CS234 - Winter 2019
Reinforcement Learning for Warfarin Dose Estimation

Andriy Syrov [*1] **Yun Zhou** [*2] **Ming-Kuang Daniel Wu** [*3]

Abstract

Artificial intelligence has enabled automatic decision making by leveraging big data. In a variety of domains where these AI-enabled decisions are made, there is increasing demand for *context-based* decision making at the individual level such as in personalized medicine and advertising. Often times, decisions need to be made by the AI agent *in the face of uncertainty* and with the challenge of the immense complexity of the environment. In this project, we set out to apply *reinforcement learning* to tackle the warfarin dosing problem. We discuss various types of *contextual bandit* algorithms and show that they can be applied to reinforcement learning models and effectively outperform two popular dosing algorithms actively used in the clinical setting. Furthermore, we show that simple *ensemble* methods can be applied to enhance the performance and most importantly mitigate the risk of severe mistakes.

1. Introduction

Warfarin is the most widely used oral anticoagulant agent worldwide; more than 30 million prescriptions were written for this drug in the United States in 2004 (Wysowski DK, 2007). Correctly dosing warfarin remains a significant challenge as the appropriate dosage is highly variable among individuals (by a factor of up to 10) due to patient clinical, demographic and genetic factors (Bastani & Bayati, 2015).

Physicians typically follow a fixed-dose strategy: they start patients on 5mg/day (the appropriate dose for the majority of patients) and slowly adjust the dose over the course of a few weeks by tracking the patients anticoagulation levels. An incorrect initial dosage can result in highly adverse consequences such as stroke (if the initial dose is too low) or internal bleeding (if the initial dose is too high). Every year, nearly 43,000 emergency department visits in the United States are due to adverse events associated with inappropriate warfarin dosing. (Budnitz, 2006).

Some previously proposed algorithms for predicting the appropriate dose of warfarin (Anderson JL, 2007; Caldwell MD, 2008; Millican EA, 2007; Sconce EA, 2005; AH., 2007) are usually based on relatively small clinical populations, and their general predictive accuracy is uncertain (Bussey HI, 2008). The International Warfarin Pharmacogenetics Consortium developed a pharmacogenetic dose algorithm for warfarin with the use of a large and diverse data set that included data from patients at centers around the world and used it to determine retrospectively whether the dosage recommendations that were based on this algorithm were significantly better than those that were based on an algorithm that used only clinical variables or those that were based on a fixed-dose strategy (Consortium, 2009).

In this project, we studied and implemented a range of contextual bandit algorithms to tackle the warfarin dosing challenge. We describe in this paper the algorithms we experimented and compare their results. We also discuss our findings based on our experiment results and suggest a few directions for future research.

2. Formulation & Related Work

2.1. Multi-armed Bandit Formulation

The problem of warfarin dose recommendation can be modeled as a **multi-armed bandit (MAB)** problem with context information. By factoring context information into the recommendation decision modeled as a MAB, we call it a *contextual bandit* (J.Langford & T.Zhang., 2007).

Formally, let T be the number of time steps, then:

Context: At each time step $t \in [1..T]$, a new patient arrives and we observe the *context information* (knowledge about the patient (e.g., gender, age, ...) in the form of individual feature vector $x_t \in \mathbb{R}^d$.

[*] Equal contribution [1] UHG (Unified Healthcare Group), Bellevue, Washington, USA [2] Google, Mountain View, California, USA [3] Intuit, Mountain View, California, USA. Correspondence to: Andriy Syrov <asyrov@stanford.edu>, Yun Zhou <21zhouyun@gmail.com>, Ming-Kuang Daniel Wu <danielwu@alumni.stanford.edu>.

Arms: The dose recommendation algorithm has access to $K = 3$ arms, each representing a bucket of the optimal dosages using the *clinically relevant* dosage differences suggested in (Consortium, 2009): (1) *Low*: under 3mg/day, (2) *Medium*: 3-7mg/day , and (3) *High*: over 7mg/day.

Reward: If the algorithm identifies the correct dosage for the patient, the reward is 0, otherwise a reward of -1 is received.

Objective: Design a MAB algorithm that learns a mapping $\pi : X \rightarrow a$ that yields the maximal expected reward. Let $\pi_t(x_t) \in [1..K]$ denote the arm chosen by policy π at time $t \in [1..T]$ for the patient feature vector x_t. Define the optimal policy π^* to be the policy that yields the maximal expected reward across all patients, and R_t^* is the reward incurred at time step t by following π^*. Then the expected *regret* incurred when the agent chooses arm i at time step t is $\mathbb{E}[R_t^* - R_t^i]$, where R_t^i is the reward received from choosing arm $i \in [1..K]$ at time step t. Therefore, our goal is to create and evaluate algorithms that minimize the ***total expected regret***: $\sum_{t=1}^{T} \mathbb{E}[R_t^* - R_t^i]$.

