Identifying Precision Treatment for Rheumatoid Arthritis with Reinforcement Learning

Chixiang Chen[1], Ashley Gannon[2], Duwani Katumullage[3], Miaoqi Li[4], Mengfei Liu[5], Rebecca North[6], Jialu Wang[7]

Industry Mentors: Grant Weller[8], Victoria Mansfield[9], Yinglong Guo[10]
Faculty Mentor: Daniel Luckett [11]

Abstract

Rheumatoid Arthritis (RA) is an autoimmune disease that causes chronic inflammation in the lining of joints leading to multiple complications including painful swelling, long-term damage from bone erosion, and joint deformity. The diagnosis and treatment of RA with precision medicine is challenging, as there is a high rate of co-morbid conditions that effect various organs and there are currently no clinically validated methods for measuring disease progression. In this work, we analyze a set of longitudinal administrative health data from more than 6,000 RA patients, and propose a framework to identify an optimal individualized dynamic treatment regime (DTR) by applying Q-learning which is a popular data-driven methodology for multistage decision problems in both randomized trials and observed data.

1 Introduction

Rheumatoid arthritis (RA) is a heterogeneous, chronic inflammatory disease that affects 1% of people worldwide [1, 2]. RA predominately affects the lining of the joints, causing pain and inflammatory flare ups. It also effects organs such as the heart, skin, eyes and lungs, causing growths, vasculitis, scleritis, Sjogren's syndrome, and a laundry list of other ailments. In addition to the discomfort and co-morbid conditions, patients diagnosed with RA have a 60% increase in the risk of heart attack one year after diagnosis and are twice as likely as the average person to develop depression. The variety of symptoms that can manifest in this patient population makes RA a difficult disease to diagnose and manage. Additionally, there is currently no single, autoantibody specific diagnostic test available to diagnose patients [3], and the American College of Rheumatology (ACR) does not provide an optimal treatment regime to follow once the patient has been diagnosed. Thus, determining which of the many treatment regimes provided by the ACR is the optimal treatment regime for RA is the focus of this work.

Because RA is a chronic illness, it is imperative to develop a long-term treatment strategy. Treatments vary from person to person and can depend on factors such as treatment history, disease state, age, personal preferences, etc.. Dynamic treatment regimes (DTRs) have been used to generate long-term individualized treatment plans for patients with a variety of chronic illnesses (e.g. as in [4]). In theory, this framework maps an individual's current characteristics to an optimized set of possible treatments at each decision point. In reality, it is an onerous task to identify optimal DTRs from large data sets. One possible way to approximate DTRs is to use a machine learning approach, specifically a reinforcement learning approach. Reinforcement learning can be used to optimize the expected outcome and produce a data-driven policy. In this work, we will implement one such method, an approximate dynamic programming method known as Q-learning, with the intent to identify an optimal treatment regime for RA patients based on certain input variables. While there are several studies aiming to identify the comparative effectiveness of DTR for RA patients, to our knowledge, no data-driven effort to identify an optimal DTR for RA has been published.

[1]Division of Biostatistics and Bioinformatics, Pennsylvania State University
[2]Department of Scientific Computing, Florida State University
[3]Department of Statistics, Sam Houston State University
[4]Department of Mathematical Sciences, University of Cincinnati
[5]Department of Statistics, The George Washington University
[6]Department of Statistics, NC State University
[7]Department of Statistics, The George Washington University
[8]UnitedHealth Group, Research & Development
[9]UnitedHealth Group, Research & Development
[10]UnitedHealth Group, Research & Development
[11]Department of Biostatistics, University of North Carolina

Q-learning has become a popular method used to identify optimal DTRs from observational studies (e.g. as in [5], [6], and [7]), and is an appropriate method to apply here. The data in this work, provided by UnitedHealth Group, are composed of observational longitudinal data, derived from health insurance claims from more than 6,500 RA patients. Most studies that implement Q-learning to identify DTRs assume that the data are collected at small, finite numbers of treatment intervals. In this data set, there exists at minimum 1 year of claims following diagnosis of RA, and a minimum of 6 months of claims prior to diagnosis for each patient. All patient attributes in this data, which include comorbidities, number of visits to medical facilities and prescribed medications, and treatment information, are updated on a weekly basis. Applying a Q-learning algorithm to this data set can be challenging due to an imbalance of observation in treatments, potential latent confounds, and unclear implementation instructions for treatment intervals. Our team will address how we handled these difficulties in Section 3.

The remainder of this paper is organized as follows. Section 1 offers a plethora of descriptive statistics of the full data set and describes the data cleaning process. Section 3 details the basic methodology of Q-learning, followed by fitting a generalized linear model to the first two months of data, then implementing the Q-learning algorithm on the first three months of data. The results of these analyses are presented in Section 4, with a concluding discussion of our findings and suggested future work in Section 5.

2 Data Processing

2.1 Descriptive Statistics

To better understand the full data set, we first perform extensive descriptive statistical analyses on the attributes recorded for each patient. These attributes include age, gender, medical appointments, specific treatments, and preexisting medical conditions. This key step will aid in justifying the models that we develop later on.

The first attributes we consider are gender, age, and comorbidities. From Table 1, we note that 74% of patients are female, and half of the population is between the ages of 41 and 56. Since RA is known to affect nearly three times as many women as it does men, and RA most commonly develops between the ages of 30 and 60, this sample appears to be quite representative of the general population of individuals diagnosed with RA. We also note that there are several comorbidities listed that are not common within the given population (<5%), namely AIDS/HIV, acute myocardial infarction (AMI), dementia, paralysis, and renal failure. On the other hand, the three most common comorbidities are hypertension, chronic obstructive pulmonary disease (COPD), and depression.

	N = 6846		N = 6846
Age at First Diagnosis		**Dementia**	
min	12	Prior	7 (0)
max	64	Post	11 (0)
mean (sd)	47.26 ± 10.99	**Diabetes**	
median (iqr)	50.00 (41.00, 56.00)	Prior	782 (11)
Gender		Post	866 (13)
Male	1,753 (26)	**Hypertension**	
Female	5,093 (74)	Prior	2182 (32)
AIDS/HIV		Post	2315 (34)
Prior	9 (0)	**Liver Disease**	
Post	13 (0)	Prior	433 (6)
Acute Myocardial Infarction		Post	557 (8)
Prior	67 (1)	**Paralysis**	
Post	81 (1)	Prior	34 (0)
Angina		Post	32 (0)
Prior	409 (6)	**Peripheral Vascular Disease**	
Post	417 (6)	Prior	243 (4)
Cancer		Post	275 (4)
Prior	294 (4)	**Renal Failure**	
Post	357 (5)	Prior	90 (1)
Cerebrovascular Disease		Post	192 (3)
Prior	303 (4)	**Ulcers**	
Post	296 (4)	Prior	119 (2)
Congestive Heart Failure		Post	97 (1)
Prior	131 (2)	**Depression**	
Post	192 (3)	Prior	1,276 (19)
COPD		Post	1,129 (16)
Prior	1,423 (21)	**Skin Ulcers**	
Post	1,296 (19)	Prior	865 (13)
		Post	565 (8)

Table 1: Descriptive table of age, gender and comorbidities. Continuous data are summarized by the minimum, maximum, mean (standard deviation), and median (interquartile range). The remaining binary data is summarized by their frequency and the percent of the population with the comorbidity (values in ()).

