of RL when Watkins on 1989 [58] introduced the *Q*-learning algorithm, which was implemented to solve multi-stage decision problems depending upon the training data trajectories. Thus, *Q*-learning algorithm is able to solve these issues of traditional DP algorithm and hence, it is also called approximate dynamic programming algorithm.

In the field of health and medical studies, RL has used in treatment of behavioral disorders where patients were administered multiple treatments in different treatment stages [8]. Similarly, *Q*-learning was implemented for defining decision rules in chronic psychiatric disorders [9] and also been successfully applied for segmenting prostrate in transrectal ultrasound images [59]. Thus, summarizing on advantages of RL we can say that this method does not rely on physical dynamics or accurate model for describing time dependent optimal treatment strategies derived through clinical training data. This feature, helps in applying heterogenic treatment across patient that captures the notion of individualized therapies. Next, the process of RL focuses on long-term benefits of a treatment decision rule to an individual by considering response of previous treatment, patient's history and also delayed effects of treatment assigned.

2.3.2 Mathematical Definition

In clinical scenario, reinforcement learning involves trying a sequence of treatment actions, recording the results of these treatments and statistically estimating the relations between these treatments and their results. The treatments assigned to the patient interacts with them and known as the "environment" which may be human body, DNA or proteomics etc. These interactions happen continuously during trial assignment and thus, the environment interacts with the actions and provides the feedback as potential outcomes. Mathematically, let's denote the environment (states) and possible actions ("treatments") as *X* and *A*, respectively. Both of these variables are random and time-dependent thus, $\overline{X_t} = \{X_1, X_2, \dots, X_t\}$. and similarly, define actions as $\overline{A_t} = \{A_1, A_2, \dots, A_t\}$. When the values of random variables *X* and *A* are realized, we denote them in lower case as $\overline{x_t} = \{x_1, x_2, \dots, x_t\}$ and $\overline{a_t} = \{a_1, a_2, \dots a_t\}$. Assume P as distribution of above finite longitudinal trajectories when sampled. The distribution of each present state X_t is conditional on previous state values of ($\overline{X_{t-1}}$, $\overline{A_{t-1}}$). Let's denote these conditional densities as $\{f_1, \dots, f_T\}$ and again the expected values for each distributions *P* is denoted as *E*.

As patients are given a treatment a_t in time t, after each of these time steps of t they receive a numerical reward say r_t which represents patient's status or well-being after that treatment. Mathematically, the reward function is depended upon: previous state $\overline{x_t}$, action $\overline{a_t}$, and current state x_{t+1}, where t= 0, 1, …, *T* and represented as:

$$r_t = R(\overline{x_t}, \overline{a_t}, x_{t+1}) \qquad (2.1)$$

In RL however, to learn what to do when similar events happen in future, first we need to map situations from state space X to actions to be taken from action space A, depending on the goal which may either be to maximize or minimize the expected value of discounted return:

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^T r_{t+T} = \sum_{k=0}^{T} \gamma^k r_{t+k} \qquad (2.2)$$

In equation (2.2), γ denotes the discount rate whose values ranges from $0 \leq \gamma \leq 1$, which means that the future rewards are discounted geometrically depending on the value of γ. The different values of discount rate affect whether the future rewards are taken into consideration or not. For example, if $\gamma = 0$, then in the same equation (2.2) we can observe that $R_t = r_t$, which means only immediate reward is considered, whereas, when $\gamma = 1$ the future rewards are strongly taken into account and under these circumstances the reward function are either maximized or minimized over the long run.

The next important factor of RL algorithm is an exploration "policy" or treatment policy in medical terms. The policy is represented as p and defined as the mapping of state $\bar{x_t}$ and action $\overline{a_{t-1}}$ to the probability $p_t(a| \bar{x_t}, \overline{a_{t-1}})$, which is the probability of action a being taken when given history is $(\bar{x_t}, \overline{a_{t-1}})$. It can be described in other way as the sequence of decision rules for e.g. $\{d_1, \ldots, d_T\}$ and it can also be considered as an action i.e. $\{d_1, \ldots, d_T\} = a_t$ in a nonstationary, non-Markovian but deterministic system. In the training data, if its distribution is denoted by P_d then the expectations with respect to these distributions can be denoted as E_d. The goal of RL study is to find out the treatment that results in maximum reward for the patient or in other words seek for the policy that yields maximum value of expectations with respect to sum of rewards over time.

Another key estimation in RL system is the estimation of value function, which is defined as state or state-action pair function that combines the total reward an agent can gather, considering all expected future reward when starting from a given state. Suppose, D is the set consisting of all policies such that $d \in D$ then the value function denoted by *V(x)* is defined as the sum of expected rewards with initial state *x* and following the policy $d \in D$. Mathematically, value function is denoted as:

$$V(x) = E_d[R_t | x_t = x] = E_d[\sum_{k=1}^{T} \gamma^k \, r_{t+k} | x_t = x] \qquad (2.3)$$

As the state or state-action pairs are time-dependent, value function for a history set $(\overline{x_t}, \overline{a_{t-1}})$ is given by:

$$V_t(\overline{x_t}, \overline{a_{t-1}}) = E_d\left[\sum_{k=1}^{T} \gamma^k r_{t+k} | \, \overline{X_t} = \overline{x_t}, \overline{A_{t-1}} = \overline{a_{t-1}}\right] \qquad (2.4)$$

The main difference between equations (2.3) and (2.4) is the function pair they define, as equation (2.3) defines the state-value functions of policy *d* whereas equation (2.4) defines action-value function for policy *d* [57].

Now, next goal is to estimate the best policy that would maximize the final reward in the long run. Thus, optimal value function can be defined as:

$$V_t^{opt}(\overline{x_t}, \overline{a_{t-1}}) = max_{d \epsilon D} V_t(\overline{x_t}, \overline{a_{t-1}})$$

$$= max_{d \epsilon D} E_d\left[\sum_{k=0}^{T} \gamma^k r_{t+k} | \, \overline{X_t} = \overline{x_t}, \overline{A_{t-1}} = \overline{a_{t-1}}\right] \qquad (2.5)$$

On the basis of equation (2.5), we can define an optimal policy as the policy that results in maximum value of value function $V_t(\overline{x_t}, \overline{a_{t-1}})$. The optimal policy is denoted as d^{opt} and if this policy exists we can further establish the Bellman optimality equation for this optimal policy. The Bellman optimality equation gives the relationship between values of the current state and its successor states and it shows the fact that optimal policy yields the best expected result or is the best action with respect to current state. The Bellman optimality equation for $V_t^{opt}(\overline{x_t},\overline{a_{t-1}})$ can be derived as follows:

$$V_t^{opt}(\overline{x_t},\overline{a_{t-1}}) = max_{a_t}E_{d^{opt}}[\sum_{k=0}^{\infty}\gamma^k r_{t+k}|\,\overline{X_t} = \overline{x_t},\overline{A_{t-1}} = \overline{a_{t-1}}]$$

$$= max_{a_t}E_{d^{opt}}[r_t + \gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|\,\overline{X_t} = \overline{x_t},\overline{A_{t-1}} = \overline{a_{t-1}}]$$

$$= max_{a_t}E\left[r_t + \gamma V_{t+1}^{opt}(\overline{X_{t+1}},\overline{A_t})|\overline{X_t} = \overline{x_t},\overline{A_{t-1}} = \overline{a_t}\right]$$

$$= max_{a_t}\sum_{x'}P_{xx'}^{a}\,[R_{xx'}^{a} + \gamma V_{t+1}^{opt}(x')] \qquad (2.6)$$

Where,

$$P_{xx'}^{a} = \Pr\{\overline{x_{t+1}} = x'|\overline{x_t} = x, \overline{a_t} = a\} \qquad (2.7)$$

$$R_{xx'}^{a} = E[r_t|\overline{x_t} = x, \overline{a_t} = ax_{t+1} = x'] \qquad (2.8)$$

Equations (2.7) and (2.8) denotes two forms of Bellman equations for $V_t^{opt}(\overline{x_t},\overline{a_{t-1}})$. Also, for a policy to be optimal i.e.d^{opt}, it must satisfy:

$$d_t^{opt}(\overline{x_t},\overline{a_{t-1}})\epsilon \arg max_{a_t} E[r_t + \gamma V_{t+1}^{opt}(\overline{X_{t+1}},\overline{A_t})|\overline{X_t} = \overline{x_t},\overline{A_{t-1}} = \overline{a_t}] \qquad (2.9)$$

Thus, above mathematical definition of Reinforcement Learning technique depicts that observing reward of present state and transition to next state does not require knowledge for model of the environment. Both of these processes are determined by the consequences of interaction between environment and the actions taken. This aspect of RL differentiates it from other form of Dynamic Programming.

2.3.3 *Q*-function Estimation

Q-learning is a reinforcement learning technique which targets on estimating and maximizing the above discussed value function, rather than minimizing regret or any other blip function. The blip function is the concept fundamental to DTR estimation and is also known as contrast function which is defined as the difference between expected outcome of a patient under two different treatments [60]. The value functions are also called Q-functions in Q-learning scenario. So, to estimate these Q-functions we should first consider the dimension of state variables in state-space (X) and treatment actions in action-space (A). In order to obtain the estimated values of these Q-functions various approaches such as linear least square regression, Support vector machines regressions, extremely randomized trees, etc. are implemented. However, it has been observed that estimation of these functions is mainly the approximation of least squares value iteration [61-63].

For estimating Q-function, first we need to define an error parameter (θ_t) for the t^{th} Q-function and this parameter should satisfy:

$$\theta_t \epsilon \arg min_\theta E_n[r_t + max_{a_{t+1}} Q_{t+1}(\overline{X_{t+1}}, \overline{A_t}, a_{t+1}; \theta_{t+1}) - Q_t(\overline{X_t}, \overline{A_t}; \theta)]^2 \quad (2.10)$$

Q-learning is a regression-based approximate dynamic programming algorithm that depends on Q-functions where, input to the system are the training data and output is the function approximation for estimating potential final outcome. The SMART design assists in providing the training data as it consists of trial assignment or treatment decision for an individual at different time interval. The number of Q-functions to be estimated depends upon the number of stages in SMART design. Therefore, for a two-stage SMART, we should follow a bottom to top (backwards) approach, i.e. initially second stage Q-function should be approximated then we should move on to do that for first stage. Further, if we consider two treatment options available at each treatment stages then we need to code these treatment options as 1 and 0. Suppose, A_1 gives the treatment decision at stage 1 and A_2 gives that at stage 2, then based on the training data, the regression model at stage 2 or Q_2 can be defined as:

$$Q_2(X, A_2; \theta) = \beta_0 + \beta_1 X + (\beta_2 + \beta_3 X)A_2 \quad (2.11)$$

Where, $\theta = (\beta_0, \beta_1, \beta_2, \beta_3)$ are the values of regression coefficient or intercept values and *X* gives the values of states indicating the response of treatment or summary of side effects up to the end of first decision point. For low dimensional action space, it is conducive to implement multiple regression models for function approximation however, as the dimension increases it becomes necessary to move towards quadratic or higher order regression analysis.

2.4 Probabilistic Framework

The number of stages in a RL problem, where in stage there is interaction between the agent and the environment can be of finite or infinite numbers. However, the infinite-horizon problem is beyond the scope of this thesis. So, let's consider a RL problem with finite number of stages say *K* and let *j* be one of the stages within *K* where, $1 \leq j \leq K$. Thus, at stage *j* suppose the agent observes a state O_j which may belong to a vector consisting of discrete or continuous variables and to have further interaction to the environment the agent then executes an action A_j which should belong to a vector of discrete variables. The interaction between agent and the environment through the executed action results in a real-valued reward say Y_j. After, this interaction the agent moves on to the next stage. Here, as the problem is of finite- horizon we can define $\bar{O}_j = (O_1, \dots, O_j)$ and $\bar{A}_j = (A_1, \dots, A_j)$. Now, the history set H_j can be defined as the vector of all the covariates information consisting the elements say $(\bar{O}_j, \bar{A}_{j-1})$ at stage j. Then, the reward can be denoted as the function of history set H_j, the current action executed A_j and the transition to next state O_{j+1} i.e.

$$Y_j = Y_j(H_j, A_j, O_{j+1}) = Y_j(\bar{O}_{j,}\bar{A}_j, O_{j+1}) \quad (2.12)$$

In statistical term, the reward is considered like potential outcome and in some cases, there can be only one ultimate reward with all previous rewards assumed to be 0.

Now, let's define a policy d as a vector of all the decision rules and are determined through mapping from history space (*H_j*) to the action space (*A_j*) i.e. $d_j: H_j \rightarrow A_j$, for $1 \leq j \leq K$. For a stochastic process the policy defines the mapping of the history space to the space of probability distributions of the action space and is denoted as $d_j(a_j|h_j)$. Also, the policy space can be defined as the function space of collection of these policies that are mapped between history and Action space, this function space is denoted as *D*.

Furthermore, let's consider a finite-horizon trajectory of training data set as $\{O_1, A_1, O_2, \ldots, A_K, O_{K+1}\}$. The training dataset consists of the records for *n* number of individuals, so, we will have *n* number of these trajectories. If the subjects are sampled randomly following some fixed probability distribution say P_π. However, the probability distributions of each O_j that are conditional on ($H_{j\text{-}1}$, $A_{j\text{-}1}$) are unknown thus, suppose these conditional densities as $\{f_1, \ldots, f_K\}$ and corresponding policies as $\pi = (\pi_1, \ldots, \pi_K)$, then depending on history H_j the probability that action a_j is taken is given by $\pi_j(a_j|H_j)$. We consider that all the actions have positive probability of being executed. Then, the likelihood of trajectory $\{o_1, a_1, o_2, \ldots, a_K, o_{K+1}\}$ under the probability distribution P_π is given by:

$$f_1(o_1)\pi_1(a_1|o_1)\prod_{j=2}^{K} f_j(o_j|h_{j-1}, a_{j-1})\pi_j(a_j|h_j)f_{K+1}(o_{K+1}|h_K, a_K) \tag{2.13}$$

Now, we denote the expectation value of the policy with respect to distribution $P_\pi$as E_π. Again, let's denote the distribution of an arbitrary policy $d = (d_1, \ldots, d_K)$ as P_d and this policy is also responsible for action generation. So, *d* is the deterministic policy and the likelihood of trajectory $\{o_1, a_1, o_2, \ldots, a_K, o_{K+1}\}$ under distribution P_d is given by:

$$f_1(o_1)\mathbb{I}[a_1 = d_1(o_1)] \prod_{j=2}^{K} f_j\left(o_j \middle| h_{j-1}, a_{j-1}\right)\mathbb{I}[a_j = d_j(h_j)] f_{K+1}(o_{K+1}|h_K, a_K) \tag{2.14}$$

Also, if we consider the policy *d* as a stochastic process then the likelihood becomes:

$$f_1(o_1)d_1(a_1|o_1) \prod_{j=2}^{K} f_j\left(o_j \middle| h_{j-1}, a_{j-1}\right) d_j(a_j|h_j) f_{K+1}(o_{K+1}|h_K, a_K) \tag{2.15}$$

The expectation with respect to the distribution P_d is denoted by E_d. Then, the goal of statistical RL is to learn an optimal policy say d^* that has the greatest possible expected value within that class.

The value function can be defined as total expected future reward from a particular starting state and then after choosing actions according to some policy. Thus, at state o_1 with respect to an arbitrary policy *d* we can denote the value function as follows:

$$V^d(o_1) = E_d[\sum_{j=1}^{K} Y_j(H_j, A_j, O_{j+1})|O_1 = o_1] \tag{2.16}$$

When considered *j* stages, the value function for history h_j is the total expected rewards from stage j $(1 \leq j \leq K)$onwards and is denoted as:

$$V_j^d\left(h_j\right) = E_d[\sum_{k=j}^{K} Y_k(H_k, A_k, O_{k+1})\middle|H_j = h_j] \tag{2.17}$$

Then, we set $V_{K+1}^d(\cdot) = 0$ and by definition we know $V_1^d(\cdot) = V^d(\cdot)$, the value functions can be now recursively expressed as:

$$V_j^d(h_j) = E_d[\sum_{k=j}^{K} Y_k(H_k, A_k, O_{k+1})|H_j = h_j] \tag{2.18}$$

$$= E_d[Y_j(H_j, A_j, O_{j+1})|H_j = h_j] + E_d[\sum_{k=j+1}^{K} Y_k(H_k, A_k, O_{k+1})|H_j = h_j]$$

$$= E_d[Y_j(H_j, A_j, O_{j+1})|H_j = h_j] + E_d[E_d[\sum_{k=j+1}^{K} Y_k(H_k, A_k, O_{k+1})|H_{j+1}]|H_j = h_j]$$

$$= E_d[Y_j(H_j, A_j, O_{j+1})|H_j = h_j] + E_d[V_{j+1}^d(H_{j+1})|H_j = h_j]$$

$$= E_d[Y_j(H_j, A_j, O_{j+1}) + V_{j+1}^d(H_{j+1})|H_j = h_j], 1 \leq j \leq K$$

Finally, the optimal treatment strategy can be defined under the value function as:

$$V_j^{opt}(h_j) = max_{d \in D} V_j^d(h_j) \tag{2.19}$$

The optimal value functions also satisfy the Bellman equation as:

$$V_j^{opt}(h_j) = max_{a_j \in A_j} E[Y_j(H_j, A_j, O_{j+1}) + V_{j+1}^{opt}(H_{j+1})|H_j = h_j, A_j = a_j] \tag{2.20}$$

Also, the value of policy *d* denoted as V_d is given by taking the average value or marginal mean outcome over all possible initial observation and can be expressed as:

$$V^d = E_{O_1}[V^d(O_1)] = E_d[\sum_{k=1}^{K} Y_k(H_k, A_k, O_{k+1})] \tag{2.21}$$

The probabilistic framework discussed above considers a classical RL approach where optimal rule is chosen as the one that maximizes the value function. However, we can consider *Q*-function that are nothing but action-value functions where "*Q*" stands for quality of actions and can be considered as a substitute to V^d. Thus, the *Q*-function at stage *j* considering the policy as *d* can be defined as the total expected future reward starting from history set h_j and undergoing actions a_j according to the policy d. Mathematically,

$$Q_j^d\left(h_j, a_j\right) = E[Y_j\left(H_j, A_j, O_{j+1}\right) + V_{j+1}^d(H_{j+1})|H_j = h_j, A_j = a_j] \tag{2.22}$$

Also, the optimal Q-function at stage j can be expressed as:

$$Q_j^{opt}\left(h_j, a_j\right) = E[Y_j\left(H_j, A_j, O_{j+1}\right) + V_{j+1}^{opt}(H_{j+1})|H_j = h_j, A_j = a_j] \tag{2.23}$$

Therefore, in medical decision-making scenario it is a subject of extreme interest in estimating the value of Q_j^{opt}, which can directly estimate the optimal policy and enables an agent for choosing an optimal treatment decision.