2.2. Related Work

The fundamental challenge in bandit problems is the need for balancing *exploration* and *exploitation*. To minimize the *regret*, an agent *exploits* its past experience to select the arm that appears best. However, this seemingly optimal arm may in fact be sub-optimal, due to the imprecise knowledge of the agent. To avoid being stuck in a sub-optimal situation, the agent needs to take a chance and *explore* seemingly sub-optimal arms to gather more information for better decision making. Inevitably, $exploration$ may increase short term regret but is essential for maximizing total rewards or minimizing total regrets. A good trade-off between *exploration* and *exploitation* is needed.

The *context-free MAB* problem has been studied extensively by statisticians (Berry & Fristedt, 1985; Robbins., 1952; Thompson., 1933). One of the simplest exploration algorithm is *ϵ-greedy* algorithm where each arm's payoff is first estimated. At each following step t, the agent chooses the arm with the highest payoff estimate (thus, *greedy*) with probability $1 - \epsilon$ (*exploitation*), and a random arm is chosen with probability ϵ (*exploration*). When each arm is tried infinitely often, the payoff estimates converge to true value of each arm which warrants less *exploration* and more *exploitation*. Therefore, it is sensible to decay ϵ appropriately over time (Robbins., 1952).

Instead of random exploration, *Upper confidence bound (UCB)* algorithms adopt a smarter strategy to balance exploration and exploitation (P. Auer & Fischer., 2002; Agrawal., 1995; Lai & Robbins., 1985). At each step t, the agent estimates both the mean payoff of each arm as well as a corresponding confidence interval so that with high probability the difference between the arm's estimated mean payoff and its true mean payoff falls within the confidence interval. The agent then select an arm that achieves the highest UCB.

In contrast to the well-studied *context-free* MAB algorithms, *contextual* MAB problems remain an active area of research. Auer considered the *contextual bandit* problem with linear payoffs (under the name *associative reinforcement learning with linear value functions*) and presented *LinRel* algorithm. Further improvements were made by subsequent UCB algorithm researchers. These algorithms use the idea of *optimism-in-the-face-of-uncertainty (OFU)*, which elegantly solves the exploration-exploitation trade-off by maintaining confidence sets for arm parameter estimates and choosing arms optimistically from within these confidence sets (Bastani & Bayati, 2015). Li et al. proposed a new, general *contextual bandit* algorithm, *LinUCB*, that is computationally efficient and well motivated from learning theory (Li et al., 2010). Chu et al. provided a theoretical analysis of a variant of *LinUCB* and proved an upper and a lower regret bounds for it (Chu, 2011).

Bastani and Bayati formulated learning a model of decision rewards conditional on individual-specific covariates as a multi-armed bandit with high-dimensional covariates. They presented a new efficient bandit algorithm based on the LASSO estimator and proved a new oracle inequality that guarantees the convergence of the LASSO estimator despite the non-i.i.d. data induced by the bandit policy (Bastani & Bayati, 2015). Furthermore, they illustrated the practical relevance of their algorithm by evaluating it on the warfarin dose problem that inspired our study.

Decision tree as suggested by (Elmachtoub et al., 2017) can also be used as computationally efficient approach. Decision trees are reliable and conceptually simple models. They do not need handcrafted features to function. which is important in our problem domain. They also provide high accuracy and have interpretable representation. Elmachtoub et al. presented decision tree based algorithms that approximate Thompson sampling by using bootstrapped data sets. They also provided an efficient decision tree heuristic algorithm of which variant we implemented and evaluated on Warfarin data set along with other algorithms.

3. Algorithms

In this section, we provide an overview of the various algorithms we studied. The experiment setup and results are described in section 5.

3.1. Baseline Algorithms

3.1.1. Fixed Dose

In current practice, a patient is typically prescribed an initial dose, the doctor then monitors how the patient responds to the dosage, and then adjusts the patients dosage. This interaction can proceed for several rounds before the best dosage is identified. Our first baseline algorithm adopts this strategy to assign a fixed *medium* dose to all patients.

Algorithm 1 Fixed Dose

```
for all $t = 1, 2, 3, ..., T$ do
    Choose arm $a_{Med}$
end for
```

3.1.2. Warfarin Clinical Dosing Algorithm

The *International Warfarin Pharmacogenetics Consortium* developed and used an algorithm for estimating the appropriate warfarin dose that is based on both clinical and genetic data from a broad population base (Consortium, 2009). This method is a linear model based on age, height, weight, race and medications that patient is taking. See appendix of (Consortium, 2009) for details on the construction of the feature vector $\mathbf{x_t}$.