Flag Number	**DMARD**	**NSAID**	**Glucocorticoid**	**Opioid**
0	628,694	1,036,441	1,002,517	1,117,019
1	499,013	91,266	125,190	10,688
Percent of Observations:	44.25%	8.09%	11.10%	0.95%

Table 2: Number of times a medication is prescribed after patient diagnosis across all weeks.

For each week, each patient has a treatment ID that denotes his/her treatment for that week. The treatment IDs are a series of five indicators, one for each of the five types of disease-modifying antirheumatic drug (DMARD) that can be prescribed: traditional synthetic DMARDs, namely hydroxychloroquine, leflunomide, and sulfasalazine; methotrexate (MTX), a synthetic DMARD that is also used to treat cancer patients; non anti-tumor necrosis factor (nTNF), biologic DMARD that is composed of monoclonal anti-tumor necrosis factor antibodies and inhibits TNF in the body; anti-tumor necrosis factor (TNF), another type of biologic DMARD that inhibits TNF, but instead contains soluble TNF receptors; and tofacitnib (tofa), a biologic DMARD that inhibits the JAK1 enzyme and disrupts cell signaling pathways. Figure 1 gives an example of a possible treatment ID, and Table 3 gives the frequency with which each combination of DMARDs was

followed as a treatment. The first item to note from Table 3 is that the treatment ID "00000", which indicates no DMARD use, has the highest frequency of all treatments. This seemingly counter-intuitive event is likely due to the skewed distribution of treatment initiation as depicted in Figure 2. Second, there are quite a few treatment groups that have significantly smaller frequencies. Upon investigation of this phenomenon, we discovered that these treatment IDs correspond to a transition between two more common treatment groups rather than to a legitimate treatment option.

Figure 1: An example of a treatment ID. In this instance, the patient has been prescribed 1 traditional synthetic DMARD and 1 TNF.

Table 3: Frequency of treatment occurrence across all weeks following patient diagnosis.

Treatment Group	Frequency	Treatment Group	Frequency
00000	413204	10000	158739
00001	2470	10001	993
00010	70072	10010	14935
00011	21	10020	54
00020	179	10100	1626
00100	5068	10101	5
00101	12	10110	16
00110	34	11000	38013
00200	3	11001	220
01000	145060	11010	5436
01001	753	11100	574
01010	31991	20000	14221
01011	7	20001	11
01020	66	20010	2113
01100	1635	20100	397
01101	2	21000	3694
01110	32	30000	916
01200	3		

Even though DMARDs are the primary drug therapy for RA patients to slow down the progression of the disease, RA is a chronic, heterogeneous disease and will thus manifest differently and at varying intensities across patients, and across time. This often leads RA patients to supplementing their primary drug therapy with a glucocorticoid, a non-steroidal anti-inflammatory drug (NSAID), or an opioid. Table 2 gives the frequency across all post-diagnosis weeks with which DMARDs, NSAIDs, glucocorticoids, and opioids were used. We see that there were more weeks where one or more types of drug were not used after diagnosis, at least initially. While this is an encouraging result for the categories of NSAID, glucocorticoid, and opioid, it is something of a surprise for DMARDs. However, examination of the empirical distribution of treatment initiation in Figure 2 allows us to infer that the inflated number of non-DMARD weeks is a result of the lag between diagnosis and treatment commencement for a many of the patients. We can also note that, of the prescribed medications, the most common prescriptions are for disease-modifying antirheumatic drugs (DMARDs), with non-steroidal anti-inflammatory drugs (NSAIDs) and glucocorticoids being prescribed at similar frequencies and opioids prescribed the least.

Next, we consider the frequency with which medications were prescribed for each classification of drug among DMARDs, NSAIDs, glucocorticoids, and opioids. From Figure 3, we find that, by category, the most frequently prescribed medications are methotrexate sodium (36%) among DMARDs, Celecoxib (30%) among NSAIDs, Prednisone (97%) among glucocorticoids, and oxycodone HCL (47%) among opioids.

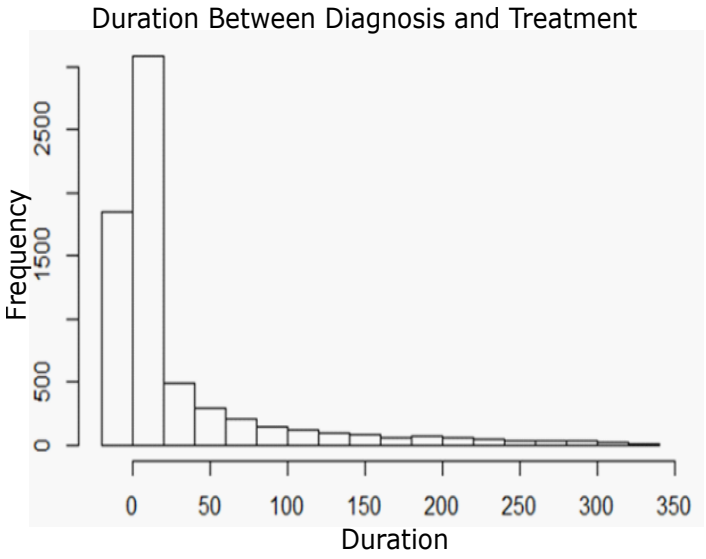
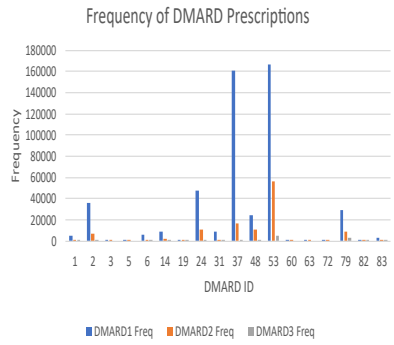
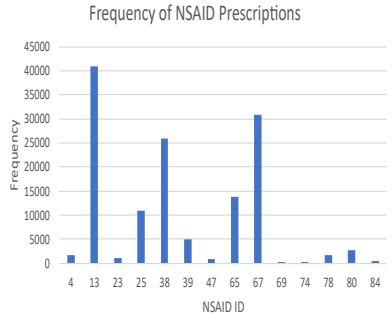