CHAPTER 3

METHODOLOGY

3.1 Training Data Acquisition

The training data are the clinical trials from Sanford profile health and they consist of body weight of patients over multiple time points, resulting in a dataset which consists of trajectories with patient baseline weight, weight after 4 months and final weight after 12 months. There are two treatment decision points at 4 months and 12 months period and the dataset also consists of other patient characteristics such as gender, age, race, etc. Although there are various possible ways for data collection Clinical trials can be taken as very reliable source in case of applying reinforcement learning approaches. Also, the block diagram for visualizing the process or methodology used in this thesis for obtaining optimal DTR can be shown as:

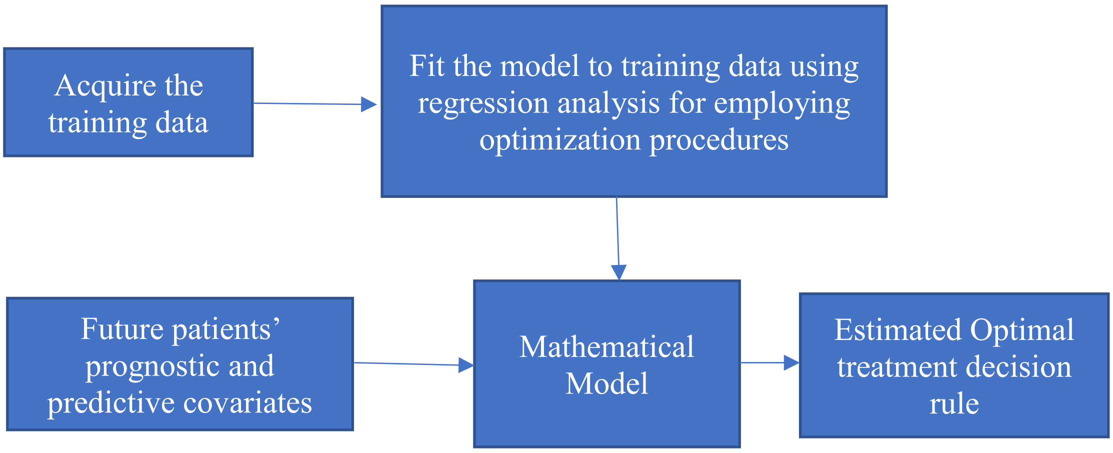

Acquire the training data
Fit the model to training data using regression analysis for employing optimization procedures
Future patients' prognostic and predictive covariates
Mathematical Model
Estimated Optimal treatment decision rule

Figure 3.1. Block Diagram showing the process of building a Mathematical Model for estimating optimal DTR.

For deriving the data using clinical trials design, "Sequential Multiple Assignment Randomized Trial" (SMART) design method is very promising as suggested by various studies [8, 31, 64]. This type of trial design method pairs the treatment decisions or in RL term, actions to their corresponding results or clinical outcomes (states). There are multiple stages where an agent or clinician should make treatment decisions, at each stage one treatment is randomly assigned with probability 0.5 to participants then the result is observed. So, an individual is randomized through different treatment plans which enables a decision-maker to observe the final outcomes considering all possible treatment patterns. For example, if we consider a trial with three stages namely pre-treatment (S_0), mid-treatment (S_1) and post-treatment (S_2) and two treatment decisions at each stage that are actions a_1 and a_2. Then, for pre-treatment stage we may randomly choose one treatment decision for some patients and another treatment for some others and evaluate the initial results observed from these treatments. After the first treatment stage in design we further randomize treatments (a_1, a_2) for patients to observe the treatment outcomes under these stages. At last, we would have then assigned all possible four patterns of treatment assignment randomly to a group of patients and observed their outcomes. Hence, after we have performed this trial assignment we can observe a training dataset which can be further used to define an optimal treatment decision for an individual and it will be discussed later in this chapter.

As discussed earlier the design of this thesis focuses on weight management treatment plan for patients enrolled in Sanford Profile Health. Therefore, the important goal here is to prescribe an optimal weight management treatment plans for an individual with certain attributes and prove that this treatment plans will work best for her/him depending upon her/his characteristics. In this scenario, for obtaining the training data the restructuring of Sanford Health Data according to SMART design is the preliminary task for data acquisition. Thus, the dataset consists of baseline_weight, month4_weight and month12_weight which are patients' pre-treatment weight, mid-treatment weight and post-treatment weight respectively. The dataset also consists of the information about treatment assigned at stages A_1 and A_2, which are the mid-treatment and post-treatment decisions respectively. As, there are two treatment options are each stage these two treatments are coded as 1 and -1. Various patient's attributes such as gender, race, heights are also available in the dataset.

3.2 Model Definition

3.2.1 Mathematical Framework

3.2.1.1 State and Action Modelling

In many medical settings representing state space is very important for defining a mathematical model as general medical outcome are mixed values of discrete categorical and continuous variables. The state space in medical scenario are typically of high dimension and this may pose several difficulties in state space representation such as:

- State that can define sufficient statistic for the problem.
- Effect in modelling due to irrelevant state variables.
- Curse of dimensionality due to high dimensional spaces.
- Need of defining appropriate state variables.

A state model can be defined as one having sufficient statistic in a statistical sense if it can specify the relevant parameters, completely of the associated distributions with the help of comprehensive information that it should contain [65]. Further, as we are considering the context of RL, a state representation should be able to sufficiently specify the distribution of future rewards and state transitions. Also in RL, policy *d* is the mapping from state to action space, therefore if the state lacks the sufficient information about the associated distributions then the policy also may lack quality. In Medical terms, it is very complicated to know the sufficient quantity or quality of state variables so more or less we have to rely on our intuition for selecting appropriate state variables.

Therefore, we need to be careful in including those state variables that are relevant in defining the overall statistic of the model and avoiding source of errors and data inefficiency. In RL problems as the number of state variables grows, the system tends to be affected by “curse of dimensionality”, which explains that the number of states increase exponentially with respect to number of dimensions in the system. The effect of this curse can also be observed while increasing the number of data to obtain a particular confidence boundary [65]. Discretizing the state variable may seem effective in tackling these drawbacks, however, it should be performed very carefully, which requires complete knowledge of specific limitations in behavior of relevant state variables. In some cases, the discretization method may not be very complicated as for e.g. mapping of blood pressure into hypotension, prehypertension, etc., on other hand, in many cases this type of categorization may not be straightforward which can introduce bias in value function estimation. Thus, in that case, methods such as function approximation and other regularization techniques are robust and are important in reducing the effect of dimensionality and overall presence of irrelevant state variables in the system, so, these methods are of great help when such obstacles are encountered [66].

The mathematical model for Sanford Profile defines five state variables namely, gender, race, parent_BMI, baseline_BMI and month4_BMI where gender and race gives the respective information about an individual coded as 0 and 1. Similarly, parent_BMI defines the averaged Body Mass Index (BMI) of patient’s parents and baseline_BMI, month4_BMI are the pre-treatment and mid-treatment BMI values of the patient. There are also two treatment decision points and two treatment options coded as 1 and -1 at each of these decision points.

After defining the state space next, we need to define the set of action space or treatment interventions to be precise. The set of actions may change as we move from one stage of randomization to another in a SMART trial design. Generally, majority of DTRs consists of small and discrete set of actions for e.g. Treatment 1 vs. Treatment 2. However, we may consider some cases where the action space is continuous which is beyond the scope of this thesis. Action space being continuous also poses numerous problems in trial design and further optimization of DTR. As we know that RL problem tries to optimize the DTR setting by maximizing the outcome over action space after each iteration. Therefore, maximizing outcome over continuous action space can introduce bias in learning as it requires numerical approximation. In addition to that, the RL algorithm randomizes all possible treatment or actions and selects the one with maximum reward, so, exploring or randomizing in continuous space is numerically infeasible. Thus, discrete action space with few dimensions generally results in rapid and confident attainment of RL solutions.

In our case, there are two treatment actions available at each stage in the SMART design. The decisions are denoted as A_1 and A_2 in the first stage and second stage decision respectively of the model and the treatments are coded -1 and 1 defining two different types of actions in an action space.

3.2.1.2 Time Horizon

After defining state and action spaces we need to define the time limit or choose the time horizon for the mapping between these spaces to continue in the system.

In RL problem the time horizon can be categorized into either “finite” or “infinite” cases. In case of finite time horizon, the problem of decision making terminates after some finite number of time period or steps. In most of the cases the number of time steps or at least the upper bound of time period are known in advance to the agent. Any kind of medical therapy that aims in moving patients from “bad” to “good” state can be thought of as a finite time horizon as in these cases the goal is to cure a disease or be at remission.

Alternatively, in case of treatments with short time steps and those with chronic conditions it is beneficial to assume an infinite time horizon of treatment decisions. The RL problem in this scenario should be able to operate and provide decision rules indefinitely. An example for infinite time horizon case can be one where response to treatment of a patient may be unstable and due to this a continual treatment is required, otherwise the patient condition may deteriorate and may end up in state of relapse.

In our case as the stages of treatment are not infinite and there are two possible treatment options at each stage thus, it is a finite horizon problem with definite number of time steps. As the problem statement of the research suggests estimating the optimal treatment decision rule for patients’ weight management, the treatments decision made should work in patients’ well-being and after the state of well-being or weight management is obtained these treatments are not continued, however, regular exercise and balanced diet are essential for sustaining the lost weight.

3.2.1.3 Reward Function

In RL problem, the result of mapping between state and action space is eventually depicted upon the reward function, from which an agent can estimate the cost or utility of employing an action at some stage of clinical trial. The reward function can be linear, nonlinear or discontinuous, however, only requirement is that it should bounded by entire state and action space. Studies also show that the choice of reward function effects the learning rates in RL algorithms [67, 68].

An efficient reward function should always clearly reflect the desired goal and can only be a simple function for e.g. in game playing an agent can define a reward function as it wins or loses the game [69, 70]. Mathematically, the reward function in this case can be simply defined as:

$$R(s,a) = \begin{cases} 1 & agent\ wins\ the\ game\ at\ state\ s, \\ 0 & otherwise. \end{cases} \tag{3.1}$$

Above equation shows a simple reward function where choosing a reward is based upon whether an agent wins or loses the game. The agent is rewarded 1 if it wins whereas it is awarded 0 otherwise.

Similarly, reward function in medical setting should define the tradeoffs between the costs of treatments and costs of symptoms. For instance, in HIV model [71] following reward function was considered:

$$R(s,a) = c_1 a_1^2 + c_2 a_2^2 + c_3 s_V + c_4 s_E \tag{3.2}$$

In above equation (3.2), a_1 and a_2 are the real-valued actions that represents the drug dosage levels similarly, s_V and s_E are the state variables that denotes viral load and immune response, respectively. Also, the coefficients c_1 to c_4 are constants whose values should define the priorities of the agent and should narrow the differences in range between state and action spaces. In this scenario, it is favorable to an agent in minimizing the treatment and viral load whereas maximizing a good immune response. Thus, to achieve this goal through above reward function coefficients c_1, c_2 and c_3 must be negative whereas, c_4 should be positive.

In case of Sanford Health Data, gender, race, baseline_BMI, parent_BMI, and month4_BMI are the state variables and two treatments each at stages A_1 and A_2 are the actions. Then here in this model we can define the reward function as:

$$R_2(S,A) = \beta_1 S_g + \beta_2 S_p + \beta_3 S_4 + \beta_4 A_2\left(S_p + S_4\right) \qquad (3.3)$$

$$R_1(S,A) = \beta_5 S_g + \beta_6 S_r + \beta_7 S_p + \beta_8 S_b + \beta_9 A_1(S_g + S_p) \qquad (3.4)$$

Equation (3.3) denotes the second stage reward function or which can also be viewed as second stage value function in Q-learning algorithm that will be discussed later. In this reward function, S_g, S_p and S_4 are the state variables representing gender, parent_BMI and month4_BMI respectively. Similarly, A_2 is the treatment action undertaken at stage 2 and $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients of respective state variables. Additionally, as we have considered a two stage SMART trial design there should be a reward function defined for stage 1 of trial assignment as well which is denoted by equation (3.4). In this function, state variables S_r and S_b representing race and baseline_BMI are included and A_1 defines the treatment action taken at stage 1 also, $\beta_5, \beta_6, \beta_7, \beta_8, \beta_9$ gives the coefficients for each of state variables and interaction between them. Now, the next goal is to estimate these coefficients through regression analysis and then approximate the above functions or estimate the values for interaction between action space and state space.

3.2.1.4 Function Approximation Algorithm

After the representation of state and action space, choice of favorable time horizon and assignment of effective reward function, we further need to implement one of the many RL algorithms for representing and estimating the value function. In case of continuous state space and discrete action space with appropriate dimension we can use a simple tabular Q-function approximation algorithm. Also, in case of high dimensional dataset it is beneficial to use methods which can eradicate the issues that can occur due to this high dimensionality and heterogeneity of the data. However, in both cases we need to be sure about selecting the proper state space variables and perform efficient trial assignment for good approximation of these value functions.

As we described previously that Q-learning is an efficient reinforcement learning technique that is used to estimate and maximize the value function and this algorithm also estimates the policy that maximizes the value of expected future reward by relating the state and action space through function approximation.

First, to elucidate the idea, we will describe Q-learning for two treatment stages and then also generalize it to K stages $(K \geq 2)$. Q-learning algorithm involves backward induction process so the function approximation of last intervention or second stage in our case is initiated at first, this serves to control for effects of both past and subsequent adaptive intervention options. In a two-stage SMART study, training data for a single patient follows the trajectory $D = \{(X_{1,i}, A_{1,i}, X_{2,i}, A_{2,i}, Y_i)\}_{i=1}^{n}$. Where the longitudinal data D consists of independent identically distributed copies of the quintuple (X_1, A_1, X_2, A_2, Y), that gives the data collected on single subject [14]. Each quintuple is called trajectory as they are time ordered, for example, if a trajectory defined as $X_1 \in \mathbb{R}^{p_1}$ is the baseline covariate information, then $A_1 \in \{-1,1\}$ is the first treatment option, $X_2 \in \mathbb{R}^{p_2}$ is the covariate information collected between first and second treatment assignments, denoting predictive variables, $A_2 \in \{-1,1\}$ is the second stage treatment and finally, $Y \in \mathbb{R}$ is the primary outcome response variable or terminal reward. Y_1 and Y_2 can be observed at the end of each stage in two-stage DTR policy, however, in case of single terminal outcome Y can be viewed as a reward where $Y_1=0$ and $Y_2=Y$. The baseline covariates are the quantitative or qualitative variables that are measured before randomization process which influences the value of primary outcome variable, Y that is to be analyzed [72]. For notational easiness and compactness, we represent information available prior to the t^{th} treatment assignment by H_t. Hence, H_1=X_1 and $H_2 = (X_1^T, A_1, X_2^T)^T$. As we consider a two-stage intervention we need to define two stage Q-functions as:

$$Q_2(h_2, a_2) = E(Y|H_2 = h_2, A_2 = a_2) \tag{3.5}$$

$$Q_1(h_1, a_1) = E(max_{a_2 \in \{-1,1\}} Q_2(H_2, a_2)|H_1 = h_1, A_1 = a_1) \tag{3.6}$$

Thus, a two-stage DTR consists of two decision rules suppose, (d_1, d_2) where $d_j(H_j) \in \{-1,1\}$. Now, to estimate the optimal DTR, $d^{opt} = (d_1^{opt}, d_2^{opt})$ first, we need to define the optimal Q-functions for two stages treatment decisions which can denoted as follows:

$$Q_2^{opt}(h_2, a_2) = E[Y_2|H_2 = h_2, A_2 = a_2] \tag{3.7}$$

$$Q_1^{opt}(h_1, a_1) = E[Y_1 + max_{a_2} Q_2^{opt}(h_2, a_2)|H_1 = h_1, A_1 = a_1] \tag{3.8}$$