Algorithm 2 Warfarin Clinical Dosing Algorithm

```
for $t = 1, 2, 3..T$ do
    Observe patient feature vector $\mathbf{x_t} \in \mathbb{R}^d$ where
        $\mathbf{x_t} = \begin{bmatrix} 1 \\ \text{age\_in\_decades} \\ \text{height\_in\_cm} \\ \text{weight\_in\_kg} \\ \text{is\_Asian} \\ \text{is\_Black} \\ \text{is\_race\_missing} \\ \text{enzyme\_inducer\_status} \\ \text{amiodarone\_status} \end{bmatrix}$
    Compute recommended daily dosage $d = (\mathbf{w}^T\mathbf{x_t})^2/7$ where $\mathbf{w}^T$ is the weight vector:
        $\mathbf{w}^T = [4.0376, -0.2546, 0.0118, 0.0134, -0.6752, 0.4060, 0.0443, 1.2799, -0.5695]$
    Based on definition in Section 2.1, map $d$ to one of the 3 arms $a_{Low}, a_{Med}, a_{High}$ and return the corresponding arm
end for
```

3.2. UCB-based Algorithm

The fundamental problem in bandits is the explore-exploit trade-off. To minimize regret, the algorithm is motivated to select the arm that appears the best. However, the current best arm might be sub-optimal due to limitation of the current knowledge. Therefore, the algorithm is also motivated to explore new arms. Many algorithms are proposed to balance exploration and exploitation. Upper Confidence Bound (UCB) (Auer, 2003) is a class of such algorithms. Intuitively, UCB algorithms are optimistic because it always choose the arm with the highest upper confidence bound. It can also be shown that with appropriately defined confidence interval, the UCB algorithms have regret that is only logarithmic to the total number of trials T, which is optimal (Lai & Robbins., 1985).

3.2.1. LinUCB with Disjoint Linear Models

LinUCB (Li et al., 2010) is a UCB algorithm where the upper confidence bound of an arm is modeled by a *ridge regression* model. Algorithm 3 gives a detailed description of the LinUCB algorithm.

Algorithm 3 LinUCB with disjoint linear models.

```
for $t = 1, 2, 3, ..., T$ do
    Observe features of all arms $a \in A_t : x_{t,a} \in R^d$
    for all $a \in A_t$ do
        if a is new then
            $A_a \leftarrow I_d$ (d-dimensional identity matrix)
            $b_a \leftarrow 0_d$ (d-dimensional zero vector)
        end if
        $\hat{\theta}_a \leftarrow A_a^{-1} b_a$
        $p_{t,a} \leftarrow \hat{\theta}_a^T x_{t,a} + \alpha\sqrt{x_{t,a}^T A_a^{-1} x_{t,a}}$
    end for
    Choose arm $a_t = argmax_{a \in A_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff $r_t$
    $A_{a_t} \leftarrow A_{a_t} + x_{t,a_t} x_{t,a_t}^T$
    $b_{a_t} \leftarrow b_{a_t} + r_t x_{t,a_t}$
end for
```

3.3. LASSO-based Algorithm

3.3.1. Lasso Bandit

LASSO Bandit (Bastani & Bayati, 2015) approaches the explore-exploit trade-off in a much simpler way. It employs a fixed schedule where each arm is forced periodically regardless of the current state of the algorithm. The gap between forced samples is exponential to the number of trial t. Therefore, the total number of forced sample is logarithmic to the total number of taisl T. Based on this schedule, it can be shown (Bastani & Bayati, 2015) that LASSO bandit has regret that is logarithmic to T.

From a high level, the algorithm maintains two set of parameters for each arm: one fitted on forced samples and the other fitted on all samples.

The algorithm also takes a *sampling parameter* $q \in Z^+$, a *localization parameter* $h > 0$, and initial regularization parameters $\lambda_1, \lambda_{2,0}$.

Forced-Sample Sets: A set of times when an arm i is forced (regardless of the observed feature X_t).

$T_i \equiv \{(2^n - 1)Kq + j | n \in \{0, 1, ...\}, j \in \{q(i-1)+1, ..., qi\}\}$

All-Sample Sets: We use $S_{i,t}$ to denote the set of times we arm i is played up to time t.

At any time t, the algorithm maintains two sets of parameters:

1. the forced-sample estimate $\hat{\beta}(T_{i,t-1}, \lambda_1)$ based on forced samples from arm i.

2. the all-sample estimate $\hat{\beta}(S_{i,t-1}, \lambda_2)$ based on all samples observed from arm i.

Algorithm 4 describes the LASSO bandit algorithm in detail.