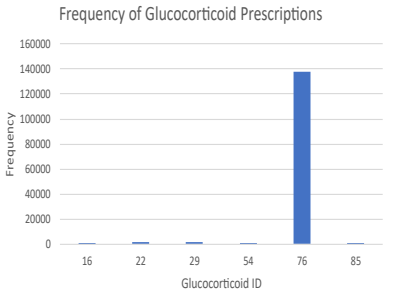
Figure 2: Empirical distribution of the number of weeks between patient diagnosis and the start of treatment. The minimum lag between diagnosis and treatment is recorded as -4 weeks, and the maximum lag is 336 weeks, or nearly 6 years. The former may be a gap in the claims data or the time it took for a full diagnosis. The median, at least, is 4 weeks, with the third quartile at 25 weeks. The ACR recommends that individuals diagnosed with RA begin DMARD treatment within the first 3 months after diagnosis [8]. This guideline appears to be followed by more than half of the RA patients in our data set, and nearly 75% began treatment in the first 6 months.



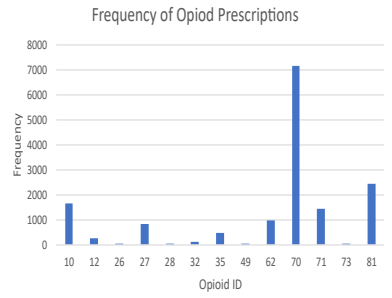
DMARD ID	Drug Name
1	Abatacept
2	Adalimumab
3	Anakinra
5	Auranofin
6	Azathioprine
14	Certolizumab Pegol
19	Cyclophosphamide
24	Etanercept
31	Golimumab
37	Hydroxychloroquine Sulfate
48	Ketoprofen
53	Methotrexate Sodium
60	Minocyc HCL/E16/SK CL4/TP.COM3
63	Mycophenolate Mofetil
72	Penicillamine
79	Sulfasalazine
82	Tocilizumab
83	Tofacitinib Citrate



NSAID ID	Drug Name
4	Aspirin
13	Celecoxib
23	Diflunisal
25	Etodolac
38	Ibuprofin
39	Indomethacin
47	Ketoprofin
65	Nabumetone
67	Naproxen
69	Oxaprozin
74	Piroxicam
78	Salsalate
80	Sulindac
84	Tolmetin Sodium



Glucocorticoid ID	Drug Name
16	Cortisone Acetate
22	Dexamethasone
29	Fludrocortisone Acetate
54	Methylprednisolone
76	Prednisone
85	Triamcinolone Acetonide



Opioid ID	Drug Name
10	Buprenorphine
12	Butorphanol Tartrate
26	Fentanyl
27	Fentanyl
28	Fentanyl Citrate
32	Hydrocodone Bitartrate
35	Hydromorphone HCL
49	Meperidine HCL
62	Morphine Sulfate
70	Oxycodone HCL
71	Oxymorphone HCL
73	Pentazocine Lactate
81	Tapentadol

Figure 3: Descriptive summary of the frequencies of medications prescribed to RA patients.

Next, we explore the frequency of transitions between the use of glucocorticoids and painkillers as well as the overall use of glucocorticoids and painkillers across all weeks, where painkillers include NSAIDs and

opioids. From Table 4, we notice that most patients do not change their supplemental treatments over time. When a change does occur, it is most often to stop taking a glucocorticoid or painkiller. From Table 5, we see that many RA patients were already using a painkiller or glucocorticoid before diagnosis, which suggests that they were either experiencing many RA symptoms before diagnosis or there were other preexisting conditions that called for the use of these drugs. This fact induces limitations in our analysis in Section 3 when using painkiller or glucocorticoid use as a surrogate response for treatment effectiveness.

Change In Treatment:	**Percentage**
No Glucocorticoid → Glucocorticoid	4.17%
Glucocorticoid → No Glucocorticoid	17.84%
No Painkiller → Painkiller	3.54%
Painkiller → No Painkiller	9.08%
Glucocorticoid → Painkiller	1.18%
Painkiller → Glucocorticoid	1.00%
No Change	63.18%

Table 4: Average percent of supplemental treatment changes that occur across all weeks.

	Painkillers	**Glucocorticoids**	**Both**
Before Diagnosis	37.2%	25.6%	18.0%
After Diagnosis	47.1%	65.6%	15.7%

Table 5: Percentage of supplemental medications prescribed to patients before and after RA diagnosis.

Finally, we consider various types of medical visits, including inpatient (IP) visits, length of IP stay, outpatient (OP) visits, emergency room (ER) visits, and doctor's (DR) visits, as well as filled prescriptions (RX scripts). Table 6 shows that the average RA patient will visit a doctor about once a month, go for an outpatient visit once every other month, and take one or two prescription medications each week.

	IP Visits	**IP Days**	**OP Visits**	**ER Visits**	**DR Visits**	**RX scripts**
mean	0.0009	0.0037	0.1484	0.0039	0.3127	1.340
sd	0.0222	0.0836	0.3528	0.0492	0.5609	1.543

Table 6: This table summarizes the average and standard deviation over all treatment IDs of medical visits and prescription fills.

2.2 Data Cleaning

To simplify the data, we first remove all weeks prior to the start of treatment for each patient, under the assumption that they do not contribute significantly to the choice of treatment. Next, since treatment decisions are typically made on a monthly basis rather than a weekly basis, we collapse the data from weekly records to monthly records. This change makes the data more informative and less computationally expensive. To convert the medical visits (IP, OP, ER, DR) and prescriptions into monthly records, we sum the number of events over the 4 weeks in each month. New treatment IDs are assigned by taking the maximum number that occurs in each column of the four original treatment IDs. We also create new variables to indicate the number of treatment changes that occurred in a given month, if the patient took a painkiller that month (0 if no, 1 if yes), and if the patient took a glucocorticoid that month (0 if no, 1 if yes).

After collapsing the data into monthly intervals, we remove 38 patients who have only one month of treatment data, as our methods require a second month of data to provide an outcome for the analysis. Further data cleaning is completed and noted as needed in the analysis sections.