After approximating above Q-functions using regression analysis which will be discussed later we can estimate the optimal DTR i.e. (d_1^{opt}, d_2^{opt}), using backward induction as in dynamic programming as:

$$d_j^{opt}(h_j) = \arg max_{a_j} Q_j^{opt}(h_j, a_j) \ , j = 1,2 \tag{3.9}$$

In general, function Q_t (h_t, at) measures the quality of treatment $a_t$when this treatment is assigned to a patient with history h_t. Here, the Q- functions at both stages are defined as unknown conditional expectations where second stage Q-function, $Q_2(h_2,a_2)$ is the conditional expectation of potential response Y when treatment a_2 is assigned to a patient with history h_2. Similarly, in stage 1 Q-function, $Q_1(h_1,a_1)$ measures the quality of assigning treatment a_1 to the patient with characteristics defined by set h_1, where, the predicted future outcome $\tilde{Y}$is given by the maximum value of $Q_2(h_2,a_2)$ i.e.$\tilde{Y} = max_{a_2 \in \{-1,1\}} \hat{Q}_2(h_2,a_2, \hat{\beta}_2)$. To obtain the values of these unknown conditional expectations we can use linear regression model for curve fitting or function approximation, it is common practice to use linear model for Q-functions represented as: $Q_t(h_t;\ a_t;\ \beta_t) =\ h_{t,0}^T\beta_{t,0} + a_t h_{t,1}^T\beta_{t,1}$, where $h_{t,0}$and $h_{t,1}$ are the same subvectors of h_t and $\beta_t = (\beta_{t,0}^T, \beta_{t,1}^T)^T$. The Q-learning algorithm using linear models for the Q-functions can be summarized in following three steps:

- Estimate β_2 and then, Q_2 via least-squares regression of Y on H_2 and A_2 using the following model:

$$\hat{\beta}_2 = \arg min_{\beta_2} \sum_{i=1}^{n} \{Y_i - Q_2(H_{2,i}, A_{2,i}; \beta_2)\}^2 \tag{3.10}$$

- Calculate predicted future outcomes $\tilde{Y}$ assuming optimal second-stage decisions,

$$\tilde{Y} = max_{a_2 \in \{1,-1\}} Q_2\left(H_2, a_2; \hat{\beta}_2\right) = H_{2,0}^T\hat{\beta}_{2,0} + |H_{2,1}^T\hat{\beta}_{2,1}| \tag{3.11}$$

  Then estimate β_1 , and hence Q_1, again using least-squares regression of $\tilde{Y}$ on H_1 and A_1 using the model,

$$\hat{\beta}_1 = \arg min_{\beta_1} \sum_{i=1}^{n} \{\tilde{Y}_i - Q_1(H_{1,i}, A_{1,i}; \beta_1)\}^2 \tag{3.12}$$

- Calculate the estimated Q-learning optimal treatment policy, $d_j^{opt} = (d_1^{opt}, d_2^{opt})$ as,

$$d_j^{opt} = \arg max_{a_2 \in \{-1,1\}}\, Q_t(h_t, a_t, \widehat{\beta_t}) \tag{3.13}$$

Now, the above process can be generalized to $K>2$ number of stages, where first we need to define the optimal Q-function by using backward induction as:

$$Q_j^{opt}\left(H_j, A_j\right) = E\left[Y_j + max_{a_{j+1}} Q_{j+1}^{opt}\left(H_{j+1}, a_{j+1}\right)\middle| H_j, A_j\right], \quad j = 1, \dots, K \tag{3.14}$$

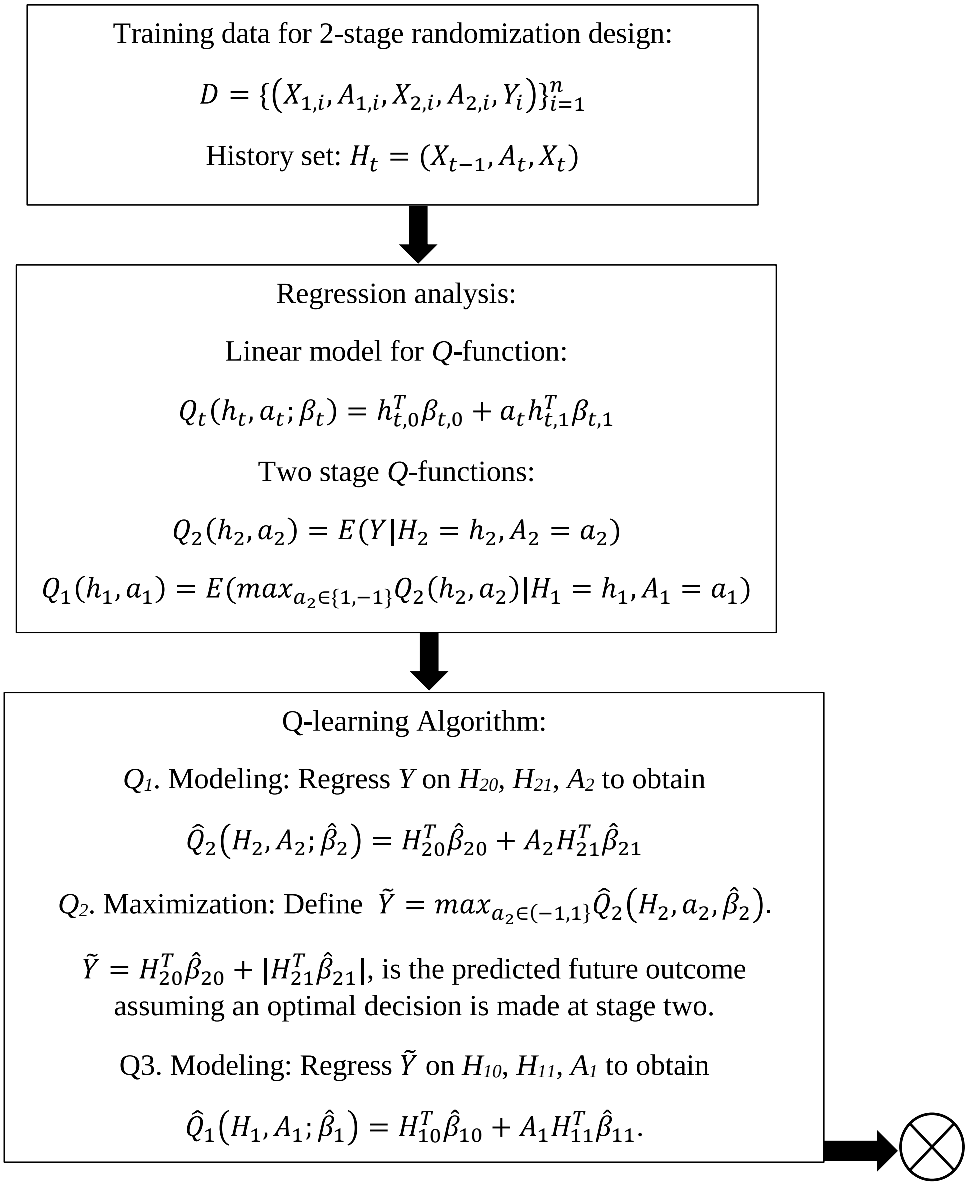
Also, for values of *j=K*, *K-1, …, 1*, as we are moving backward through the stages the regression parameter can be estimated using:

$$\widehat{\beta_J} = \arg min_{\beta_j} \frac{1}{n}\sum_{i=1}^{n}(Y_{ji} + max_{a_{j+1}} Q_{j+1}^{opt}\left(H_{j+1}, a_{j+1}; \hat{\beta}_{j+1}\right) - Q_j^{opt}(H_{ji}, A_{ji}; \beta_j))^2 \quad (3.15)$$

Finally, the optimal DTR for K stages, i.e. $(\hat{d}_1^{opt}, \dots, \hat{d}_K^{opt})$ can be obtained as:

$$\hat{d}_j^{opt}(h_j) = \arg max_{a_j} Q_j^{opt}\left(h_j, a_j; \hat{\beta}_j\right), \quad j = 1, \dots, K \qquad (3.16)$$

The Flowchart for above mentioned algorithm is presented as follows:



Training data for 2-stage randomization design:

$D = \{\left(X_{1,i}, A_{1,i}, X_{2,i}, A_{2,i}, Y_i\right)\}_{i=1}^{n}$

History set: $H_t = (X_{t-1}, A_t, X_t)$

Regression analysis:

Linear model for *Q*-function:

$Q_t(h_t, a_t; \beta_t) = h_{t,0}^T\beta_{t,0} + a_t h_{t,1}^T\beta_{t,1}$

Two stage *Q*-functions:

$Q_2(h_2, a_2) = E(Y|H_2 = h_2, A_2 = a_2)$

$Q_1(h_1, a_1) = E(max_{a_2 \in \{1,-1\}} Q_2(h_2, a_2)|H_1 = h_1, A_1 = a_1)$

Q-learning Algorithm:

Q1. Modeling: Regress *Y* on *H20*, *H21*, *A2* to obtain

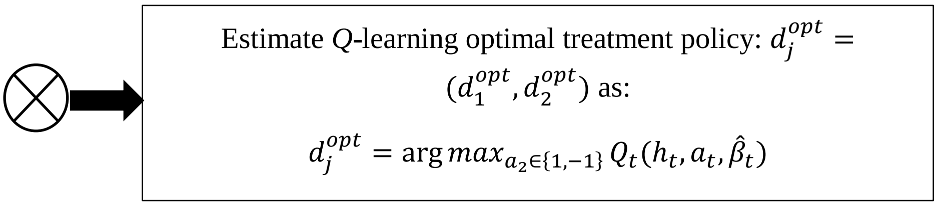
$\hat{Q}_2\left(H_2, A_2; \hat{\beta}_2\right) = H_{20}^T\hat{\beta}_{20} + A_2 H_{21}^T\hat{\beta}_{21}$

Q2. Maximization: Define $\tilde{Y} = max_{a_2 \in (-1,1\}} \hat{Q}_2\left(H_2, a_2, \hat{\beta}_2\right)$.

$\tilde{Y} = H_{20}^T\hat{\beta}_{20} + |H_{21}^T\hat{\beta}_{21}|$, is the predicted future outcome assuming an optimal decision is made at stage two.

Q3. Modeling: Regress $\tilde{Y}$ on *H10*, *H11*, *A1* to obtain

$\hat{Q}_1\left(H_1, A_1; \hat{\beta}_1\right) = H_{10}^T\hat{\beta}_{10} + A_1 H_{11}^T\hat{\beta}_{11}$.



Estimate Q-learning optimal treatment policy: $d_j^{opt} = (d_1^{opt}, d_2^{opt})$ as:

$$d_j^{opt} = \arg max_{a_2 \in \{1,-1\}} Q_t(h_t, a_t, \hat{\beta}_t)$$

Figure 3.2. Flowchart for Regression analysis and Q-learning estimation

3.3 Model Implementation

3.3.1 Framework on Sanford Health Data

Profile by Sanford is a personalized weight management plans that combines healthy grocery food with nutritious meal replacement products. These plans are created by physicians and researchers. In this weight management plans, there are mainly three core principles, which are:

- Nutrition
- Activity
- Lifestyle

The Profile is a personalized plan as the meal plan and activities are developed for each profile member. The main steps involved in this weight management plans are:

- Reduce, in which food with low carbohydrate but high protein is given to the patients.
- Adapt, where more healthy foods are introduced after certain period of time.
- Sustain, where the lost weight is maintained by means of exercise and careful diet.

Major protocols in personalized meal plan under Profile by Sanford can be listed as follows:

- Reboot Protocol:

  The reboot Protocol focuses the body to burn the stored fats by means of physical exercise or meal replacements. There are mainly five profile meal replacement plans consisting of lean and green evening meals, snacks and vegetables.

- Balance Protocol:

  The Balance protocol takes in consideration the members or patients with special medical or Dietary requirements and takes a balanced approach accordingly.

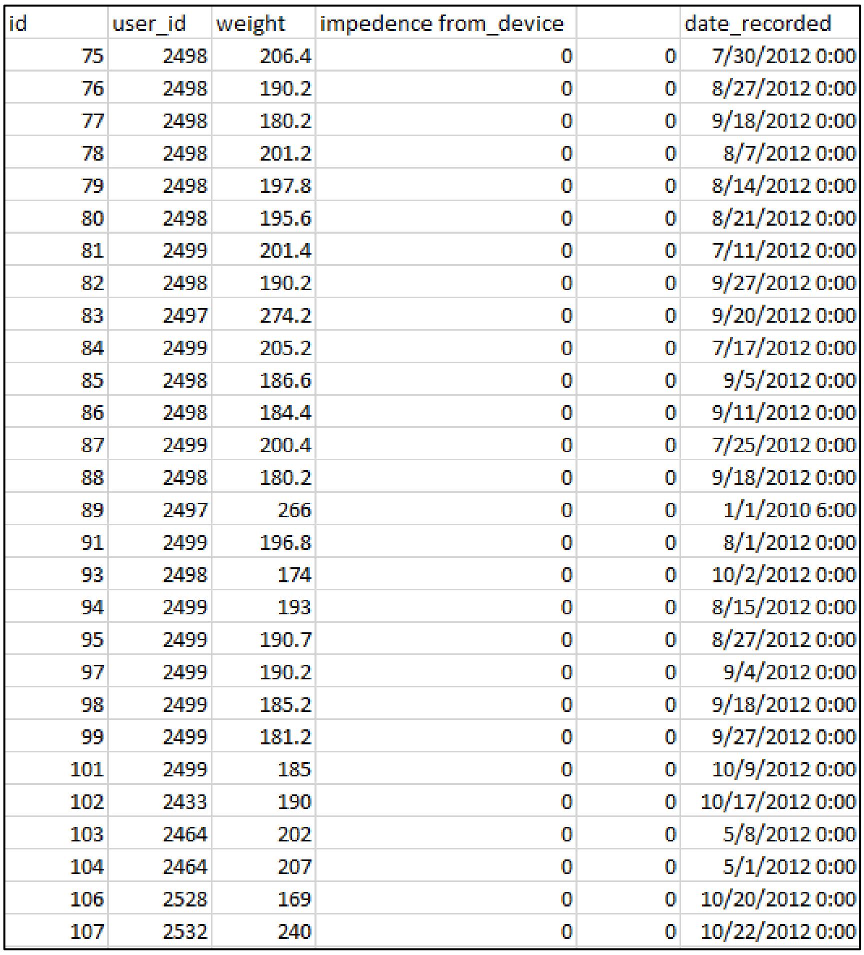
- Additional Protocol:

  This protocol provides special plans for pregnant and nursing moms as well as teens who are obese.

The Sanford Profile data used for this research needed a serious restructuring for obtaining an appropriate training dataset which would be conducive for application of algorithm that can estimate the optimal DTR. Initially, the dataset consisted of patients' id, their respective weights and the date these weights were recorded. On the top of that the weights were not even arranged according to patients' id, so, one patient's weight taken at point t_1 may appear at top of dataset whereas that taken at different point t_2 may appear later in it. Thus, the first modification needed was on the dataset, to arrange the weights according to their respective user-id or patient's id. After this modification was performed using MATLAB different state variables were further added along with weights measured at baseline (before any treatment is administered), after 4 months (after treatment at first stage) and after 12 months (after second stage treatment). Similarly, the BMI of each patients for each measured weight was also calculated using equation (3.17) and also the treatment options at each treatment stages were randomized for every subject.

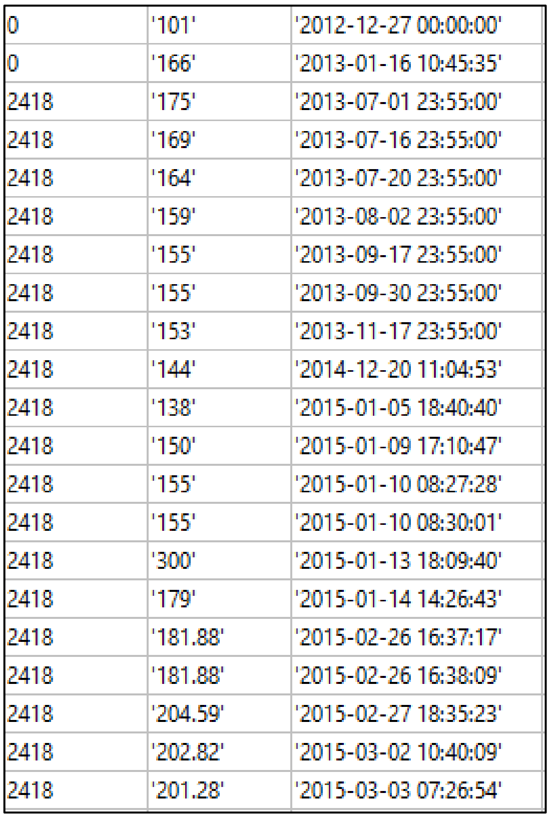
$$BMI = \frac{mass_{lb}}{height_{in}^{2}} \times 703 \tag{3.17}$$

Figure 3.2, shows the initial dataset by Sanford Profile which required a serious arrangements and restructuring. Arrangement of weights according to respective patient’s id was performed using Matlab programming which involved using cell format for each patient and also noting the weights after 4 months and 12 months according to the dates of weight measurement available in the dataset. After, the arrangement the restructuring process the dataset is shown in figure 3.3, where in first column the patient id was arranged and second column gave the respective weight in lb taken at date given by the third column.



id	user_id	weight	impedence from_device		date_recorded
75	2498	206.4	0	0	7/30/2012 0:00
76	2498	190.2	0	0	8/27/2012 0:00
77	2498	180.2	0	0	9/18/2012 0:00
78	2498	201.2	0	0	8/7/2012 0:00
79	2498	197.8	0	0	8/14/2012 0:00
80	2498	195.6	0	0	8/21/2012 0:00
81	2499	201.4	0	0	7/11/2012 0:00
82	2498	190.2	0	0	9/27/2012 0:00
83	2497	274.2	0	0	9/20/2012 0:00
84	2499	205.2	0	0	7/17/2012 0:00
85	2498	186.6	0	0	9/5/2012 0:00
86	2498	184.4	0	0	9/11/2012 0:00
87	2499	200.4	0	0	7/25/2012 0:00
88	2498	180.2	0	0	9/18/2012 0:00
89	2497	266	0	0	1/1/2010 6:00
91	2499	196.8	0	0	8/1/2012 0:00
93	2498	174	0	0	10/2/2012 0:00
94	2499	193	0	0	8/15/2012 0:00
95	2499	190.7	0	0	8/27/2012 0:00
97	2499	190.2	0	0	9/4/2012 0:00
98	2499	185.2	0	0	9/18/2012 0:00
99	2499	181.2	0	0	9/27/2012 0:00
101	2499	185	0	0	10/9/2012 0:00
102	2433	190	0	0	10/17/2012 0:00
103	2464	202	0	0	5/8/2012 0:00
104	2464	207	0	0	5/1/2012 0:00
106	2528	169	0	0	10/20/2012 0:00
107	2532	240	0	0	10/22/2012 0:00

Figure 3.3. Initial Profile by Sanford dataset that required a serious restructuring.