Algorithm 4 LASSO Bandit

```
Input parameters $q, h, \lambda_1, \lambda_{2,0}$
Initialize $T_{i,0}$ and $S_{i,0}$ by the empty set, and $\hat{\beta}(T_{i,0}, \lambda_1)$
and $\hat{\beta}(S_{i,0}, \lambda_{2,0})$ by 0 in $R^d$ for all arms.
Use $q$ to construct force-sample sets $T_i$ for all arms.
for $t = 1, 2, 3..T$ do
    Observe $X_t$
    if $t \in T_i$ for any $i$ then
        $\pi_t \leftarrow i$
    else
        $\hat{K} = \{i \in [K] | X_t^T \hat{\beta}(T_{i,t-1}, \lambda_1) \leq \max_{j \in [K]} X_t^T \hat{\beta}(T_{j,t-1}, \lambda_1) - h/2\}$
        $\pi_t \leftarrow argmax_{i \in \hat{K}} X_t^T \hat{\beta}(S_{i,t-1}, \lambda_{2,t-1})$
    end if
    $S_{\pi_t,t} \leftarrow \S_{\pi_t,t-1} \cup t$
    $\lambda_{2,t} \leftarrow \lambda_{2,0}\sqrt{\frac{logt+logd}{t}}$
    Play arm $\pi_t$ and observe reward.
end for
```

3.4. Tree-based Algorithm

3.4.1. Decision Tree Bandit

Decision Trees have a number of advantages [1] for our problem domain such as they (a) are interpretable; (b) require less feature engineering, can learn non-linear features; (c) handle both continuous and binary data, outliers, categorical and missing features.

Below is the algorithm as proposed in (Elmachtoub et al., 2017):

S_0, F_0 - prior numbers of successes and failures for Beta. We set both to 1

$\hat{\theta}_a$- decision tree for current step t and arm a pair. These trees have boolean label type (success/failure);

$N_S(\hat{\theta}_a, x_t), N_F(\hat{\theta}_a, x_t)$ - functions returning number of successes and failures for passed tree $\hat{\theta}_a$ and feature vector x_t;

Algorithm 5 Decision Tree MAB Heuristic

```
for $t = 1, 2, 3..T$ do
    for all $a \in A$ do
        $\{$sample probability $a$ is successful for $x_t\}$
        $s_a =$Beta$(N_S(\hat{\theta}_a, x_t) + S_0, N_F(\hat{\theta}_a, x_t) + F_0)$
    end for
    $\{$choose $a$ with highest probability of success$\}$
    $a_t = \arg\max_a(s_a)$
    $\{$test if dosage $a_t$ correct for $x_t\}$
    $r_t =$Correct-Dose$(x_t, a_t)$
    $\{$update tree for arm $a_t$ with new data point$\}$
    Refit-Tree$(\hat{\theta}_{a_t}, (x_t, r_t))$
end for
```

A few notes: (a) the algorithm uses Correct-Dose function that returns boolean value if selected dosage is correct (i.e. reward structure is not used); (b) sklearn implementation of decision tree was used instead of CART (proposed in (Elmachtoub et al., 2017)), with the parameter settings proposed in (Fraj, 2017);

3.5. Ensemble Method

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. An ensemble usually produces more accurate solutions than a single model would. One of the simplest ensemble method is *Majority Voting* where every model makes a prediction (votes) and the final output prediction is the one that receives more than half of the votes.

[1] For completeness we also need to mention the disadvantages, such as (a) need for efficient incremental implementation; (b) possible duplication of sub-trees and (c) some difficulty in learning functions.

3.5.1. Ensemble by Majority Vote

The top 3 performers from our experiments are chosen to form a *voting committee*: *LinUCBDisjoint*, *Lasso*, and *DTree-Alt*. We apply ***majority voting*** method to choose the arm. In the event of an even split between the three arms, we tie break using the *Fixed Dose* algorithm (i.e., choose the *Medium Dose* arm.) The fact that the *voting committee* fails to reach a majority implies the higher uncertainty in the decision. Therefore, it makes sense to fall back to choosing the *Medium Dose* to avoid making severe dosing mistakes (choosing *Low* when the true dosage is *High* or vice versa.)

Algorithm 6 Ensemble by Majority Vote

```
for $t = 1, 2, 3..T$ do
    Use LinUCBDisjoint to choose arm $a_{LinUCB}$
    Use LASSO Bandit to choose arm $a_{LASSO}$
    Use Tree-Alt to choose arm $a_{DTree-Alt}$
    if $a_{LinUCB} \neq a_{LASSO} \neq a_{DTree-Alt}$ then
        Choose arm $a_{MED}$
    else
        Choose arm $a_{Majority}$ where $a_{Majority}$ equals the majority of the 3 arms returned by the above algorithms
    end if
end for
```

4. Evaluation Methodology

In the standard *supervised learning* setting, the training and evaluation of the model can be performed *offline* with labeled dataset. In contrast, for a *contextual bandit* setting, we would like to measure the performance of a *bandit* algorithm for selecting an arm at each time step based on the preceding interactions. Because of the interactive nature of the problem, we need to simulate this *online* setting to evaluate our algorithms. Therefore, even though the the warfarin dosing dataset contains the ground truth label (*Therapeutic Dose of Warfarin*), we suppress this counterfactual information to the bandit algorithms, thereby keeping an optimal algorithm unknown. This lets us benchmark the performance of our algorithms in an unbiased manner.