3 Methodology

Let's start at the general notation. Our goal is to minimize the following the reward function to get the corresponding optimal decision rules:

$$(\ d_1(\boldsymbol{X}_1),\ d_2(\boldsymbol{X}_2)...\ d_n(\boldsymbol{X}_n)) = \underset{a_1,a_2...a_n}{\operatorname{argmin}} \mathbb{E}\{Y|A_1 = a_1, \boldsymbol{X}_1, A_2 = a_2, \boldsymbol{X}_2, ..., A_n = a_n, \boldsymbol{X}_n\} \tag{1}$$

Where

$$\begin{aligned} i = &\quad 1, 2, ...n \\ \boldsymbol{X}_i \equiv &\quad \text{covariate matrix at time point i} \\ d_i(\boldsymbol{X}_i) \equiv &\quad \text{optimal decision at time point i} \\ Y \equiv &\quad \text{last stage outcome} \\ A_i \equiv &\quad \text{treatment/decision at time point i.} \end{aligned}$$

In this paper, let Y denote the outcome that indicates treatment efficacy. The true outcome of interest is the frequency of flareups. However, since claims data does not provide an exact measurement of these events, we chose the use of painkillers or glucocorticoids to be a surrogate response. The covariates we use to predict this outcome could be any subset of the patient attributes that are provided in the claims. We allow some overlap among these covariates $\boldsymbol{X}_i$ at different stages. The decision space for a_i, $i = 1, ..., n$, can also vary across time. The key technique we apply to solve (1) is Q-learning. In the next subsection, we illustrate our method by first starting at the simplest situation that only involves a one-stage decision rule, then extend it to a two-stage decision-making process.

3.1 A One-Decision Model

This is the fundamental stage of the DTR, when we only consider the baseline cross sectional data for each patient at the initial stage of their disease. Hence, the dynamic aspect of RA need not be taken into account yet.

In (2), we conform to the general notation given at the beginning of this section to discover the treatment groups that minimize the expected value of the outcome, which we call the optimal decision rule.

$$\hat{d}(x) = \underset{a}{\operatorname{argmin}}\ \mathbb{E}(\hat{Y}|\boldsymbol{X} = \boldsymbol{x}, A = a). \tag{2}$$

Since only one decision is considered here, we can apply logistic regression to model $\mathbb{E}(Y|\boldsymbol{X}, A)$ with the training data, then directly obtain the optimal decision rule by minimizing the estimated model from the test data.

To evaluate the performance of the optimal decision rule, we need to calculate and compare the predicted $\mathbb{E}(Y|A)$ based on the optimized decision rule and observed treatment combination in the test data, respectively. A naive prediction formula is given by (3):

$$\tilde{Y} = \frac{\left(\sum_i^n Y_i \mathbb{1}\{A_i = \hat{d}(\boldsymbol{x}_i)\}\right)}{\left(\sum_i^n \mathbb{1}\{A_i = \hat{d}(\boldsymbol{x}_i)\}\right)},\quad \tilde{Y}_{obs} = \frac{\left(\sum_i^n Y_i \mathbb{1}\{A_i = a_i\}\right)}{\left(\sum_i^n \mathbb{1}\{A_i = a_i\}\right)}. \tag{3}$$

However, (3) is not typically acceptable due to the fact that treatment assignment is not randomized in an observational study, and thus cannot guarantee an unbiased estimator for $\mathbb{E}(Y|A)$. To account for the bias, we can apply the method of *inverse probability treatment weights* (IPTW), as in [9]:

$$\tilde{Y}_{IPTW} = \frac{\left(\sum_i^n \frac{Y_i \mathbb{1}\{A_i = \hat{d}(\boldsymbol{x}_i)\}}{\hat{P}(A_i|\boldsymbol{x}_i)}\right)}{\left(\sum_i^n \frac{\mathbb{1}\{A_i = \hat{d}(\boldsymbol{x}_i)\}}{\hat{P}(A_i|\boldsymbol{x}_i)}\right)},\quad \tilde{Y}_{IPTW-obs} = \frac{\left(\sum_i^n \frac{Y_i \mathbb{1}\{A_i = a_i\}}{\hat{P}(A_i = a_i|\boldsymbol{x}_i)}\right)}{\left(\sum_i^n \frac{\mathbb{1}\{A_i = a_i\}}{\hat{P}(A_i = a_i|\boldsymbol{x}_i)}\right)}. \tag{4}$$

Here, $\tilde{Y}_{\text{IPTW}}$ is the IPTW estimator based on the optimized decision rule, and $\tilde{Y}_{\text{IPTW}-obs}$ is the IPTW estimator based on the observed treatment combination. The quantity $P(A|\boldsymbol{X})$ is the propensity score utilized to adjust for bias, which can be modeled by multinomial logistic regression.

However, in cases where $P(A|\boldsymbol{X})$ is misspecified, the estimation will still carry some bias. Thus, with the aim of making the estimation more resilient, we introduce a more robust formula that incorporates both the IPTW and the conditional expectation of the outcome, $\mathbb{E}(Y|\boldsymbol{X}, A)$, known as the double robust estimator [10]. The estimator in (5) allows for the misspecification of either $P(A|\boldsymbol{X})$ or $\mathbb{E}(Y|\boldsymbol{X}, A)$, but not both.

$$\tilde{Y}_{\text{IPTW-rob}} = \frac{1}{n}\sum_{i}^{n}\left\{\frac{Y_i\mathbb{1}(A_i = \hat{d}(\boldsymbol{x}_i))}{\hat{p}(A_i = \hat{d}(\boldsymbol{x}_i)|\boldsymbol{x}_i)} - \frac{\mathbb{1}(A_i = \hat{d}(\boldsymbol{x}_i)) - \hat{p}(A_i = \hat{d}(\boldsymbol{x}_i)|\boldsymbol{x}_i)}{\hat{p}(A_i = \hat{d}(\boldsymbol{x}_i)|\boldsymbol{x}_i)}\hat{\mathbb{E}}\{Y_i|\boldsymbol{x}_i, A_i = \hat{d}(\boldsymbol{x}_i)\}\right\}. \quad (5)$$

3.1.1 Data Set

To proceed with the aforementioned analysis, we extract the first month of data for all patients. We create a new outcome variable, a glucocorticoid-or-painkiller flag, by combining the individual indicators for glucocorticoid and painkiller use. We clean the data by eliminating uncommon treatment IDs. We define uncommon treatment IDs as an ID corresponding to fewer than 100 patients. The remaining treatment IDs for this section can be observed in Table 7. We remove these patients from this data set for a couple of reasons. The first is we want to avoid creating singular covariate matrices. The second is that we do not have enough replicates of these patients to make significant statistical conclusions regarding their treatment IDs. The final data set generated is then randomly divided into training data (70%) and testing data (30%).

Treatment ID
00010
01000
10000
11000

Table 7: Remaining treatment IDs after data cleaning.

3.1.2 Model Selection

To narrow down the covariates that we will use, we implemented two variable selection methods: Random Forest and backward selection of a logistic regression model. The random forest algorithm is implemented using the Random Forest package in R, and the logistic regression model is fitted using glm()in R. From these two methods and our descriptive statistical analyses, we examine the top covariates and determine which covariates, X, to be used in the approach.