0	'101'	'2012-12-27 00:00:00'
0	'166'	'2013-01-16 10:45:35'
2418	'175'	'2013-07-01 23:55:00'
2418	'169'	'2013-07-16 23:55:00'
2418	'164'	'2013-07-20 23:55:00'
2418	'159'	'2013-08-02 23:55:00'
2418	'155'	'2013-09-17 23:55:00'
2418	'155'	'2013-09-30 23:55:00'
2418	'153'	'2013-11-17 23:55:00'
2418	'144'	'2014-12-20 11:04:53'
2418	'138'	'2015-01-05 18:40:40'
2418	'150'	'2015-01-09 17:10:47'
2418	'155'	'2015-01-10 08:27:28'
2418	'155'	'2015-01-10 08:30:01'
2418	'300'	'2015-01-13 18:09:40'
2418	'179'	'2015-01-14 14:26:43'
2418	'181.88'	'2015-02-26 16:37:17'
2418	'181.88'	'2015-02-26 16:38:09'
2418	'204.59'	'2015-02-27 18:35:23'
2418	'202.82'	'2015-03-02 10:40:09'
2418	'201.28'	'2015-03-03 07:26:54'

Figure 3.4. Arranged and restructured Sanford dataset.

Figure 3.3, only shows the raw form of arranged dataset with only weights as possible state variable. However, there is need of additional state variables and further the BMI for each patient so that the relationship between these variables and the response can be implemented using RL algorithm and *Q*-learning for estimating optimal DTR. The whole training dataset that meets all above requirements along with the results from employing the algorithm will be described in the result and discussion chapter, Chapter 4.

3.3.2 R environment and coding

The Q-learning algorithm is implemented in R environment for estimating optimal DTR using the package iqLearn [75], which can be used with dataset from two-stage SMART trial design with binary treatments at each treatment stages. The dataset used was obtained from Sanford Health that contained the weight of patients measured at different time periods and the study was to observe the effect of meal replacement plans on adolescent obesity. This dataset consists of four covariates information at the start of first stage namely, gender, race, parent_BMI and baseline_BMI. Similarly, at second-stage or after the first treatment is administered, co-variate "month4_BMI" is collected. The treatment variables are denoted by "A_1" and "A_2" for first and second treatment stages, respectively. The primary outcome "month12_BMI" is observed at the end of stage two.

In *Q*-learning the function "qLearn*Q*1" recommends the estimated optimal treatment for first-stage with history set h_1. Similarly, function "qLearn*Q*2" recommends the optimal treatment for second-stage having history, h_2. The residual plots can be accessed for regression using "plot.qLearnS1" and "plot.qLearnS2" for first stage and second stage regression, respectively. The outcome of these residual plots will be discussed in chapter 4. Further, the plug-in value of any treatment decision rule can be estimated using the function "value". This function gives the estimated values of all possible treatment decisions rules embedded in the SMART design. Thus, the decision rule yielding maximum plug-in value is chosen as the optimal decision rule. Similarly, The adequateness of regression analysis or closeness of regression line fit can be observed and analyzed by using "summary" command. This will give us all the values of regression coefficients or parameters involved in the regression equation along with the R-squared value.

Therefore, for *Q*-function approximation and estimation of optimal decision rule R-software environment was favorable as it enables one to perform both graphical and statistical analysis of fit adequacy of a regression model and verify the underlying assumptions [73].

3.4 Optimization Assumptions and Residual Analysis

The multiple regression analysis is a statistical tool for analyzing and modeling the relationship between dependent and independent variables. The simple model for linear regression can be considered as $y = \beta_0 + \beta_1 x + \varepsilon$, where x is the independent variable and y the dependent variable. The independent variable is considered as predictor or regressor variable whereas dependent variable is the response variable. The term ε in the model gives the difference between the observed value and the predicted value by the model and is known as error. To obtain close fit of regression line the error term ε should be minimized.

Now, there are some important assumptions that are needed to be established before employing simple linear regression in the mathematical model. These assumptions help to define the criteria for verifying the results and also to underpin the notion that errors are independent random variables and requirement of hypothesis testing and interval estimation. The main assumptions in the study of simple linear regression analysis are listed below [74]:

- The relationship between the response variable y and the corresponding regressor variables should be approximately linear.
- The error term ε has zero mean.
- The error term ε has approximately constant variance σ^2.
- The errors are independent.
- The errors follow normal distribution.

For examining the adequacy of the model, the validity of above assumptions should be met and if these assumptions are violated the linear model can be infected by various model inadequacies resulting serious consequences in model fit. The violation of these assumptions can lead us to an unstable model where providing different sample leads to varying model or results with conflicting conclusions. One of the diagnostic method for examining the violations of these regression assumptions if to study the residuals of the model.

Therefore, residual analysis is not only a prominent way for checking the violation of linear regression assumption and adequacy of model fit but also a standard approach that should be followed while using regression based method, such as Q-learning, for function approximation to estimate DTR [75, 76]. So, first to define the residuals in regression analysis, consider first the following expression:

$$e_i = y_i - \hat{y}_i, \quad i = 1,2, \dots, n \tag{3.18}$$

Where y_i is the observed or real value of the dependent variable that is obtained from the training dataset, and $\hat{y}_i$ is the corresponding fitted value or the predicted value by the model. Then, as equation (3.18) suggests the residual can be defined as the deviation between the value of the response variable that is obtained from the training data and that obtained from the fitted value in the regression model. Thus, analysis of residuals can help in examining above assumptions as residuals define the error between the realized value from the model and the observed value in the data. Several model inadequacies can be detected by plotting residuals and observing the violation of assumptions which leads to an effective investigation of if the regression model fits the training data satisfactorily and if the assumptions of linear regression analysis are met.

To properly understand the process of residual analysis the basic properties of residuals should be understood. The important property of residual is that their mean is zero, and the approximate average variance is estimated as:

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res} \tag{3.19}$$

Where $n - p$ gives the degree of freedom associated with the n residuals and p is the number of parameters. The residuals are independent of each other and the residual values can be scaled. Standardized residuals are one of the process for scaling the residuals which is useful for finding the observations that are ouliers or extreme values in which the observations are separated from the core part of data in some way.

The average variance of residuals in data is approximated by MS_{Res}, which is given by equation (3.19), using the value of MS_{Res} from it the values of residuals using standardized residuals, can be scaled as:

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}\ , \quad i = 1,2,\dots,n \tag{3.20}$$

The values of standardized residuals have approximately unit variance and contain a mean of zero. The data point with large standardized residual say, $d_i>3$ denotes a potential outlier.

Model checking using Residual Diagnostic plots was introduced by Henderson et al. in 2010 [77], which was used for checking model misspecifications for estimating optimal DTR. As mentioned earlier, graphical analysis of residuals is very effective way to analyze the fit adequacy of a regression model and check the underlying assumptions. The residual diagnostic plots at each stage of regression presents the plots between residuals and fitted values, normal Q-Q plot, scale location plot and residual vs leverage plot. The plot between residual and fitted values shows if residuals have non-linear patterns and normal Q-Q plot shows if residuals are normally distributed. Similarly, scale-location plot checks the assumption of constant variance of residuals and residuals vs leverage checks to find out influential cases if any. These all plots are meant to check if above listed assumptions for linear regression models are violated or not to access a good fit. In the model, there are two stages and two subsequent regression analysis with residuals defined as difference between potential final outcome and outcome form the estimated Q-functions.

3.5 Sampling for Model Validation

The model validation is one of the most important step in a mathematical model building process which includes the process of measuring the extent of clinical benefit while applying the treatment rule for future patients. Generally, there are two ways to validate a model which are, external validation process and internal validation process. The external validation process employs the training data for model building whereas uses test or validation data to validate the model. On the other hand, the internal validation uses the same single dataset for both model building and validating agendas. Bootstrapping is one of the internal model validation process where samples are generated from population dataset in which the samples are drawn with replacement. Also, the sample size of both dataset is same and the validation process begins by testing the model on these bootstrap samples.

The bootstrap as discussed earlier involves random sampling with, replacement of data points from original dataset which are then later used for establishing statistical inferences. Along with it the bootstrapping method can be used to approximate the confidence intervals (CI) for estimated regression coefficients in a regression analysis. For example, the 95% CI of a sample mean can be obtained by using following steps:

- Let’s consider n observations with sample data points $(Y_1, Y_2, \dots, Y_n)$, where $\bar{Y}$ gives the sample mean of this sample dataset.

- If SD is the standard deviation of sample then SE, the standard error of sample mean is then:

$$SE = \frac{SD}{\sqrt{n}} \tag{3.21}$$

  The value of above SE gives the closeness of sample mean to the unknown population mean.

- Now, the 95% CI can be obtained by using the expression $(\bar{Y} - 1.96 * SE, \bar{Y} + 1.96 * SE)$.

- As the sample size increases and as the sampling distribution of sampling mean is closer to normality the CI moves closer towards the validity.

One of the very popular bootstrapping methods for computationally constructing CIs is the double bootstrap method which was explained by Davison and Hinkley in 1997 [75] and further implemented by Nankervis in 2005 [78]. In context of estimating optimal DTR using Q-learning, Chakraborty et al. in 2010 [79] used double bootstrapping method for estimating the CIs of the regression coefficients in multiple regression model of Q-functions.

Now, in double bootstrap method first an estimator of a parameter and its bootstrapped counterpart are defined. So, let $\hat{\theta}$ be the estimator of parameter θ and $\hat{\theta}^*$ be the bootstrap version of that estimator. Then as it is known from above that the $100(1-\alpha)\%$ percentile bootstrap CI is given by $\left(\hat{\theta}^*{}_{\left(\frac{\alpha}{2}\right)}, \hat{\theta}^*{}_{\left(1-\frac{\alpha}{2}\right)}\right)$, where $\hat{\theta}^*{}_{\gamma}$ is the $100\gamma^{th}$ percentile of the bootstrap distribution. Then the double bootstrap CI was calculated as follows:

- A first set of bootstrap samples say B_1 from original dataset was constructed. For this sample, the bootstrap version of estimator $\hat{\theta}^{*b}$ was estimated, where $b=1,\ldots, B_1$.
- Depending on first set of bootstrap samples, i.e. B_1 the second set of bootstrap samples, B_2 was constructed and the double bootstrap version of estimator, $\hat{\theta}^{**bm}$ was calculated, where, $b= 1,\ldots, B_1$ and $m=1,\ldots,B_2$.
- The value of $u^{*b} = \frac{1}{B_2}\sum_{m=1}^{B_2} \mathrm{I}[\hat{\theta}^{**bm} \leq \hat{\theta}]$ was estimated, where $\hat{\theta}$ is the estimator obtained from original data.
- Now, lastly the double bootstrap CI was obtained by calculating the interval $\left(\hat{\theta}^{*}{}_{\hat{q}\left(\frac{\alpha}{2}\right)}, \hat{\theta}^{*}{}_{\hat{q}\left(1-\frac{\alpha}{2}\right)}\right)$, where $\hat{q}(\gamma) = u^{*}_{(\gamma)}$ or the $100\gamma - th$ percentile of distribution u^{*b}, $b = 1,\ldots, B_1$.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results from Sanford Health Data

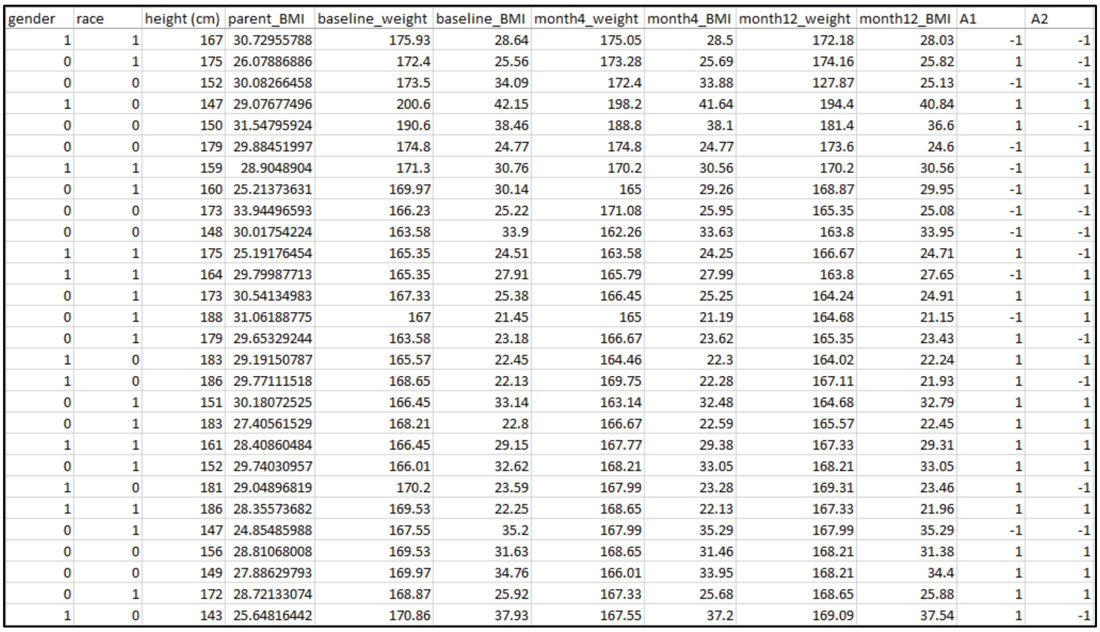
4.1.1 Data Restructuring

The data obtained from Sanford Health needed a critical restructuring to make it conducive for implementing SMART design and further *Q*-learning for optimization process. As mentioned in previous chapter the received dataset only consisted of user_id or patient's id along with their respective weights and date the weight was measured, however, the data were not arranged and required the weights taken after 4 months and 12 months. As, all patients were not able to continue the treatment for whole year or in other words there were some dropouts before 12 months period, only 210 patients data were selected for analysis who continued their treatment until 1 year period or more.

Furthermore, covariates such as gender, race, height, parent's BMI and treatment decisions at each stage were randomized and annexed to the restructured dataset. According to the heights assigned to each patients the BMI after 4 months and after 12 months of treatment were calculated using the equation (3.17). Thus, the restructured data consists of 210 rows of patients and 9 columns of covariates and some head rows of the dataset can be observed below in figure 4.1.