In order to make sure any performance gain or loss over baselines is not because of ordering of the patients, we run all algorithms multiple iterations on different random permutations of the patients. Furthermore, we ensure that all algorithms are trained and evaluated on the same permutations of the patients with the same number of iterations so the results can be analyzed fairly. We tested iteration numbers from 3 to 1000 and saw no significant differences in the evaluation results once we run 10 or above iterations. Therefore, we decided to test all our algorithms with the same 10 random permutations of the data set.

5. Experiments

5.1. Experiment Setup

This subsection provides details for our experiment setup including the data set we use, feature engineering, performance metric, and the competing algorithms.

5.1.1. Dataset

We use a publicly available patient dataset that was collected by staff at the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) for 5700 patients who were treated with warfarin from 21 research groups spanning 9 countries and 4 continents. Importantly, this data contains the true patient-specific optimal warfarin doses (which are initially unknown but are eventually found through the physician-guided dose adjustment process over the course of a few weeks) for 5528 patients. It also includes patient-level contextual information such as clinical factors, demographic variables, and genetic information that have been found to be predictive of the optimal warfarin dosage (Consortium, 2009). There are missing values in the dataset, so we further filter out records that do not have values for all three required features we need, namely: *Age*, *Height (cm)*, and *Weight (kg)*. This reduces our dataset size to 4386 patients.

5.1.2. Feature Engineering

Our dataset consists of 63 titled columns covering the following patient contextual information (Consortium, 2009):

- *Demographics*: gender, race, ethnicity, age, height, weight
- *Diagnosis*: reason for treatment (e.g. deep vein thrombosis, pulmonary embolism, etc.)
- *Pre-existing diagnoses*: indicators for diabetes, congestive heart failure or cardiomyopathy, valve replacement, smoker status
- *Medications*: indicators for potentially interacting drugs (e.g. aspirin, Tylenol, Zocor, etc.)
- *Genetics*: presence of genotype variants of CYP2C9 and VKORC1

After filtering out patient records with missing values in the required columns as described in 5.1.1., we first perform imputation of missing value of *VKORC1 SNPs* following the instructions given in the appendix Section S4 of (Consortium, 2009). We use both the *Medications* column that contains a list of medication names and individual binary-valued medication columns to construct features covering patient's medications. For all categorical data, we perform *One-Hot encoding* and treat unknown and missing values as an extra possible feature value. For numeric data (*i.e., Height, Weight, Target INR, INR on Reported Therapeutic Dose of Warfarin*), we perform *MinMax scaling* to ensure the data is scaled to a fixed 0 to 1 range. Table 1 summarizes the number of features we use in each algorithm.

Algorithm	Feature Count	Algorithm	Feature Count
Fixed Dose	0	Clinical Dose*	8
LinUCBDisjointBasic*	8	LinUCBDisjoint*	133
DTree	8	DTree-Alt	20
Lasso*	133	**also adds a bias term*	

Table 1. Feature Count by Algorithm

5.1.3. Performance Metrics

We apply the following metrics to evaluate the algorithm performance:

- **Total Regret**: As defined in section 2.1, *regret* is the opportunity loss for one step in the bandit formulation. For each dosing decision made in iteration i for patient t, we denote the *regret* as l_t^i. When a correct dose is recommended, $l_t^i = 0$, otherwise $l_t^i = 1$. Therefore, we compute *total regret* $L_t = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} l_t^i$ where n is the number of iterations we run the algorithm.
- **Average Fraction of Incorrect Dosing Decisions**: This quantity is computed cumulatively at each step of an iteration. Specifically, at step t, we calculate the average faction of mistakes made from the start of the iteration up to patient t. We then average this quantity over all iterations.

5.1.4. Compared Algorithms

The algorithms that are empirically evaluated in our experiments can be categorized into two groups.

Context-free Algorithm: One of our baseline algorithm is *FixedDose* (3.1.1), which always recommends medium dosage without taking into consideration the contextual patient information. This is also widely practiced in the clinical setting to avoid the risk of severe dosing errors.

Contextual Algorithms: All the other algorithms we studied take advantage of the contextual patient information.