Random Forest	**Backward selection**
Age at diagnosis	Gender
Treatment group for month 1	Acute myocardial infarction
COPD	Hypertension
Subscriber	Angina
Gender	Paralysis
Hypertension	Skin
	Renal failure
	Peripheral vascular disease
	Treatment group for month 1

Table 8: Covariates selected by Random Forest and backward selection.

In the logistic regression model, we notice the presence of covariates that rarely occur in the observed population ($< 5\%$), as given in Table 9. Since such a small proportion of patients suffer from such ailments, we do not consider them in our approach. We also remove the subscriber index, to avoid over-fitting the model and since this covariate is less informative than the others.

Covariate	Percent of Occurrence
Acute myocardial infarction	1%
Hypertension	34%
Angina	6%
Paralysis	1%
Skin ulcers	14%
Renal failure	1%
Peripheral vascular disease	1%

Table 9: Occurrence of covariates in two month data set.

The final covariates we use in our model can be seen in Table 10. The logistic regression model for the expected conditional outcome will consist of the main effects of each of these covariates, as well as the interaction effects between these covariates and the treatment group for month 1. Moreover, for the propensity scores, the treatment group for month 1 will be modelled against the rest of the covariates mentioned in Table 10. The detailed result of this analysis will be discussed in Section 4.

Final List of Covariates for the One Decision Model
Age at diagnosis
Treatment group for month 1
COPD
Hypertension
Angina
Skin ulcers
Gender

Table 10: Final list of covariates for this model.

3.2 A Two-Decision Model

The two-stage analysis is an extension of the one-stage analysis, and can easily be generalized to more than two stages based on the developed methodology and algorithms. Q-learning, a type of reinforcement learning, can be considered as one of the primary tools used in developing dynamic treatment regimes [5]. Let us define the Q-functions for a two stage analysis as follows [9]:

$$f_2(\boldsymbol{X}_2, A_2, \boldsymbol{X}_1, A_1) = \mathbb{E}[Y|\boldsymbol{X}_2, A_2, \boldsymbol{X}_1, A_1], \tag{6}$$

$$f_1(\boldsymbol{X}_1, A_1) = \mathbb{E}[\underset{a_2}{min}\ f_2(\boldsymbol{X}_2, a_2, \boldsymbol{X}_1, A_1)|\boldsymbol{X}_1, A_1]. \tag{7}$$

This algorithm implements backwards recursion to determine the optimal regime. We start with the final stage, stage 2 here, to optimize the treatment in that stage. Data are randomly divided into two parts, training and testing sets, composed of 70% and 30% of the patient population, respectively. Logistic regression is applied to the training set, and the individualized optimal treatments for stage 2 are determined based on (8):

$$\hat{d}_2(\boldsymbol{X}_2, \boldsymbol{X}_1, A_1) = \underset{a_2}{\text{argmin}}\ \hat{f}_2(\boldsymbol{X}_2, a_2, \boldsymbol{X}_1, A_1). \tag{8}$$

Next, for each patient in the training set, the predicted outcome, $Y_i^* \in (0, 1)$, is ascertained under the individualized optimal treatment regime in stage 2, and is used as the "pseudo-outcome" for stage 1. These pseudo-outcomes are transformed using the logit function, then fit to the covariates in the training set via linear regression. Finally, the individualized optimal regime for stage 1 is determined by (9):

$$\hat{d}_1(\boldsymbol{X}_1) = \underset{a_1}{\text{argmin}}\ \hat{f}_1(a_1, \boldsymbol{X}_1). \tag{9}$$

To evaluate the performance of the two-stage Q-learning model, similar criteria to that in Section 3.1 are utilized on the testing set, with exact expressions for these criteria given below:

$$
\begin{aligned}
\tilde{Y}_{\text{IPTW-obs}} =& \frac{\left(\sum_i^n \frac{Y_i \mathbb{1}\{A_{1i}=a_{1i}\}\mathbb{1}\{A_{2i}=a_{2i}\}}{\hat{P}(A_{1i}=a_{1i}|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=a_{2i}|\boldsymbol{x}_{2i})} \right)}{\left(\sum_i^n \frac{\mathbb{1}\{A_{1i}=a_{1i}\}\mathbb{1}\{A_{2i}=a_{2i}\}}{\hat{P}(A_{1i}=a_{1i}|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=a_{2i}|\boldsymbol{x}_{2i})} \right)}, \\
\tilde{Y}_{\text{IPTW}} =& \frac{\left(\sum_i^n \frac{Y_i \mathbb{1}\{A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})\}\mathbb{1}\{A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})\}}{\hat{P}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})|\boldsymbol{x}_{2i})} \right)}{\left(\sum_i^n \frac{\mathbb{1}\{A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})\}\mathbb{1}\{A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})\}}{\hat{P}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})|\boldsymbol{x}_{2i})} \right)}, \\
\tilde{Y}_{\text{IPTW-rob}} =& \frac{1}{n}\sum_i^n \Bigg\{ \frac{Y_i \mathbb{1}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i}))\mathbb{1}(A_{2i}=\hat{d}_1(\boldsymbol{x}_{2i}))}{\hat{P}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})|\boldsymbol{x}_{2i})} - \\
& \frac{\mathbb{1}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i}))\mathbb{1}(A_{2i}=\hat{d}_1(\boldsymbol{x}_{2i})) - \hat{P}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})|\boldsymbol{x}_{2i})}{\hat{P}(A_{1i}=\hat{d}_1(\boldsymbol{x}_{1i})|\boldsymbol{x}_{1i})\hat{P}(A_{2i}=\hat{d}_2(\boldsymbol{x}_{2i})|\boldsymbol{x}_{2i})} \\
& \times \hat{f}_1(\hat{d}_1(\boldsymbol{x}_{1i}),\boldsymbol{x}_{1i}) \Bigg\}.
\end{aligned}
\tag{10}
$$

Here, we assume that $A_1|\boldsymbol{X}_1$ is independent of $A_2|\boldsymbol{X}_2$ so that each propensity score can be separately fitted by a multinomial logistic model This assumption could be dropped by fitting multivariate multinomial model, but for convenience, we keep this assumption in the following analysis. To the author's knowledge, this may be the first proposal of the double robust estimation of $\mathbb{E}(Y|A_1,A_2)$ in a dynamic treatment regime among current literature.

3.2.1 Data Set

The data is again cleaned by eliminating uncommon treatment IDs, using similar criteria to that in Section 3.1.1. The remaining treatment IDs for this section are given in Table 11.

Treatment ID
00000
00010
01000
10000
11000

Table 11: Remaining treatment IDs after data cleaning.

3.2.2 Model Selection

Random Forest and backward selection of a logistic regression model are again utilized to detect the most important covariates for the analyses in stage 2. Table 12 summarizes the top variables selected using these two methods, respectively.