The restructured dataset of figure 4.1, consists of required input covariates and treatment decisions for each stages of a two-stage SMART design. The data are restructured in a way where inputs to the first stage regression analysis are gender, race, parent_BMI and baseline_BMI and that to the second stage regression analysis are again gender, parent_BMI and month4_BMI along with the treatment decisions A_1 and

A_2 for first stage and second stage treatments respectively. After, restructuring of data the multiple regression was employed for approximating the functions for second stage and first stage of SMART design. The summary and discussion of the regression analysis is described next.



gender	race	height (cm)	parent_BMI	baseline_weight	baseline_BMI	month4_weight	month4_BMI	month12_weight	month12_BMI	A1	A2
1	1	167	30.72955788	175.93	28.64	175.05	28.5	172.18	28.03	-1	-1
0	1	175	26.07886886	172.4	25.56	173.28	25.69	174.16	25.82	1	-1
0	0	152	30.08266458	173.5	34.09	172.4	33.88	127.87	25.13	-1	-1
1	0	147	29.07677496	200.6	42.15	198.2	41.64	194.4	40.84	1	1
0	0	150	31.54795924	190.6	38.46	188.8	38.1	181.4	36.6	1	-1
0	0	179	29.88451997	174.8	24.77	174.8	24.77	173.6	24.6	-1	1
1	1	159	28.9048904	171.3	30.76	170.2	30.56	170.2	30.56	-1	1
0	1	160	25.21373631	169.97	30.14	165	29.26	168.87	29.95	-1	1
0	0	173	33.94496593	166.23	25.22	171.08	25.95	165.35	25.08	-1	-1
0	0	148	30.01754224	163.58	33.9	162.26	33.63	163.8	33.95	-1	-1
1	1	175	25.19176454	165.35	24.51	163.58	24.25	166.67	24.71	1	-1
1	1	164	29.79987713	165.35	27.91	165.79	27.99	163.8	27.65	-1	1
0	1	173	30.54134983	167.33	25.38	166.45	25.25	164.24	24.91	1	1
0	1	188	31.06188775	167	21.45	165	21.19	164.68	21.15	-1	1
0	1	179	29.65329244	163.58	23.18	166.67	23.62	165.35	23.43	1	-1
1	0	183	29.19150787	165.57	22.45	164.46	22.3	164.02	22.24	1	1
1	0	186	29.77111518	168.65	22.13	169.75	22.28	167.11	21.93	1	-1
0	1	151	30.18072525	166.45	33.14	163.14	32.48	164.68	32.79	1	1
0	1	183	27.40561529	168.21	22.8	166.67	22.59	165.57	22.45	1	1
1	1	161	28.40860484	166.45	29.15	167.77	29.38	167.33	29.31	1	1
0	1	152	29.74030957	166.01	32.62	168.21	33.05	168.21	33.05	1	1
1	0	181	29.04896819	170.2	23.59	167.99	23.28	169.31	23.46	1	-1
1	1	186	28.35573682	169.53	22.25	168.65	22.13	167.33	21.96	1	1
0	1	147	24.85485988	167.55	35.2	167.99	35.29	167.99	35.29	-1	-1
0	0	156	28.81068008	169.53	31.63	168.65	31.46	168.21	31.38	1	1
0	0	149	27.88629793	169.97	34.76	166.01	33.95	168.21	34.4	1	1
0	1	172	28.72133074	168.87	25.92	167.33	25.68	168.65	25.88	1	1
1	0	143	25.64816442	170.86	37.93	167.55	37.2	169.09	37.54	1	-1

Figure 4.1. The head rows of restructured dataset consisting of randomized covariates and required BMI information for implementing SMART design.

4.1.2 Initial Data Assessment

The initial assessment was performed on the restructured dataset where first response after 12 months was observed for female and male and race A and race B (races under comparison) using box-plots. Similarly, box-plots were again used to observe the effect of treatments (e.g. Augment and Switch) on response after 4 months and after 12 months.

In figure 4.2, for data following a normal distribution, the mean value of BMI after 12 months for female is lower than that for male. Similarly, race B has lower mean value for BMI after 12 months compared to race A. Additionally, comparing response after 4 months for treatment stage 1, there is not much difference in average

BMI between treatments augment and switch. However, for second stage the average BMI after 12 months decreases for treatment augment compared to switch.

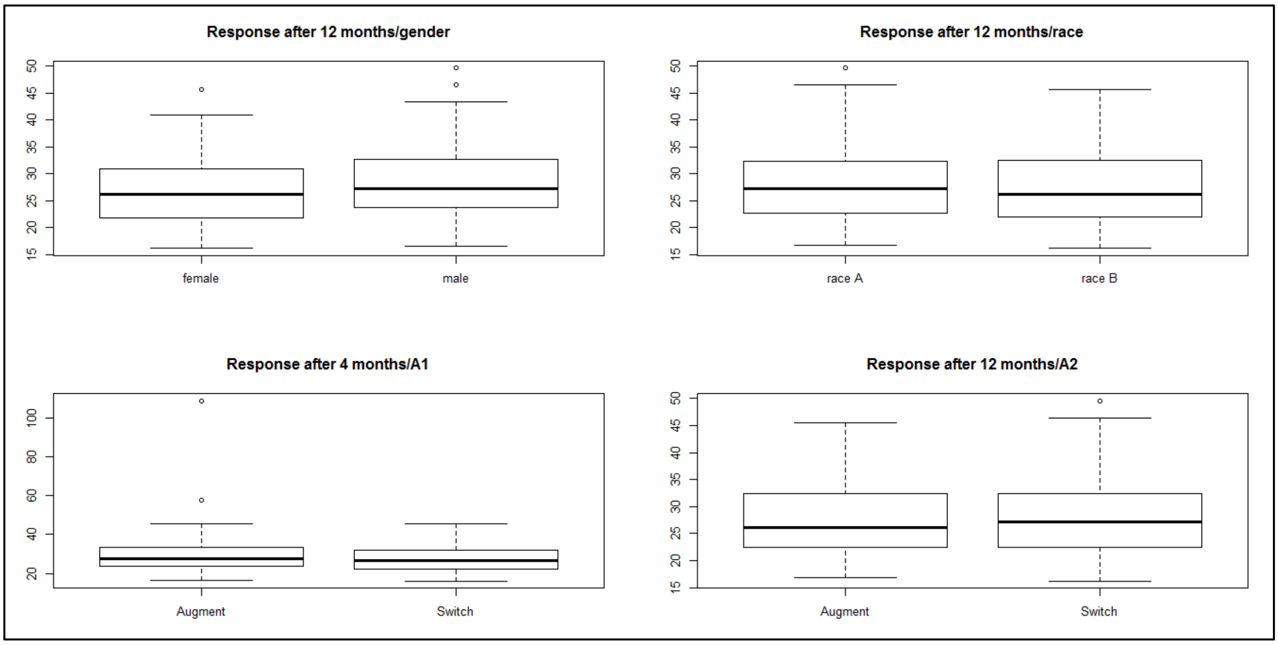


Figure 4.2. Box-plots of response after 4 months and 12 months according to gender, race and treatment decisions at each stage.

After average BMI for 4 months and 12 months response were observed according to gender, race and treatment stages using Box-plots, scatter-plots matrix was used to observe relationship between response and predictors. First the relation between second stage response variable, month12_BMI was compared with predictors parent_BMI, baseline_BMI and month4_BMI.

From figure 4.3, it can be clearly observed that the response month12_BMI has a strong linear relationship with the predictors baseline_BMI and month4_BMI. This also validates the case of linear model being employed for estimating *Q*-function for stage 2 regression analysis. Similarly, figure 4.4, below shows the relation between response i.e. month4_BMI and predictors which also has a strong linear relationship and linear model is defined for first stage regression analysis too.

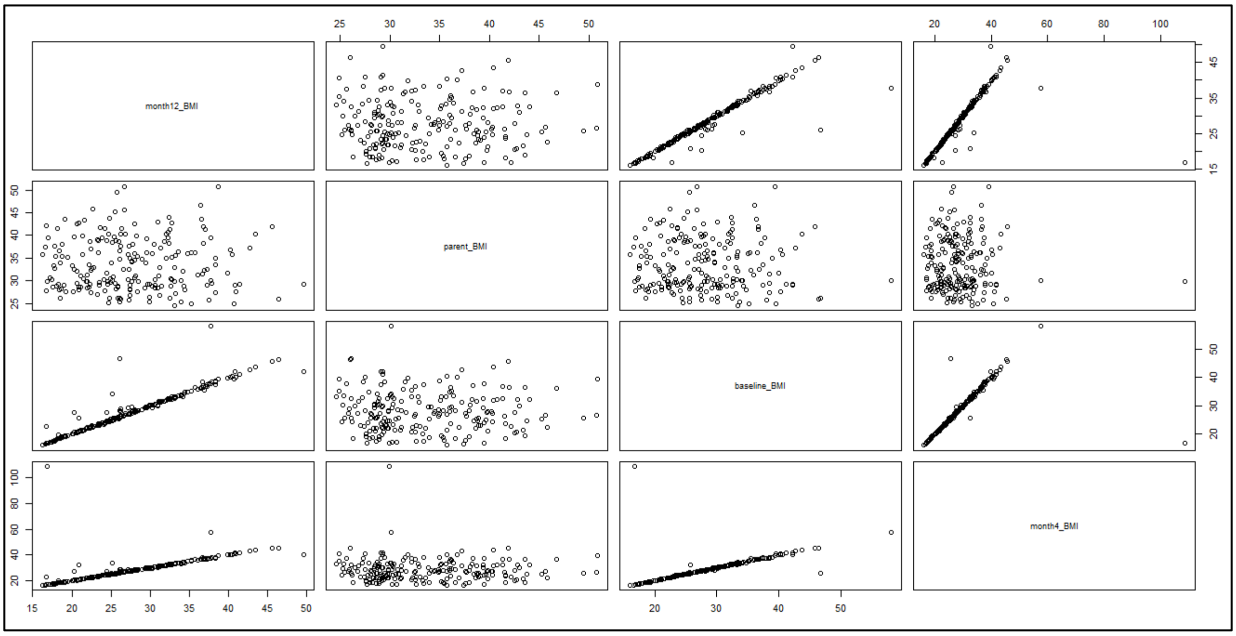


Figure 4.3. Scatter-matrix plots for second-stage regression predictor and response variables showing a linear relationship.

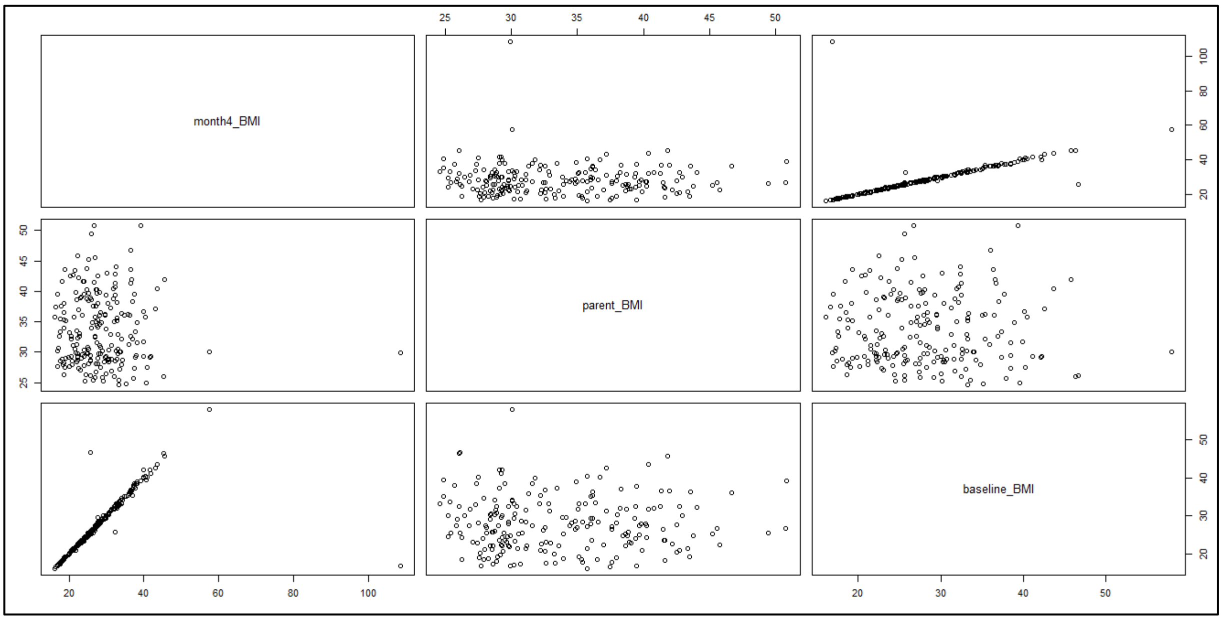


Figure 4.4. Scatter-matrix plots for first-stage regression predictor and response variables showing a linear relationship.

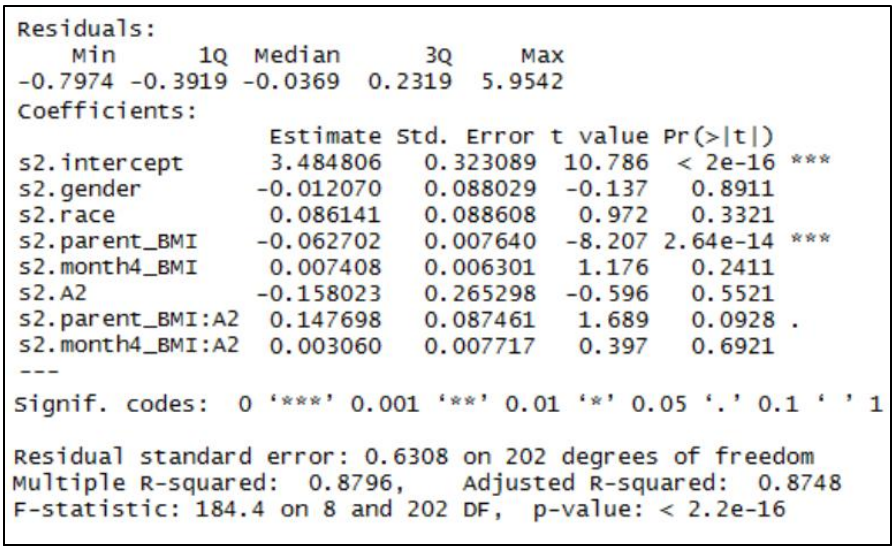
4.1.3 Regression Analysis

The multiple regression is implemented initially to second stage of SMART design as Q-learning is a backward induction method. In second stage, as mentioned earlier the input to the regression formula are gender, parent_BMI, month4_BMI and treatment decision (A_2). The stage two multiple regression model equation is represented below as:

$$y = \beta_0+\beta_1 gender+\beta_2\ parent\_BMI+\beta_3 month4\_BMI+A_2*(\beta_4\ parent\_BMI+\beta_5 month4\_BMI) \qquad (4.1)$$

In above equation $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients of the independent variables gender, parent_BMI, month4_BMI and interactions between parent_BMI and month4_BMI respectively. Also, y is defined as the negative percent change in BMI at month 12 from baseline BMI i.e., $y = -100 * (\frac{month12_{BMI} - baseline_{BMI}}{baseline_{BMI}})$.Now, the goal is to estimate these coefficients so that the function given by the equation (4.1) can be approximated. After applying the multiple regression, following summary for second stage was obtained:

Table 4.1 Summary table for stage 2 regression analysis



```
Residuals:
    Min      1Q  Median      3Q     Max
-0.7974 -0.3919 -0.0369  0.2319  5.9542
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
s2.intercept      3.484806   0.323089  10.786  < 2e-16 ***
s2.gender        -0.012070   0.088029  -0.137   0.8911
s2.race           0.086141   0.088608   0.972   0.3321
s2.parent_BMI    -0.062702   0.007640  -8.207 2.64e-14 ***
s2.month4_BMI     0.007408   0.006301   1.176   0.2411
s2.A2            -0.158023   0.265298  -0.596   0.5521
s2.parent_BMI:A2  0.147698   0.087461   1.689   0.0928 .
s2.month4_BMI:A2  0.003060   0.007717   0.397   0.6921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6308 on 202 degrees of freedom
Multiple R-squared:  0.8796,    Adjusted R-squared:  0.8748
F-statistic: 184.4 on 8 and 202 DF,  p-value: < 2.2e-16
```

The information about estimated values of regression coefficients, residuals, standard error of estimated coefficients and other such as R-squared error can be obtained from above Table 4.1.

Furthermore, in simple linear regression model the coefficients are the constants that represent the intercept and the slope of the linear model. The first column *"Estimate"* of the table coefficients gives the estimates of all the expected values of the coefficients as described by equation (4.1). Similarly, the second column *"Std. Error"* gives the error in estimated coefficients. The lower values of this standard error suggest good quality of regression line fit. The *t value* on the third column gives the measure of how many standard deviations is the estimated coefficients away from 0. Further the distance closer to decision rule of rejecting null hypothesis, enabling the declaration of strong relationship between predictor variables and the response variable. The asterisk (*) alongside the values represent the level of significance three being the most significant estimate. Finally, the column *"Pr (>t)"* describes the probability of observing the value equal or greater than t value. The smaller p values in this column suggests that the relationship that is observed between the predictor and response variable is not by chance or fluke. Thus, these small p-values for estimates slope or intercepts suggests that the null hypothesis can be rejected and concludes a good relationship between treatment decisions and the patient co-variates.

Next, is the residual standard error that measures the quality of regression line fit and is defined as the average distance or deviation of the response (treatment outcome) variable from the linear regression line. As, the regression line cannot be perfect, and every model is presumed to have some error term *E* and this error term should be as minimum as possible so that the prediction is accurate and consistent. From Table 4.1, after observing the error value of 0.63, it can be deduced that the predictor variable can deviate from the regression line approximately by this error term on average during the prediction of response variable. Also, the degree of freedom defines the number of data points that was taken into consideration while estimating the regression parameters.

Similarly, the Multiple R-squared and Adjusted R-squared statistics also provides the measure of closeness of fit between the model and its fitting to the actual data. So, R^2 term defines the measure of linear relationship between the independent and response variable whose values lie between 0 and 1. If the value of this term is closer to 0 than the regression line will poorly explain the variance in response variables whereas values closer to 1 will provide good regression line fit. In case suggested by above Table 4.1, approximately 88% of variance observed in response variable can be well explained by the predictor variables. However, in multiple regression setting the value of R^2 increases as the number of predictor variable increases or as more variable are introduces to the model. So, to minimize this effect the adjusted R^2 is preferred more as it considers and adjusts the effect of number of variables considered for regression analysis.