- *ClinicalDose*: This is the second baseline we studied as described in 3.1.2. It is a linear model using 8 patient features plus 1 bias term to predict the appropriate dosage.
- *LinUCBDisjointBasic* & *LinUCBDisjoint*: These are linear UCB based algorithm as described in 3.2.1. *LinUCBDisjointBasic* uses the same set of 8 features as *Clinical Dose* while *LinUCBDisjoint* uses 133 patient features. In both variations, we add a bias term and initialize its parameter $\alpha = 0.01$.
- *DTree* & *DTree-Alt*: These are *decision tree* based implementations as described in Algorithm 5. *DTree* uses the same set of 8 features as *Clinical Dose* while *DTree-Alt* uses 20 patient features. In both variations, we initialize: $tree\_depth = 4$, $min\_samples\_split = 37$, $min\_samples\_leaf = 11$, $max\_leaf\_nodes = 4$, and $criterion = "gini"$.
- *LASSO*: An implementation based on the *LASSO Bandit* algorithm described in 3.3.1. It uses same set of 133 patient features used by *LinUCBDisjoint*. We initialize: $q = 1, n = 10, h = 5, \lambda_1 = 0.05, \lambda_2 = 0.05$
- *Majority3*: This implements the majority voting ensemble algorithm described in 3.5.1. The voting committee consists of our top 3 performers: *LASSO*, *LinUCBDisjoint*, and *DTree-Alt*.

5.2. Experiment Results

As described in Section 4, all algorithms were executed on 10 randomly shuffled warfarin dose data set and the results are averaged. We ensure that each algorithm is trained and evaluated on the same permutations for a fair comparison. Table 2 shows our results, and the visualization of our results are in Figures 1 and 2

Algorithm	Average Fraction of Incorrect Decisions (lower better)	Total Regret (lower better)
Fixed Dose (baseline)	0.3842	1685.0
DTree	0.3728	1635.0
LinUCBDisjointBasic	0.3719	1631.1
Clinical Dose (baseline)	0.3484	1528.0
DTree-Alt	0.3468	1521.0
LinUCBDisjoint	0.3384	1484.2
Lasso Bandit	0.3337	1463.6
Majority3 Ensemble	**0.3275**	**1436.2**

Table 2. Results of All Algorithm Runs

Figure 1 shows the cumulative average fraction of incorrect decisions (i.e. incorrect dosage) as function of number of patients seen. For better legibility, the full size version of the figure is reproduced in the Appendix (Figure 10). Figure 3 shows close-up views of the same plot broken into four sections along the x-axis, which we further discuss in the Size of Data section below. Figure 4 is the same plot with the error band showing the *standard deviation* across all iterations. At the end, our ensemble algorithm *Majority3* outperforms the *FixedDose* baseline by almost 6 percentage points, the *ClinicalDose* baseline by over 2

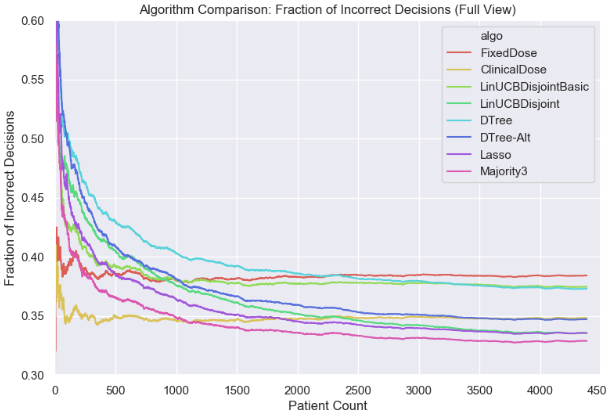

Figure 1. Average Fraction of Incorrect Dosing Decisions

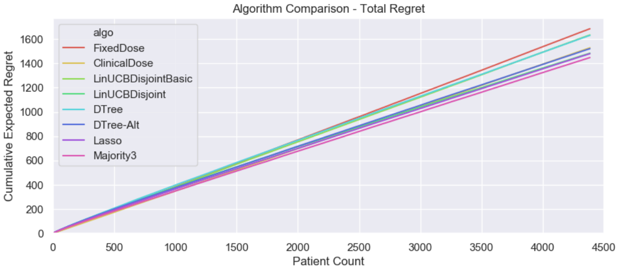

Figure 2. Total Regret

percentage points, and its constituents by up to 2 percentage points. Similarly, Figure 2 shows a complete view of the total regret as function of patients seen (full size version in the Appendix Figure 11) and Figure 5 shows its associated close-up views.