Random Forest	**Backward selection**
Glucocorticoid or painkiller in month 1	Treatment group in month 2
Age at diagnosis	Glucocorticoid or painkiller in month 1
Prescriptions in month 1	OP visits in month 1
Dr Visits in month 1	Prescriptions in month 1
Treatment group in month 2	
OP visits in month 1	
Treatment group in month 1	

Table 12: Covariates selected by Random Forest and backward selection.

We notice that all the covariates selected by the logistic regression method are a subset of covariates selected by the random forest method. We keep all of the covariates selected by both methods in addition to the age at diagnosis and the Dr Visits in month 1. The final covariates we use in our model can be seen in Table 13. It is worthwhile to mention that when applying those important covariates in stage 2 and stage 1, we also include the interaction terms between the Treatment group in month 2 and month 1 respectively with all these other covariates. Moreover, the propensity scores of treatments are obtained by fitting a multinomial logistic model in stages 1 and 2, using covariates listed on Table 10 and Table 13, respectively.

Final List of Covariates for the Two Decision Model
Treatment group in month 2
Glucocorticoid or painkiller in month 1
Age at diagnosis
Prescriptions in month 1
Dr Visits in month 1
OP visits in month 1

Table 13: Final list of covariates for this model.

4 Computational Results

4.1 A One-Decision Model

As discussed in Section 3.1.2. we use the logistic regression model to estimate the expected outcome, and use the corresponding test data to predict the outcome values for each patient under each treatment group observed. We then obtain the optimal rule (the optimal treatment group) for each patient based on the minimum predicted outcome.

Ultimately the set of optimal treatment groups we observed for this stage are can be referenced in table 7. The optimal treatments for patients belonging to different categories based on gender and comorbidity status, can be seen in Figure 4.

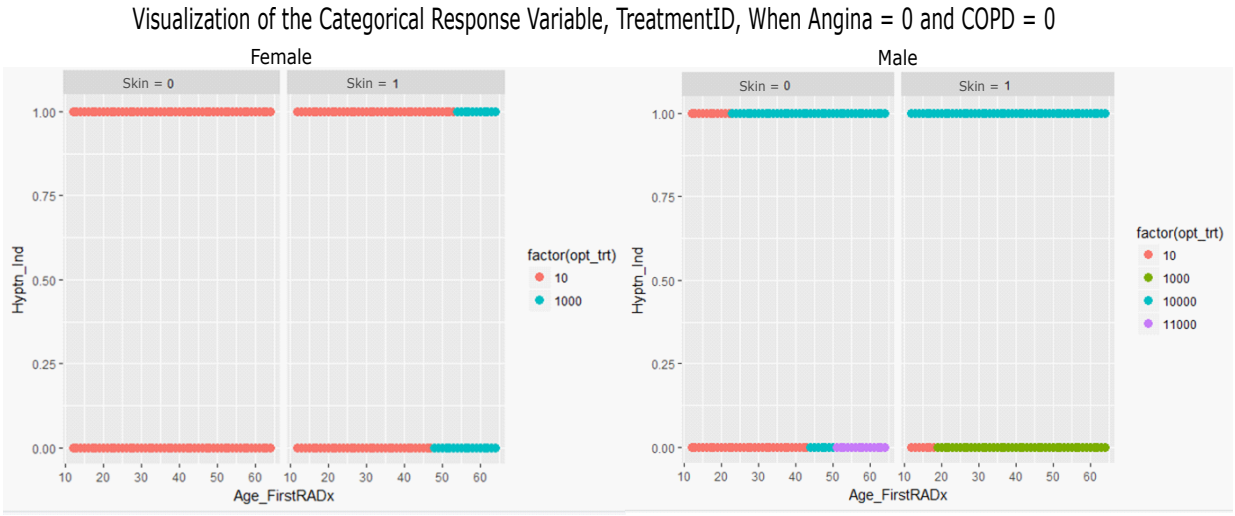


Figure 4: This figure displays the optimal treatments for patients who do not have COPD and Angina. For female, if the patients do not have skin ulcers, the treatment method consists of 1 TNF for all ages, regardless whether they have hypertension or not. If the female patients have Skin Ulcers, the treatment changes to 1 MTX at 46 years old for patients without hypertension, and at 53 years old with for paatients hypertension. For male patients, if they do not have skin ulcers, the treatment changes from 1 TNF to 1 DMARD for hypertension patients at around 21 years old. For those patients who do not have hypertension, the treatments change twice. At 42 years old, the treatment changes from one TNF to one DMARD, and at 50 years old, the treatment changes from one DMARD to a combination of one DMARD and one MTX. If the male patients have Skin Ulcers, the treatment is one DMARD for every patient who has hypertension. Whereas treatment changes from one TNF to one MTX at 19 years old for patients without hypertension.

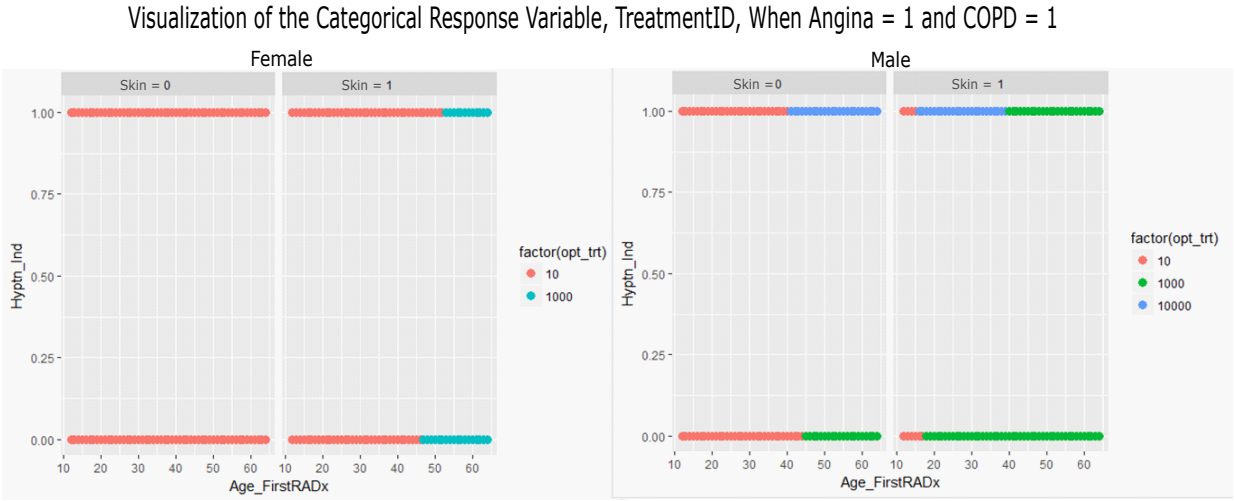


Figure 5: This figure displays treatments for patients who have COPD and Angina. For all female patients without skin ulcers, the treatment is 1 TNF. For the female patients who have Skin Ulcers, the treatment changes from 1 TNF to 1 MTX at 53 years old with hypertension, and the treatment changes from 1 TNF to 1 MTX at 46 years old without hypertension. For male patients, if they do not have Skin Ulcers, the treatment changes from 1 TNF to 1 DMARD at 40 years old with hypertension, and the treatment changes to 1 MTX at 43 years old without hypertension. If they have Skin Ulcers, the treatment changes from 1 TNF to 1 DMARD at 15 years old, and then changes to 1 MTX at 39 years old with hypertension. For those male patients who do not have hypertension, the treatment changes from 1 TNF to 1 MTX at 17 years old.