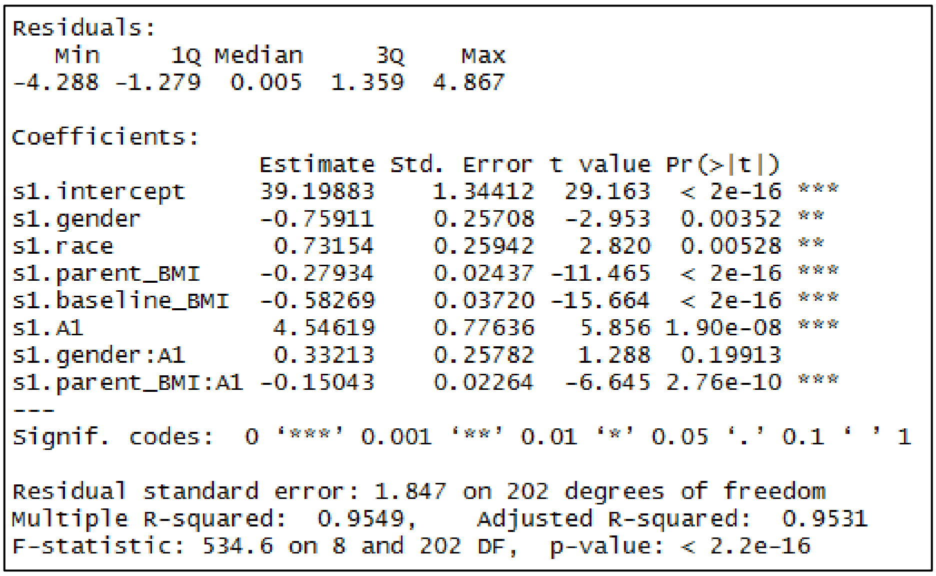
Lastly, the F-statistic, as shown in table 4.1, can also be a good estimator of relationship between the dependent and independent variables in a regression model. As the value of this statistic moves further from 1 or is greater than 1 the better is the model or it well explains the relationship. However, its value is also dependent upon the number of variables considered on the model. Generally, if the number of data points are small the value of F-statistic little bigger than 1 is sufficient in rejecting the null hypothesis and accepting the notion that there is a good relationship between predictor or response variable or good fit of regression model is obtained through regression analysis. In case of second stage regression analysis for Q – function approximation 184.4, F-statistic value was obtained that is larger than 1 in relation to size of data employed. So, from the results of second stage multiple regression model it can be posited that there is good linear fit and the model can be implemented for predicting treatment rule for future patients.

Now, Table 4.2, below shows the summary of regression analysis for function approximation in stage 1 of SMART design. As Q-learning process follows backward induction, the predicted future outcomes $\tilde{Y}$ assuming an optimal decision was prescribed in the second-stage, was calculated as follows:

$$\tilde{Y} = max_{a_2 \in \{-1,1\}} Q_2\left(H_2, a_2; \hat{\beta}_2\right) = H_{2,0}^T \hat{\beta}_{2,0} + |H_{2,1}^T \hat{\beta}_{2,1}| \qquad (4.2)$$

After predicting the future outcomes, the value of coefficients and error minimization process was undertaken using least-square regression method and following summary table was obtained:

Table 4.2 Summary table for stage 1 regression analysis



```
Residuals:
   Min     1Q Median     3Q    Max
-4.288 -1.279  0.005  1.359  4.867

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
s1.intercept      39.19883    1.34412  29.163  < 2e-16 ***
s1.gender         -0.75911    0.25708  -2.953  0.00352 **
s1.race            0.73154    0.25942   2.820  0.00528 **
s1.parent_BMI     -0.27934    0.02437 -11.465  < 2e-16 ***
s1.baseline_BMI   -0.58269    0.03720 -15.664  < 2e-16 ***
s1.A1              4.54619    0.77636   5.856 1.90e-08 ***
s1.gender:A1       0.33213    0.25782   1.288  0.19913
s1.parent_BMI:A1  -0.15043    0.02264  -6.645 2.76e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.847 on 202 degrees of freedom
Multiple R-squared:  0.9549,    Adjusted R-squared:  0.9531
F-statistic: 534.6 on 8 and 202 DF,  p-value: < 2.2e-16
```

As discussed for table (4.1), summary for second stage regression analysis, table (4.2) shows the summary table for first stage regression. Here also, first the coefficients of independent variables were estimated for the equation (4.2) as shown below:

$$y = \beta_0 + \beta_1 gender + \beta_2 race + \beta_3\, parent\_BMI + \beta_4 baseline\_BMI + A_1 * (\beta_5\, gender + \beta_6\, parent\_BMI) \qquad (4.3)$$

Similar to summary table (4.1), in table (4.2) the estimated coefficients of equation (4.3) are listed. Using these estimates, the function can be approximated and used to compare the *Q*-values for different set of covariates and treatment decisions and then finally the optimal treatment decision can be estimated as it is the set of inputs that results in maximum *Q*-value.

From table (4.2), the values of coefficient estimates, error while estimating these coefficients and various other information about adequacy of regression line fit such as Residual standard error and adjusted R-squared values can be obtained. The value of 1.847 was obtained for the residual standard error value during first stage regression, this means that the values of independent variables can deviate from regression line by value of 1.847. Similarly, the multiple R-squared error was 0.9549 and that for adjusted one was 0.9531, referring to the statement that approximately, 95% of variable that is observed in response variable can be well explained by the predictor variables. Also, the F-statistic value of 534.6 was observed which is much greater than 1 and p-value obtained was significantly lower than 1. Therefore, from these results it can be stated that good linear fit was observed during regression analysis for first stage treatment decision in SMART design and this estimated function can be implement while prescribing first stage treatment for patients with totally different sets of input covariates.

4.1.4 Regression Diagnostic Plots

As mentioned in earlier chapters, the graphical analysis of residuals is very effective way to analyze the adequacy of fit and to check the underlying assumptions of any regression model. As residuals are the difference between the observed value and predicted value from the model, there are mainly four plots that comes into consideration while describing residual diagnostics and they are plot between residuals and fitted values, the normal Q-Q plot, the scale-location plot and the residual versus leverage plot. Now, each plot obtained for first and second stage multiple regression analysis is described below.

First, the plot between residuals and fitted values shows if the residuals have nonrandom patterns or not. This plot is also useful for verifying the assumptions made for linearity and homoscedasticity (constant variance assumption). In this scenario, the model said to not meet the linear model assumption, if very large residuals with big positive or negative value were observed. So, these residuals should not be very far from 0 so that the assumption of linearity is met, similarly, it was also needed to make sure that there is no pattern observed or the residuals are equally spread around line *y=0* for accessing the assumption of homoscedasticity. The residual versus fitted plot for second stage regression analysis is shown below:

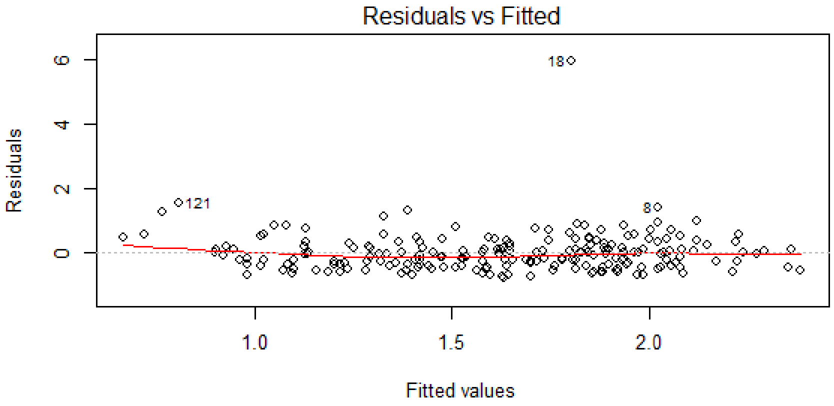


Figure 4.5 Residual versus fitted plot for second stage multiple regression

Figure 4.5, shows the residuals versus fitted plot for multiple regression while prescribing second stage treatment decision in SMART design. From the plot, it can be observed that the residuals are homogenously spread above and below 0 and no pattern can be observed for residual vs fitted line. Thus, both the assumptions of linearity and homoscedasticity were met. Again, for first stage regression analysis following plot was obtained:

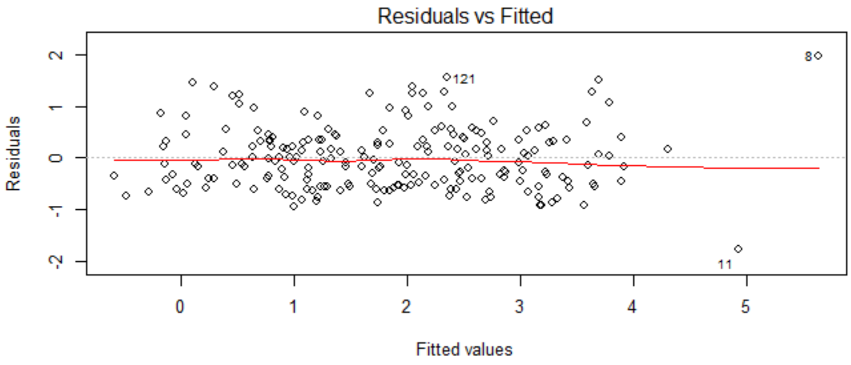


Figure 4.6 Residual versus fitted plot for first stage multiple regression

Figure 4.6, shows the plot between residual and fitted values for multiple regression in first stage of treatment decision using SMART design. From this plot, it can be observed that again the residuals are uniformly distributed along 0 and there is no pattern if a line is drawn for fitting the data points. Thus, in case of first stage regression also the assumptions of linearity and homoscedasticity were met.

Next, the normal Q-Q plot for both first and second stage regression were investigated for evaluating the normality assumption of linear regression which basically, compares the standardized residuals to theoretical quantiles or normal observations. If the observations follows or lies along the 45-degree line then it can deduced that the normality assumptions hold. It can also be observed from following figure 4.7, which shows the Q-Q normal plot for second stage regression analysis.

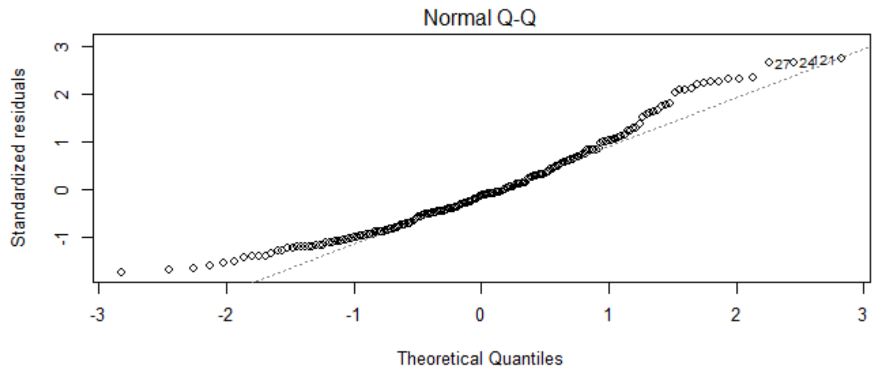


Figure 4.7 Normal Q-Q plot for second stage multiple regression analysis

Figure 4.7, shows that most of the observations follows or lies on the 45-degree dotted line hence, the normality assumption is validated for linear regression. Also, it can be asserted that for a linear regression analysis the model fitting is good and assumption of normality is observed. Again, the normal Q-Q plot for first stage regression analysis is shown below:

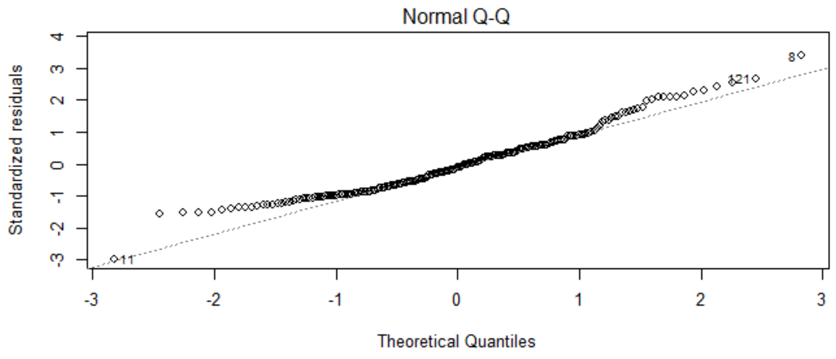


Figure 4.8 Normal Q-Q plot for first stage multiple regression analysis

Figure 4.8, shows the normal Q-Q plot for multiple regression analysis while prescribing treatments for first stage in SMART design. Here, also much of the observations are in the 45-degree line and the analysis can be asserted to the point that normality assumption of the linear regression holds in this case too and the regression fit is good.

Further, to validate the assumption of homoscedasticity which states that, the variance of residuals is constant for any different predictor values, in linear regression model fitting procedure scale-location plot can be used to check if there is some pattern or if the variance is constant for values of independent variables. The scale-location plot is plotted between the square rooted values of standardized residuals and predicted values from the model. Following figure, figure 4.6 shows the scale-location plot for predictor values of second treatment decision stage in SMART design which is used to estimate optimal DTR.

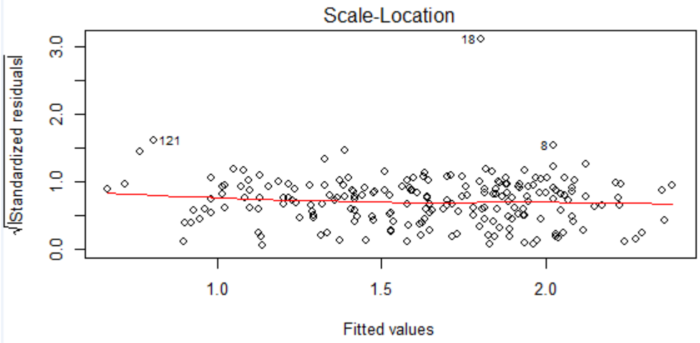


Figure 4.9 Scale-Location plot for second stage regression analysis

In Figure 4.9, the plot examines if the residuals are equally spread along the different ranges of predictor variables. Therefore, as it can be observed in figure 4.6, a horizontal line with equally spread random points along it can be conjectured that the variance of residuals is constant and hence, the assumption of homoscedasticity is validated. Also, the scale-location plot for regression analysis in first stage of treatment decision is shown in figure 4.10 below. In the figure 4.10, it can be observed that the variance is again constant along different values of predictor variables and the assumption of homoscedasticity is valid in this case too. Thus, both of the plots strengthen the conjecture of good regression model fit and underpins the significance of approximated functions, while implementing these functions as base of prescribing treatments in two consecutive stages suggested by the SMART design in our model.

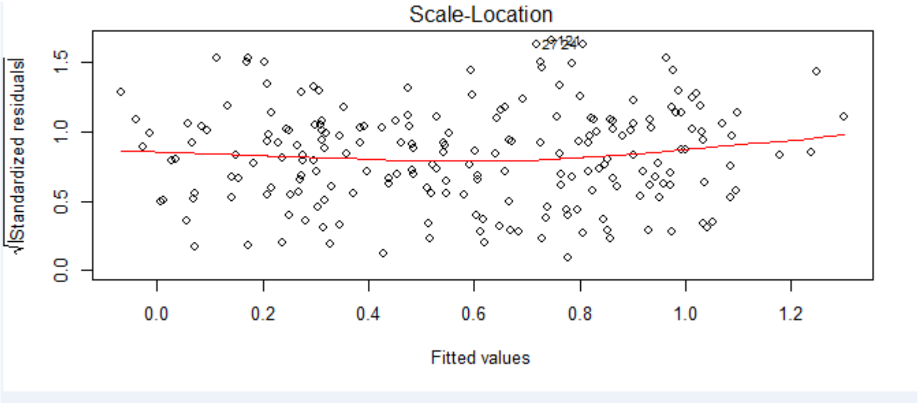


Figure 4.10 Scale-Location plot for first stage regression analysis

Lastly, for the next graphical analysis which considers the “Cook’s distance” to measure the influence of each observation while predicting the values of regression coefficients, the following terms should be defined:

- The observations with large residuals compared to other observations in the model are known as outliers. For example, if the observed value of one of the observation is very much different than that of the predicted value obtained using regression model, then this observation can be categorized as an outlier.
- Next, the leverage points of each observation are needed to be considered while analyzing the goodness of model fit. A leverage point can be defined as the distance of an observation from its mean value.
- Also, the observation with significant leverage can change the slope of regression line resulting this observation to be very influential. Hence, these

influential points have substantial influence on goodness of fit in any regression model.

Now, the graphical analysis of plots between leverage and standardized residuals also defines a statistic measure term called “Cook’s distance”. The Cook’s distance measures the influence of an observation on the overall regression model for example change in regression coefficients. Thus, this statistical tool analyzes the amount or extent of changes that occurs in model if an observation is omitted. Generally, the observations with high influence on the model has cook’s distance close to one or larger compared to other observations. Figure 4.11 below shows the residual versus leverage plot for second stage regression analysis in SMART design. From the plot, it can be observed that the model is not affected by influential points majority of observations lie within 1 cook’s distance. Therefore, the model for regression analysis in second stage treatment decision is not affected by the influential points.