5.2.1. On Learning Speed

Our top 4 performers are clearly fast learners from the start. Against *FixedDose* baseline: *Majority3* surpasses it around $t = 180$, *Lasso* around 500, *LinUCBDisjoint* around 900, and *DTree-Alt* around 1000. Against *ClinicalDose* baseline: *Majority3* surpasses it around 1100, *Lasso* around 1850, *LinUCBDisjoint* around 2400, and *DTree-Alt* around 3500. Clearly *Majority3* benefits from the complimentary strengths of its constituents.

5.2.2. On Use of Features

Our results confirm our hypothesis that *contextual information* is very valuable in the *MAB* setting as it provides clues for the *contextual bandit* algorithms to learn good policies. Looking at the end game, it is clear that all our *contextual* algorithms outperform the only *context-free* algorithm (*FixedDose*.) In addition, more relevant features in terms of quality and quantity are beneficial. We purposefully design our experiments to form pairs of same algorithms with small and larger feature sets: *LinUCBDisjointBasic (8)* vs *LinUCBDisjoint (133)* and *DTree (8)* vs *DTree-Alt (20)*. The performance gap between *DTree (8)* and *DTree-Alt (20)* is around 2.5 percentage points; we observed even bigger gap between *LinUCBDisjointBasic (8)* and *LinUCBDisjoint (133)* by almost 4 percentage points!

Another interesting observation is that the *ClinicalDose* (Consortium, 2009) baseline **??**definitely benefits from the supervised learning setting thus only requires 8 pre-engineered features to have its average fraction of incorrect dosing decision reach below 35%. However, our *online reinforcement learning* setting is closer to the true clinical setting where the ground truth of medication dosage is unknown beforehand. We would like to highlight this as indeed one of the core challenges for reinforcement learning.

5.2.3. On Size of Data

Figure 3 shows that more data helps learning, especially in the ramp-up phase of the learning. In our experiments, learning progresses the fastest during the first 1000 patients and starts to flatten out afterwards. Our fastest learning models are able to learn well even with less than 10% of the data set. So the size of data definitely matters to a point where the model's expressiveness and the effectiveness of the feature set become the more determining factors in model performance.

5.2.4. On Risk Sensitivity

One of the key concerns about applying artificial intelligence in medicine is the risk factor. Serious medical decision mistakes can cause injuries or even deaths. Therefore, we need to analyze the types of errors our algorithms make and seek mitigation for such risks.

Figure 6 shows the distribution of all the dosing decisions made, grouped by algorithm. The diagonal (circled in green) shows correct dosing decisions while the cells circled in red represent serious dosing mistakes where the dose is off by 2 buckets. It can be seen that *LinUCBDisjointBasic* has the lowest rate of serious mistakes, which is not a surprise due to the nature of UCB based algorithms. To further analyze the types of mistakes made by each algorithm, we construct Figure 7.

There are two interesting observations here: (1) even though *LinUCBDisjoint* with 133 features outperforms *LinUCBDisjointBasic* with 8 features by a good margin, *LinUCBDisjointBasic* actually makes fewer serious mistakes (*severe overdose* or *severe underdose*). This reminds us of *safe reinforcement learning* where sometimes trading off overall

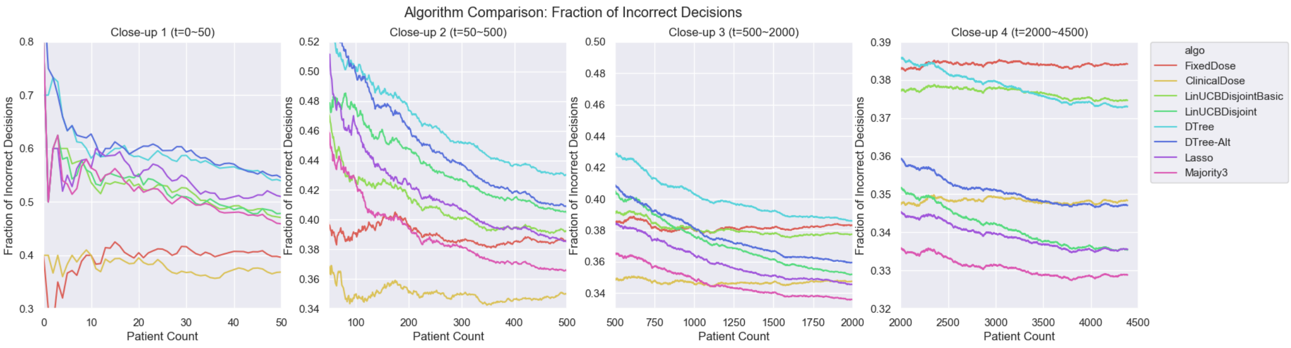

Figure 3. Close-up View of Average Fraction of Incorrect Dosing Decisions w/o Error Band