From Figures 5 and 5, we see that there are two optimal treatment groups in female patients: one TNF and one DMARD. For the female patients who do not have Skin Ulcers, the treatment with one TNF is suitable for all of them. And for those female patients who have Skin Ulcers, the treatment changes to one DMARD at 45 years old. The male patients' treatments are more diverse. For the male patients who have Skin Ulcers, the treatments change at very young ages, approximately 15 to 20 years old. The treatments change to one DMARD or one MTX based on whether they have hypertension or not. The most interesting group for male patients is that the patients who are having hypertension and Skin Ulcers, the treatment for this group stays the same no matter what is the age of patient, which is 1 DMARD.

We also compare the different estimators for the expected mean outcome of our two-month model that we discussed at Section 3.1 in Table 14. The estimated expected outcome under the estimator that does not account for bias is inaccurate. The more reliable estimators are the double robust estimator and the adjusted estimator that incorporates the IPTW. In either case, it is evident that the estimated expected value of outcome under the optimal treatment groups is smaller than that under the observed treatment groups.

Estimator	**Estimated Expected Value of Outcome**	**Observed Mean Outcome**
Without IPTW	0.141	0.186
Adjusted (With IPTW)	0.151	0.190
Double robust	0.150	0.184

Table 14: Comparison of estimates of the Expected Mean Outcome of Two-Month Model and the Observed Mean Outcome

4.2 A Two-Decision Model

As discussed in Section 3.2. we now obtain the individualized optimal treatment groups for two stages (stage 1 and stage 2) incorporating the dynamic aspect of the disease. The set of optimal decision rules we obtain for stage 1 can be referred to in table 7 and the set of optimal decision rules for stage 2 can be referred to in table 11. As in section 4, we visualize the optimal treatments for patients in, based on the covariates of the respective model.

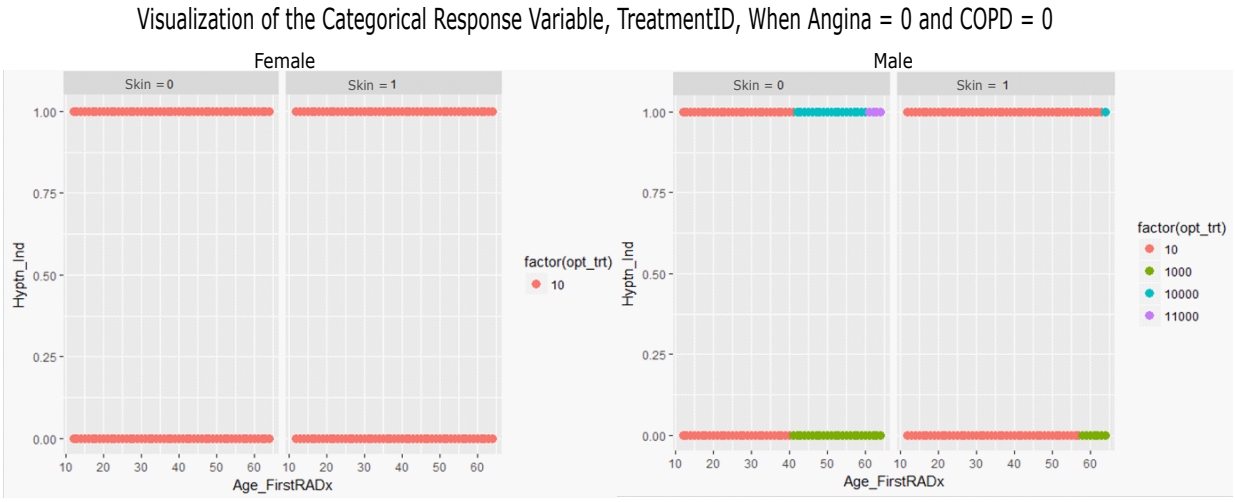


Figure 6: The figure represents patients who are not diagnosed with COPD and Angina. We can see for all female patients, whether they have hypertension or Skin Ulcers, they can be prescribed 1 TNF. However, the male patients' treatment groups are more diversity. For the male patients who do not have Skin Ulcers, their treatment groups change at age 40 when they first diagnosed RA. If they do not have hypertension, the treatment changes from 1 TNF to 1 MTX at age 40. But for the patients who have hypertension, the treatments need to be changed twice. At age 40, the treatment changes from 1 TNF to 1 DMARD. At 60, the treatment changes from 1 DMARD to 1 DMARD and 1 MTX. And for male patients who have Skin Ulcer, the treatment changes from 1 TNF to 1 DMARD if he is diagnosed RA at around 64 years old with hypertension. Without hypertension, the treatment changes from 1 TNF to 1 MTX at 56 years old.