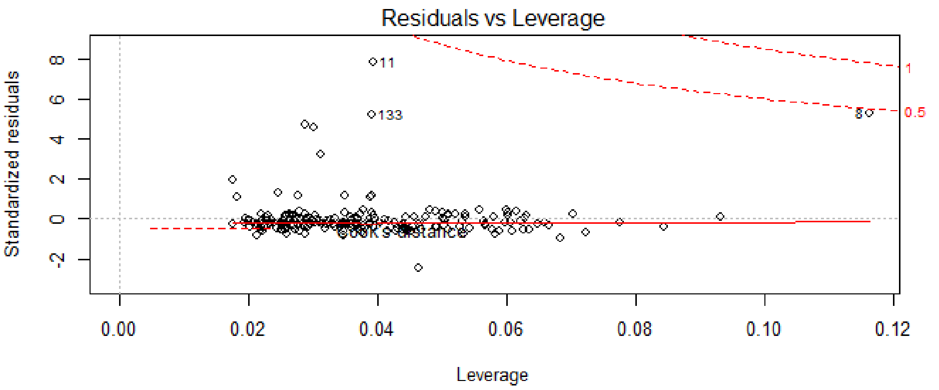


Figure 4.11 Residuals vs Leverage plot for second stage multiple regression analysis

Similarly, the residual versus leverage plot for first stage regression is also shown in figure 4.12, below. Also, from figure 4.12, it can be observed that no observations have cook’s distance greater or equal to 1 and there are no major influential observations. However, the regression line is stretched due to the observations with high leverage, but the influential points are minimal so that they will not affect the overall regression model.

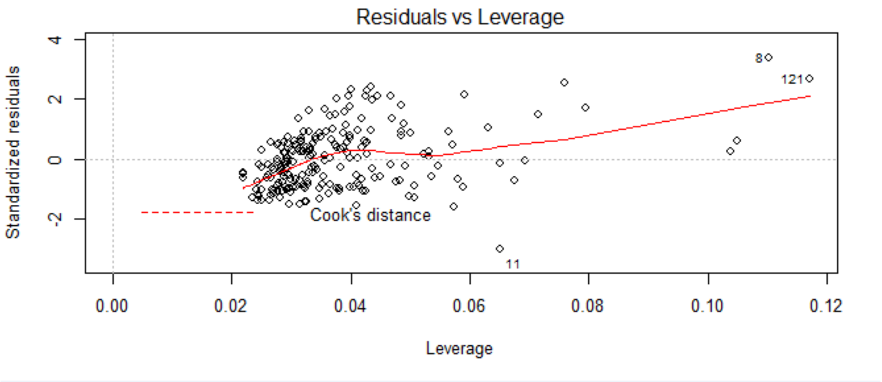


Figure 4.12 Residuals vs Leverage plot for first stage multiple regression analysis

Therefore, from above residual diagnostics, it can be deduced that the coefficients which were estimated for regression function are correctly specified as the assumptions of linear regression analysis are verified using residual diagnostic plots.

4.1.5 Optimal Treatment Decision Rule

Now, after the *Q*-functions at both treatment stages are approximated using multiple regression method, the other set of input covariates from new patient can result in output which is treatment resulting maximum *Q*-value at respective stages. For example, in stage 1 the problem is to prescribe an optimal treatment for patient with new set of input covariates, thus, to do that first the values of *Q*-functions are approximated for each available treatment options and the one resulting maximum *Q*-value is defined as the optimal treatment decision.

Table 4.3 Optimization of adaptive decision rule according to maximum value of *Q*-functions

Combination of history set	Treatment = 1	Treatment = -1	Optimal Treatment
Q_2 value, c (1,30,24)	0.3803014	-0.01900676	1

Q_1 value, c (1,0,34,30)	1.023668	1.23254	-1
Q_2 value, c (1,30,45)	-0.9962029	-0.8904261	-1
Q_1 value, c (1,0,25,30)	-2.055502	2.866527	-1

Table 4.3, above shows different values of Q-functions at stages Q_2 and Q_1, where the inputs are distinct set of biomarkers belonging to a fresh patient or patient having first clinic visit. Thus, the set of biomarkers or history set for these patients are given as input to the model and defined as the argument of set c, for example, c (1,30,24) in first row gives the information about gender, BMI after four months and BMI after 12 months, respectively of a new patient in program. Using this information, a history set is constructed and is used for estimating the values of Q-functions by keying the history set values as inputs or values of predictor variables in the approximated Q-functions of the model. As the values of regression coefficients are already approximated through regression analysis, the Q-functions values for newly constructed history set can be easily estimated using these biomarkers of a new patient.

Furthermore, the Q-values at each treatment stages are noted for assigned treatment decisions (1 or -1) in that particular stage and the treatment resulting in maximum Q-value is selected as optimal treatment decision. It can be observed in table 4.3 that in stage 2 for c (1,30,24), treatment coded 1 is chosen as optimal treatment decision because it has greater estimated value of Q_2, which is 0.3803014 in comparison to Q_2 value from treatment decision -1. Similarly, for stage 1 for c (1,0,34,30), treatment -1 results in maximum Q_1 value of 1.23254, hence, treatment -1 is chosen as optimal treatment for this stage. So, at last the optimal treatment decision rule can be assigned as (-1,1) for the patient with covariates c (1,0,34,30) and c (1,30,34), in first and second stages respectively. Thus, as it can be assumed that different patients have different history set so, the model will follow different optimal path resulting the model to be a personalized treatment regimen model.

4.1.6 Estimating Regime Values

The next comparison that can be done is between the estimated optimal regime and the standard care decision rule or a constant regime that recommends same treatment regime for all patients. Thus, the way to do this comparison is by estimating the value function that is defined as:

$$\hat{V}^{\pi} = \frac{\sum_{i=1}^{n} Y_i \mathbb{I}\{A_{1i} = \pi_1(h_{1i})\}\mathbb{I}\{A_{2i} = \pi_2(h_{2i})\}}{\sum_{i=1}^{n} \mathbb{I}\{A_{1i} = \pi_1(h_{1i})\}\mathbb{I}\{A_{2i} = \pi_2(h_{2i})\}} \tag{4.4}$$

In above equation (4.4), the value Y_i is the response for i^{th} patient, (A_{1i}, A_{2i}) are the randomized treatment decision and (h_{1i}, h_{2i}) are the histories that are observed before treatment. Thus, the value estimator defined by equation (4.4) is nothing but the weighted average of outcomes observed from patients in the trial that received treatment according to the decision rule π. This estimator is also known as the Horvitz-Thompson estimator [80].

In R-environment, the function *value*( ) within package iqLearn is used to estimate the regime values which returns the value estimated of all regimes in the design. Figure 4.10, below shows the bar graph for the estimated value of each possible regimens, namely A_1A_1, A_1A_2, A_2A_1 and A_2A_2.

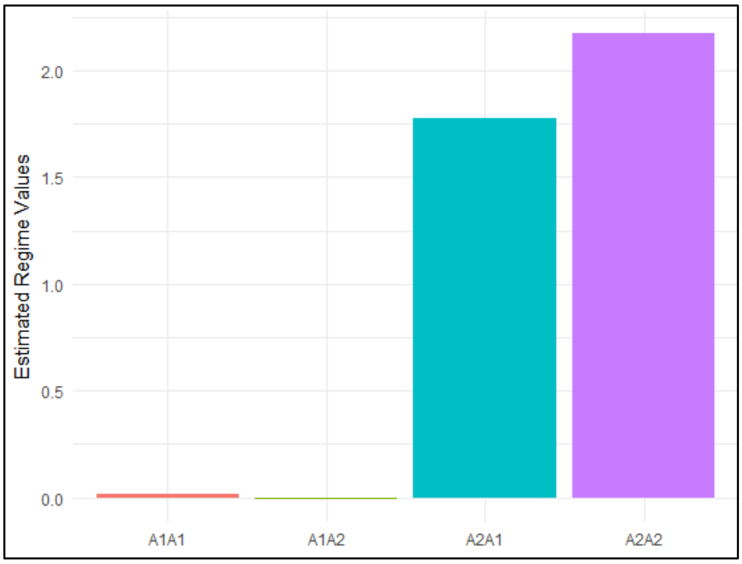


Figure 4.13 Estimated regime values for different decision rules

Thus, from above figure 4.13, it can be observed that the regime value for decision regime A_1A_1 is very small and that for A_1A_2 is negative, meaning that the patient's health is degrading. The decision rule with maximum regime value is A_2A_2, so, this decision rule is assigned as the optimal decision rule and will perform very well for patient's well- being if prescribed. Again, for the patient with different set of observed histories following estimated regime values were obtained:

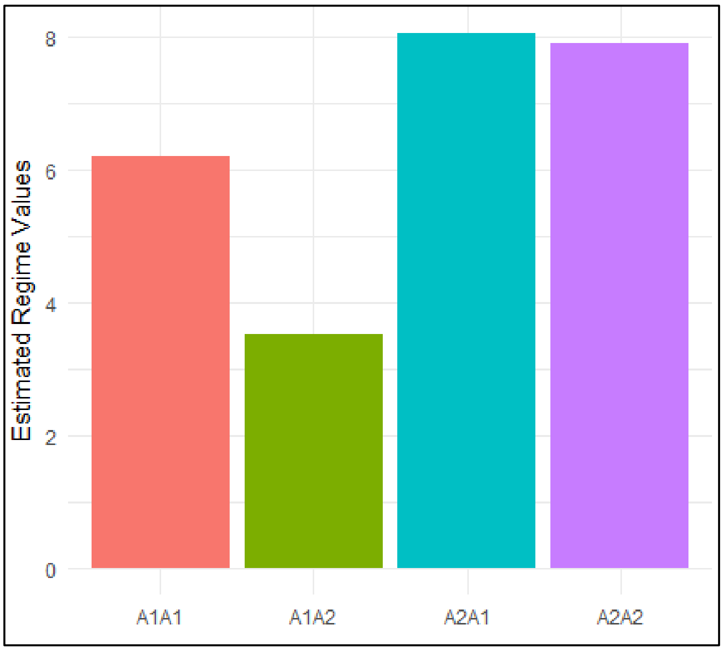


Figure 4.14 Estimated regimes values for patient with separate set of observed history

In figure 4.14, all the decision rule can result in the well-being of the patient as no decision rule has estimated regime value below zero. However, the regime with A_2 and A_1 as treatment decisions in stage 1 and stage 2 respectively has the maximum estimated regime values. Therefore, for this patient with new set of history treatment regime A_2A_1 can be considered as the optimal decision rule.

4.2 Discussions

This section of the chapter discusses the results that was obtained above and how these results can be interpreted for future implementations. So, first the results were obtained in Data restructuring section, which highlights the importance of acquiring the training data and remodeling it so that it could be used for training the model. The raw data that was obtained from Sanford Health needed serious restructuring so that it can be conducive to SMART design which is an essential part of the mathematical model. The restructured dataset as shown in figure 4.1, can be easily implemented for training the model as the input covariates and output response are clearly defined and available. Similarly, the gender, race, height and treatments which are randomized in the model are done so that the results explaining relationships between these predictors to response is obtained during regression analysis.

Next, results were about the multiple regression analysis for both first and second treatment stages in SMART design. The regression summary tables obtained are shown on table 4.1 and table 4.2, which gives the values of estimated regression coefficients, error during the estimation, the residual standard error and quantiles of the residual. The regression analysis was chosen as the method of inference and the values obtained from regression summary tables were analyzed for goodness of regression fit.

These tables not only provide the values of estimated regression coefficients but also gives the information about ability of regression analysis to account for total variation in the dependent variable or in other words the quality of fit. The Residual standard error for stage 2 and stage 1 regression were 0.63 and 1.847, respectively which can be considered minimal so, as residual standard error is the standard deviation of the residuals, it's minimal value results in good quality of regression line fit. Similarly, the Multiple R-squared value, also known as coefficient of determination is the proportion of variance in data that is explained by the model and it is proportional to number of predictor variables. The value of 0.8796 and 0.9549 were obtained for Multiple R-squared valued in second stage and first stage regression analysis respectively, indicating that the independent variable explains an estimated 88% variation in dependent variable of second stage and 95% variation in that of first stage regression. Therefore, the regression summary table indicated toward good regression line fit and the estimated values of coefficients can be further used in approximating the *Q*-functions.

After the restructured training data was implemented to obtain the estimated values of regression coefficients and analysis of goodness of fit, the graphical analysis of residual diagnostics was performed to analyze the fit adequacy of the regression model and also to check the underlying model assumptions. The assumptions that are needed to be verified for linear regression are assumptions of linearity, Homoscedasticity, Independence and Normality. Figure 4.5 and 4.6, shows the residual versus fitted values plots for second stage and first stage regression analysis respectively. It can be observed from these plots that the residuals which is the difference between observed final outcome and predicted outcome from the model, are not very large as there is not big positive or negative values. Also, to verify the assumption of linearity, the residuals from residuals versus fitted values plots are not too far away or in other words close to 0. Next, the normal Q-Q plots are shown in figure 4.7 and 4.8 for again second stage and first stage regression analysis respectively. Usually, the normal Q-Q plots are used to evaluate the normality assumption of linear regression by comparison of residuals to normal observations. As in both figures (4.7 & 4.8) the observations lie along the 45-degree dotted line, hence, it can be assumed that the normality assumptions hold in both cases. Similarly, the third plots given in figures 4.9 and 4.10 shows the scale-location plots which are used for checking the assumption of homoscedasticity which means there is constant variance in residuals. Thus, to verify the homoscedasticity assumption it was observed and made sure that there was no significant trend or pattern in the residuals and as in figures 4.9 and 4.10 the fitted line is approximately horizontal that describes no pattern for both cases, verifying the assumption of constant variance in residuals. The fourth and final plots in figures 4.11 and 4.12 are the residual versus leverage plots for second stage and first stage regression respectively. This plot is used to observe the Cook's distance which measures the influence of each observation on the regression coefficients. Thus, it can be observed in figure 4.11 that the fitted line is flat, and no influential points are affecting the model, also the cook's distance for each observation are below 1 and not significant, indicating lack of influential data points. However, figure 4.12 shows that the fitted line is somewhat stretched by the influential observations in the dataset, resulting in significant cook's distance of those data points which required further investigation. Although, some influential points are affecting the first stage regression analysis it does not affect the goodness of fit or fit adequacy of the model as all the model assumptions are verified.

Now, after fitting the model to training data and evaluating the performance by assessing model goodness-of-fit and prediction, the model should be applied to prescribe treatment rules for future patients. This application can be obtained either by estimating optimal treatment decision rules as in Table 4.3 or estimating the regime values, shown in figures 4.13 and 4.14. In Table 4.3, two new patients with different history set are selected and the values of *Q*-functions prescribing both treatment decisions are obtained for these two regression stages. The *Q*-values for both treatment decisions are recorded and the one resulting maximum *Q*-value is selected as the optimal treatment. In this scenario, the first patient should follow the treatment regime (-1, 1) and second should follow (-1, -1). Again, figures 4.10 and 4.11 shows the estimated values of all possible regimes in a bar-plot. Here, the regime with maximum regime value is selected as optimal decision rule. Thus, from figure 4.10 as regime A_2A_2 has highest estimated value, it should be selected as optimal decision rule, whereas in figure 4.11 which is for another patient, regime A_2A_1 has maximum estimated regime value so, it should be selected as optimal decision rule in this case. Therefore, various statistical analysis was implemented upon the restructured training dataset and the model performance along with model goodness-of-fit were also predicted, giving overall good model fit. Hence, the constructed mathematical model can be used for acquiring prognostic and predictive covariates which can be implemented for selecting optimal treatment decision rule.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Summary

Personalized medicine emphasizes on the fact that there is a great variability among individuals which plays a vital role in health and disease control. Individuals vary from one another in many ways such as the food they it, environmental factors, DNA and other physical conditions. Thus, the nature of diseases or disease control also varies from person to person as these factors affect the drug dosage needed or treatment decisions conducive to treat the disease. Therefore, it is only through personalized care that the medical institution can provide the right drug to the right patient for the right disease at the right time with the right dosage. So, it would not be an overstatement to state that personalized medicine is the future of medicine. Also, for employing the idea of personalized medicine for prescribing personalized treatment rules, a mathematical model using statistical inference and machine learning techniques can serve as a building base for drugs and treatment of the future. Following key things were considered for building the mathematical model:

- Acquiring of the training data.
- Selecting the method of inference based upon clinical covariates and data dimension.
- Identification of individualized treatment rules.
- Linear model fitting to the training data.
- Evaluation of model performance.
- Application of model for prescribing treatment rules to future patients.

Above listed steps were followed for model building, validation and implementation process. So, first some part of training data was acquired from Sanford profile weight management profile dataset. This initial dataset required some critical restructuring so that it can be used for analysis in two stage SMART design. So, the covariates were randomized from a comparative trial and BMI of each individual patients were calculated before, after 4 months and after 12 months of treatment. The two treatments were coded as either 1 or -1, hence two stages with two treatments at each stage groups out to four treatment decision rules. Thus, the training data consisted of 210 rows of patients or observations and 9 columns of covariates.

After acquiring the training dataset, *Q*-function approximation with regression analysis was chosen as method of inference to identify the individualized treatment rule. The model fitting was obtained using multiple regression model and the model parameters were estimated. *Q*-learning algorithm, a reinforcement learning technique that is based upon approximation of *Q*-functions using regression analysis was applied for performance evaluation of treatment decisions. Important covariates that could really impact the patient’s condition were selected as input to the regression model. The summary tables giving the values of each regression coefficients and value of error were obtained for regression analysis in two stages of SMART design. Multiple R-squared values of 88% and 90% were obtained for second stage and first stage regression analysis respectively.

Further, the model goodness-of-fit was evaluated using residual diagnostic plots to check if the assumption of linear model is met. From the residual diagnostic plots the assumptions of Linearity, Homoscedasticity, Independence and Normality were established for the multiple regression model, fulfilling the conditions and adequacy of regression fit. Finally, the model was analyzed for prescribing the treatment rule to future patients where, first the prognostic and predictive covariates for new patient was considered and was provided as input to the model. The model output was *Q*-values for both treatment stages and the treatment that resulted in maximum *Q*-values for both stages was selected as optimal treatment. Hence, the value of each treatment decision rule was also calculated using Horvitz-Thompson estimator and the one with maximum estimated regime value was selected as optimal treatment decision rule.

5.2 Conclusion

In conclusion, a mathematical model was developed and implemented to prescribe optimal treatment decisions to a patient depending upon his individual medical covariates. The developed model was used to administer treatments for obese patients enrolled in Sanford health weight management profile. Application of reinforcement learning algorithm in Sanford Profile weight management dataset is unprecedented and through model performance evaluation it can be inferenced that the model can be applied for prescribing treatment rule for future patients. A minimal residual standard error of 0.63 and 1.847 was obtained for second stage and first stage regression analysis respectively. The *Q*-values for each stage were determined using the estimated coefficients values from regression equation and the treatment decision that resulted in maximum *Q*-value was selected as optimal treatment. For example, a patient with prognostic covariates of *c* (1,0,34,30) in first stage of SMART design had *Q*-values of 1.023668 and 1.23254 when treatment 1 and -1 was administered respectively. So, in this case as treatment which is coded as -1 results in maximum *Q*-value, it is selected as optimal treatment in that stage. Similarly, same patient with *c* (1,30,24) as covariates in second stage had *Q*-values of 0.3803014 and -0.01900676, again for treatment 1 and -1 respectively. However, in this case treatment coded as 1 was chosen to be the optimal one as it resulted in maximum *Q*-value for that stage. Lastly, the value estimator was used to estimate the weighted average of each regime and the bar plots enables to assign the decision rule that results in maximum estimated regime values as optimal decision rule. Two cases were analyzed for two patients with different input covariates and it was found that A_2A_2 and A_2A_1 were the optimal decision rule for these patients. Therefore, it can be interpreted that the developed model using reinforcement learning and function approximation algorithm can be employed in estimating dynamic treatment regimens for an individual to provide a personalized healthcare.

5.3 Future Works

Many more modifications and enhancement can be included in above mathematical model for estimating optimal DTR. First, the training data structure can be well maintained by constructing a database that stores the values of all essential covariates such as gender, race, height and weights at fixed intervals. So, Sanford Health can be suggested to construct a well-maintained database for future use such as training a mathematical model. Secondly, the problem can be generalized by extending two stage SMART design to *n* stage randomization, however, for this scenario there should be knowledge of covariates value until *n* stages also, focusing the importance of data acquisition and database maintenance. Next, the binary nature of treatment decision can be extended and analyzed where treatment decisions can be coded with more values than only as 1 and -1, for example a treatment decision can be coded as 0 and applied as an input to the model. Lastly, other regression analysis method such as non-linear regression can be used and also more robust function approximation methods can be implemented, for example *Q*-learning with Mixed Residuals (*Q*L-MR).

Thus, these future works can further improve the model performance and accuracy for selecting the optimal treatment decision rules.

References:

[1] H. Lakkaraju and C. Rudin, "Learning Cost-Effective and Interpretable Regimes for Treatment Recommendation," *arXiv preprint arXiv:1611.07663*, 2016.

[2] M. P. Wallace and E. E. Moodie, "Personalizing medicine: a review of adaptive treatment strategies," *Pharmacoepidemiology and drug safety*, vol. 23, no. 6, pp. 580-585, 2014.

[3] E. J. Topol and D. Hill, *The creative destruction of medicine: How the digital revolution will create better health care*. Basic Books New York, 2012.

[4] L. M. Collins, S. A. Murphy, and K. L. Bierman, "A conceptual framework for adaptive preventive interventions," *Prevention science*, vol. 5, no. 3, pp. 185-196, 2004.

[5] P. W. Lavori and R. Dawson, "A design for testing clinical strategies: biased adaptive within-subject randomization," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 163, no. 1, pp. 29-38, 2000.

[6] I. Nahum-Shani *et al.*, "Experimental design and primary data analysis methods for comparing adaptive interventions," *Psychological methods*, vol. 17, no. 4, p. 457, 2012.

[7] Y. Zhao, D. Zeng, M. A. Socinski, and M. R. Kosorok, "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer," *Biometrics*, vol. 67, no. 4, pp. 1422-1433, 2011.

[8] J. Pineau, M. G. Bellemare, A. J. Rush, A. Ghizaru, and S. A. Murphy, "Constructing evidence-based treatment strategies using methods from computer science," *Drug & Alcohol Dependence*, vol. 88, pp. S52-S60, 2007.

[9] S. A. Murphy, D. W. Oslin, A. J. Rush, and J. Zhu, "Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders," *Neuropsychopharmacology*, vol. 32, no. 2, p. 257, 2007.

[10] J. Ma, B. P. Hobbs, and F. C. Stingo, "Statistical methods for establishing personalized treatment rules in oncology," *BioMed research international,* vol. 2015, 2015.

[11] J. Ma, F. C. Stingo, and B. P. Hobbs, "Bayesian predictive modeling for genomic based personalized treatment selection," *Biometrics,* vol. 72, no. 2, pp. 575-583, 2016.

[12] I. Nahum-Shani *et al.*, "Q-learning: A data analysis method for constructing adaptive interventions," *Psychological methods,* vol. 17, no. 4, p. 478, 2012.

[13] E. E. Moodie, B. Chakraborty, and M. S. Kramer, "Q-learning for estimating optimal dynamic treatment rules from observational data," *Canadian Journal of Statistics,* vol. 40, no. 4, pp. 629-645, 2012.

[14] E. B. Laber, K. A. Linn, and L. A. Stefanski, "Interactive model building for Q-learning," *Biometrika,* vol. 101, no. 4, pp. 831-847, 2014.

[15] A. J. Rush *et al.*, "Sequenced treatment alternatives to relieve depression (STAR* D): rationale and design," *Contemporary Clinical Trials,* vol. 25, no. 1, pp. 119-142, 2004.

[16] P. J. Schulte, A. A. Tsiatis, E. B. Laber, and M. Davidian, "Q-and A-learning methods for estimating optimal dynamic treatment regimes," *Statistical science: a review journal of the Institute of Mathematical Statistics,* vol. 29, no. 4, p. 640, 2014.

[17] A. A. Agyeman and R. Ofori-Asenso, "Perspective: Does personalized medicine hold the future for medicine?," *Journal of pharmacy & bioallied sciences,* vol. 7, no. 3, p. 239, 2015.

[18] K. K. Jain, "Future of personalized medicine," in *Textbook of Personalized Medicine*: Springer, 2015, pp. 693-708.

[19] B. Obama, "The genomics and Personalized Medicine Act of 2006," *Clinical advances in hematology & oncology: H&O,* vol. 5, no. 1, pp. 39-40, 2007.

[20] G. Parmigiani, *Modeling in medical decision making: a Bayesian approach*. Wiley, 2002.

[21] A. Wald, "Statistical decision functions," *The Annals of Mathematical Statistics,* pp. 165-205, 1949.

[22] C. F. Manski, "Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice," *Journal of Econometrics,* vol. 95, no. 2, pp. 415-442, 2000.

[23] C. F. Manski, "Treatment choice under ambiguity induced by inferential problems," *Journal of Statistical Planning and Inference,* vol. 105, no. 1, pp. 67-82, 2002.

[24] C. F. Manski, "Statistical treatment rules for heterogeneous populations," *Econometrica,* vol. 72, no. 4, pp. 1221-1246, 2004.

[25] R. H. Dehejia, "Program evaluation as a decision problem," *Journal of Econometrics,* vol. 125, no. 1-2, pp. 141-173, 2005.

[26] K. Hirano and J. R. Porter, "Asymptotics for statistical treatment rules," *Econometrica,* vol. 77, no. 5, pp. 1683-1701, 2009.

[27] M. Qian and S. A. Murphy, "Performance guarantees for individualized treatment rules," *Annals of statistics,* vol. 39, no. 2, p. 1180, 2011.

[28] J. M. Robins, "Optimal structural nested models for optimal sequential decisions," in *Proceedings of the second seattle Symposium in Biostatistics,* 2004, pp. 189-326: Springer.

[29] P. F. Thall, R. E. Millikan, and H.-G. Sung, "Evaluating multiple treatment courses in clinical trials," *Statistics in medicine,* vol. 19, no. 8, pp. 1011-1028, 2000.

[30] P. F. Thall, H.-G. Sung, and E. H. Estey, "Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials," *Journal of the American Statistical Association,* vol. 97, no. 457, pp. 29-39, 2002.

[31] P. F. Thall, L. H. Wooten, C. J. Logothetis, R. E. Millikan, and N. M. Tannir, "Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring," *Statistics in medicine,* vol. 26, no. 26, pp. 4687-4702, 2007.

[32] S. A. Murphy, "An experimental design for the development of adaptive treatment strategies," *Statistics in medicine,* vol. 24, no. 10, pp. 1455-1481, 2005.

[33] P. W. Lavori and R. Dawson, "Adaptive treatment strategies in chronic disease," *Annu. Rev. Med.,* vol. 59, pp. 443-453, 2008.

[34] J. K. Lunceford, M. Davidian, and A. A. Tsiatis, "Estimation of Survival Distributions of Treatment Policies in Two-Stage Randomization Designs in Clinical Trials," *Biometrics,* vol. 58, no. 1, pp. 48-57, 2002.

[35] A. S. Wahed and A. A. Tsiatis, "Optimal Estimator for the Survival Distribution and Related Quantities for Treatment Policies in Two-Stage Randomization Designs in Clinical Trials," *Biometrics,* vol. 60, no. 1, pp. 124-133, 2004.

[36] A. S. Wahed and A. A. Tsiatis, "Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data," *Biometrika,* vol. 93, no. 1, pp. 163-177, 2006.

[37] S. A. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 65, no. 2, pp. 331-355, 2003.

[38] J. Neyman and K. Iwaszkiewicz, "Statistical problems in agricultural experimentation," *Supplement to the Journal of the Royal Statistical Society,* vol. 2, no. 2, pp. 107-180, 1935.

[39] D. B. Rubin, "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics,* pp. 34-58, 1978.

[40] J. Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect," *Mathematical modelling,* vol. 7, no. 9-12, pp. 1393-1512, 1986.

[41] R. Bellman, "Dynamic Programming, Princeton, NJ: Princeton Univ," ed: versity Press. BellmanDynamic Programming1957, 1957.

[42] M. R. Kosorok and E. E. Moodie, *Adaptive TreatmentStrategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. SIAM, 2015.

[43] P. W. Lavori and R. Dawson, "Dynamic treatment regimes: practical design considerations," *Clinical trials,* vol. 1, no. 1, pp. 9-20, 2004.

[44] D. Oslin, "Managing alcoholism in people who do not respond to naltrexone (EXTEND)," *National Institutes of Health,* 2005.

[45] W. E. Pelham Jr and G. A. Fabiano, "Evidence-based psychosocial treatments for attention-deficit/hyperactivity disorder," *Journal of Clinical Child & Adolescent Psychology,* vol. 37, no. 1, pp. 184-214, 2008.

[46] R. M. Stone *et al.*, "Granulocyte–macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia," *New England Journal of Medicine,* vol. 332, no. 25, pp. 1671-1677, 1995.

[47] R. M. Stone *et al.*, "Postremission therapy in older patients with de novo acute myeloid leukemia: a randomized trial comparing mitoxantrone and intermediate-dose cytarabine with standard-dose cytarabine," *Blood,* vol. 98, no. 3, pp. 548-553, 2001.

[48] D. Tummarello *et al.*, "A randomized, controlled phase III study of cyclophosphamide, doxorubicin, and vincristine with etoposide (CAV-E) or teniposide (CAV-T), followed by recombinant interferon-α maintenance therapy or observation, in small cell lung carcinoma patients with complete responses," *Cancer,* vol. 80, no. 12, pp. 2222-2229, 1997.

[49] K. K. Matthay *et al.*, "Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid," *New England Journal of Medicine,* vol. 341, no. 16, pp. 1165-1173, 1999.

[50] K. K. Matthay *et al.*, "Long-term results for children with high-risk neuroblastoma treated on a randomized trial of myeloablative therapy followed by 13-cis-retinoic acid: a children's oncology group study," *Journal of clinical oncology,* vol. 27, no. 7, pp. 1007-1013, 2009.

[51] T. M. Habermann *et al.*, "Rituximab-CHOP versus CHOP alone or with maintenance rituximab in older patients with diffuse large B-cell lymphoma," *Journal of Clinical Oncology,* vol. 24, no. 19, pp. 3121-3127, 2006.

[52] M.-V. Mateos *et al.*, "Bortezomib, melphalan, and prednisone versus bortezomib, thalidomide, and prednisone as induction therapy followed by maintenance treatment with bortezomib and thalidomide versus bortezomib and prednisone in elderly patients with untreated multiple myeloma: a randomised trial," *The lancet oncology,* vol. 11, no. 10, pp. 934-941, 2010.

[53] S. F. Auyeung *et al.*, "Sequential multiple-assignment randomized trial design of neurobehavioral treatment for patients with metastatic malignant melanoma undergoing high-dose interferon-alpha therapy," *Clinical Trials,* vol. 6, no. 5, pp. 480-490, 2009.

[54] C. Kasari *et al.*, "Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial," *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 53, no. 6, pp. 635-646, 2014.

[55] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (no. 1). MIT press Cambridge, 1998.

[56] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research,* vol. 4, pp. 237-285, 1996.

[57] B. Chakraborty and E. Moodie, *Statistical methods for dynamic treatment regimes*. Springer, 2013.

[58] C. J. C. H. Watkins, "Learning from delayed rewards," King's College, Cambridge, 1989.

[59] F. Sahba, H. R. Tizhoosh, and M. M. Salama, "Application of reinforcement learning for segmentation of transrectal ultrasound images," *BMC medical imaging,* vol. 8, no. 1, p. 8, 2008.

[60] M. P. Wallace and E. E. Moodie, "Doubly-robust dynamic treatment regimen estimation via weighted least squares," *Biometrics,* vol. 71, no. 3, pp. 636-644, 2015.

[61] S. A. Murphy, "A generalization error for Q-learning," *Journal of Machine Learning Research*, vol. 6, no. Jul, pp. 1073-1097, 2005.

[62] D. Blatt, S. Murphy, and J. Zhu, "A-learning for approximate planning," *Ann Arbor*, vol. 1001, pp. 48109-2122, 2004.

[63] J. N. Tsitsiklis and B. Van Roy, "Feature-based methods for large scale dynamic programming," *Machine Learning*, vol. 22, no. 1-3, pp. 59-94, 1996.

[64] S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy, "Informing sequential clinical decision-making through reinforcement learning: an empirical study," *Machine learning*, vol. 84, no. 1-2, pp. 109-136, 2011.

[65] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[66] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 503-556, 2005.

[67] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *ICML*, 1999, vol. 99, pp. 278-287.

[68] M. J. Matarić, "Learning in behavior-based multi-robot systems: Policies, models, and other agents," *Cognitive Systems Research*, vol. 2, no. 1, pp. 81-93, 2001.

[69] G. Tesauro, "Practical issues in temporal difference learning," in *Advances in neural information processing systems*, 1992, pp. 259-266.

[70] S. Gelly and D. Silver, "Combining online and offline knowledge in UCT," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 273-280: ACM.

[71] B. Adams *et al.*, "HIV dynamics: modeling, data analysis, and optimal treatment protocols," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 10-49, 2005.

[72] C. f. M. P. f. H. Use, "Guideline on adjustment for baseline covariates," ed: European Medicines Agency, 2014.

[73] K. A. Linn, E. B. Laber, and L. A. Stefanski, "iqLearn: Interactive Q-learning in R," *Journal of statistical software,* vol. 64, no. 1, 2015.

[74] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012.

[75] A. C. Davison and D. V. Hinkley, "Bootstrap methods and their application (Vol. 1)," 1997.

[76] D. Almirall, T. Ten Have, and S. A. Murphy, "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation," *Biometrics,* vol. 66, no. 1, pp. 131-139, 2010.

[77] R. Henderson, P. Ansell, and D. Alshibani, "Regret-Regression for Optimal Dynamic Treatment Regimes," *Biometrics,* vol. 66, no. 4, pp. 1192-1201, 2010.

[78] J. C. Nankervis, "Computational algorithms for double bootstrap confidence intervals," *Computational statistics & data analysis,* vol. 49, no. 2, pp. 461-475, 2005.

[79] B. Chakraborty, S. Murphy, and V. Strecher, "Inference for non-regular parameters in optimal dynamic treatment regimes," *Statistical methods in medical research,* vol. 19, no. 3, pp. 317-343, 2010.

[80] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association,* vol. 47, no. 260, pp. 663-685, 1952.