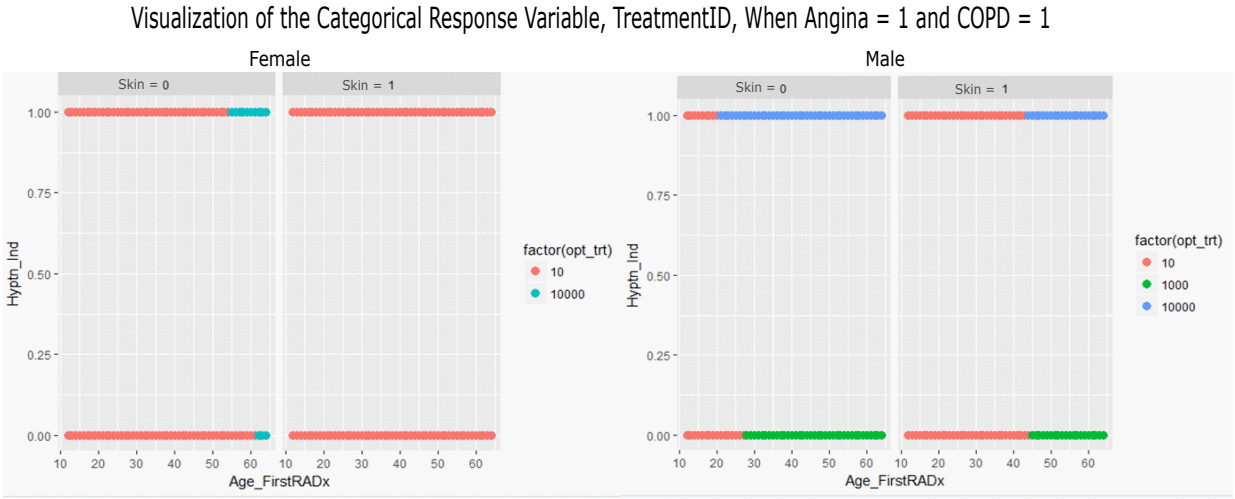


Figure 7: This figure represents the patients who have COPD and Angina. For female patients, if they have Skin Ulcers, regardless of hypertension and age at dagnosis, they are tretaed with 1 TNF. If they do not have Skin Ulcers, the treatment will change to 1 DMARD at 60 years old without hypertension, and at 55 years old with hypertension. For male patient, 1 DMARD and 1 MTX is the most common. If the patients do not have Skin Ulcers, the treatment changes from 1 TNF to 1 MTX at 26 years old without hypertension, and the treatments change from 1 TNF to 1 DMARD at 20 years old with hypertension. If they have Skin Ulcers, the treatments change from 1 TNF to 1 DMARD at 43 years old with hypertension, and from 1 TNF to 1 MTX at around 43 years old without hypertension.

From Figures 6 and 7, the treatment group appears straightforward for women. That is, 1 TNF will work

for most of the female patients. Only a small group of female patients who have COPD and Angina but no Skin Ulcers, use a different treatment: 1 DMARD. However, the treatments for male patients are slightly more complicated. For these two situations, there are four treatments being used: 1 TNF, 1 MTX, 1 DMARD and 1 MTX with 1 DMARD. On average, 1 DMARD treatment and 1 MTX treatment are more suitable for older patients. Except for the male patients who have COPD and angina, the treatment change begins earlier, at approximately 20 years old, for the patients who have hypertension, and at 26 years old for the patients who do not have hypertension. The treatment consisting of 1 MTX and 1 DMARD is only reccomended for a small group of male patients who only have hypertension.

Similar to the Section 4.1. we compare estimates of the expected value of outcome using the set of optimal decision rules with the observed mean outcome in Table 15. As expected, we can see that former is smaller than latter. Note how we have once again incorporated two different estimators; the adjusted and the double robust estimators.

Estimator	**Estimated Expected Value of Outcome**	**Observed Mean Outcome**
Adjusted (With IPTW)	0.040	0.122
Double robust	0.059	0.112

Table 15: Comparison of estimates of the Expected Mean Outcome of Three-Month Model and the Observed Mean Outcome

5 Summary and Future Work

We studied a framework for one and two decision problems in which the rewards are the expected probability of implementing treatment regimes. We proposed applicable Q-learning algorithm for precision medicine of RA patients and novel measurements for expected outcome to evaluate the effect of therapy. The work as presented can be extended to real-world multi-stage decision problems on observational data.

Interesting results are derived from our analysis, which might be instructive for future research work on RA. For example, in our one decision-two month model, we find that there are only two optimized treatment groups in female patients: for the female patients without skin ulcers, the treatment 1 TNF is suitable for all of them, and for the others with skin ulcers, the treatment changes to 1 DMARD at 45 years old. Similarly, We can see in the two decision model that all of the female patients, regardless of if they have hypertension or Skin Ulcers, are optimally prescribed 1 TNF. In contrast, the treatments of male patients are more diverse in both models. Finally, older patients appear to be more suitable for traditional DMARD treatments.

Despite the advantages of the our proposed method, there are theoretical limitations. First, the model we fit relies on the specified likelihood, which is not robust if parametric assumption is violated in practice. It will become more serious when there are more stages to deal with. However, this issue could be properly solved by applying a semi-parametric approach, such as generalized estimating equation method, or non-parametric one, such as random forest, to fit data. Second, the model selection procedure in multiple stages of the decision-making is tricky, as the outcomes in different stages are not observed until the last stage. Selecting the important covariates in each stage remains be an open research question both in theory and application. Third, we discard the patients who have only two months of observational data when we implement our 2 stage Q-learning algorithm, due to very low censoring rate. If we were to expand our model beyond 2 stages, we might lose some information if the censoring is not completely random. Thus, some techniques should be applied to deal with this issue, especially when the censoring rate is high. The standard error for predicted expectation $\tilde{Y}$ is also of our interest, which might be obtained by asymptotic derivation or bootstrap methods. Last but not the least, the global response, i.e. the overall effect of treatment combinations during the study, is more of interest to researchers instead of the outcome from the last stage alone. Thus, Finding a more efficient algorithm is also an open question.

There are also practical limitations for our model that should be dealt with in future work. For example, the outcome we defined in this study might not be the best indicator for the treatment More specified information, such as clinical information, is needed to correctly identify the response and covariates in the analysis. For example, levels of C-reactive protein (CRP) from a blood test, might be a better outcome for predicting the effectiveness of the therapy.

References

[1] Christopher B Atzinger. Biologic disease-modifying antirheumatic drugs in a national, privately insured population: Utilization, expenditures, and price trends. 10(1):10.

[2] C.A. Wijbrandts and P.P. Tak. Prediction of response to targeted treatment in rheumatoid arthritis. 92(7):1129–1143.

[3] Daniel Aletaha and Stephan Blüml. Therapeutic implications of autoantibodies in rheumatoid arthritis. 2(1):e000009.

[4] S. A. Murphy. Optimal dynamic treatment regimes. 65(2):331–355.

[5] Yair Goldberg and Michael R. Kosorok. Q-learning with censored data. 40(1):529–560.

[6] Rui Song, Weiwei Wang, Donglin Zeng, and Michael R. Kosorok. Penalized q-learning for dynamic treatment regimens.

[7] Yufan Zhao, Michael R. Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. 28(26):3294–3315.

[8] Katrina E. Donahue, Gerald Gartlehner, Daniel E. Jonas, Linda J. Lux, Patricia Thieda, Beth L. Jonas, Richard A. Hansen, Laura C. Morgan, and Kathleen N. Lohr. Systematic review: Comparative effectiveness and harms of disease-modifying medications for rheumatoid arthritis. 148(2):124.

[9] Erica E. M. Moodie, Bibhas Chakraborty, and Michael S. Kramer. Q-learning for estimating optimal dynamic treatment rules from observational data. 40(4):629–645.

[10] Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. 68(4):1010–1018.