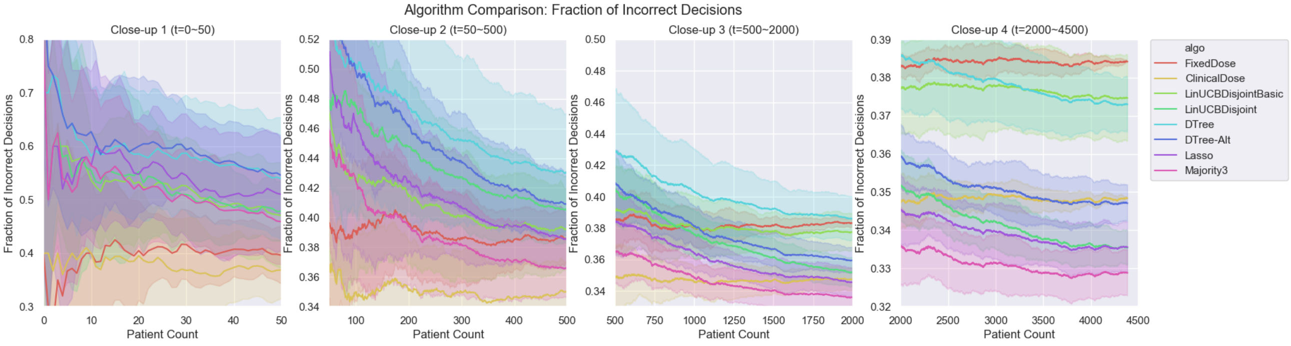

Figure 4. Close-up View of Average Fraction of Incorrect Dosing Decisions w/ Error Band

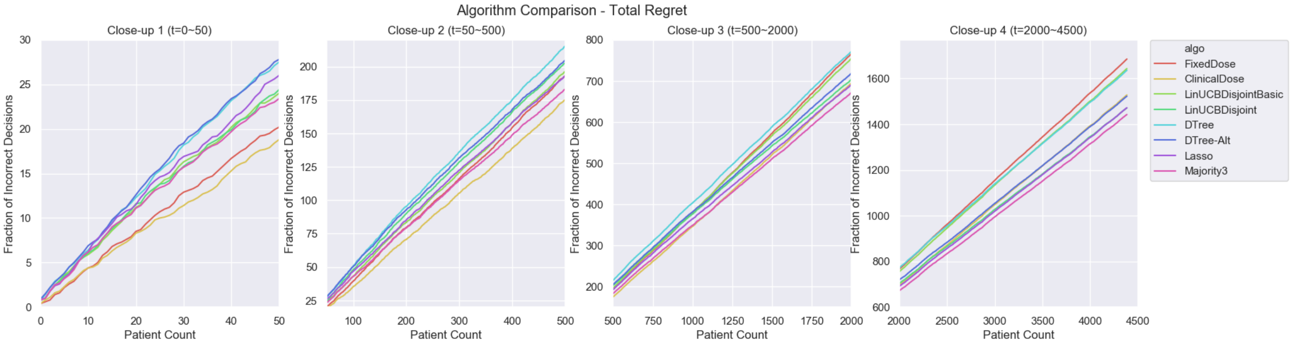

Figure 5. Close-up View of Total Regret

performance for fewer serious mistakes may be necessary. (2) A closer look at the results reveals that our ensemble *Majority3* make fewer severe mistakes when compared to its constituents, which is very interesting because it not only achieves best overall performance but also makes the least amount of severe mistakes. We think this is no coincidence. Intuitively ensemble hedges the risks by incorporating more diversified information into its decision making. This can be a promising direction for future research.

5.2.5. ON ENSEMBLE

Our ensemble model definitely benefits from its constituents in achieving superior overall performance and reducing the risk of making serious mistakes. Figure 8 shows the types of mistakes made by the ensemble and its constituents. Full size version is in Appendix (Figure 14.) The cells circled in green provide further insights into the types of mistakes that are prevented by the ensemble while the cells circled in red

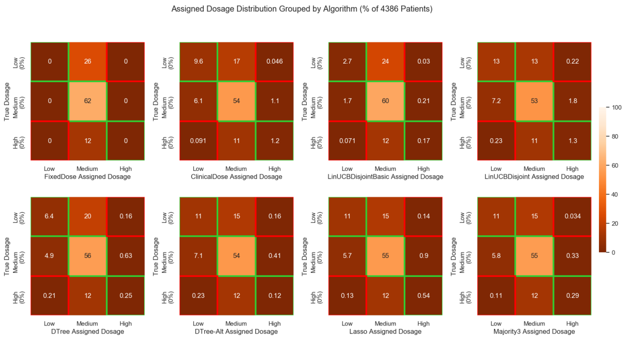

Figure 6. Assigned Dosage Distribution Grouped by Algorithm

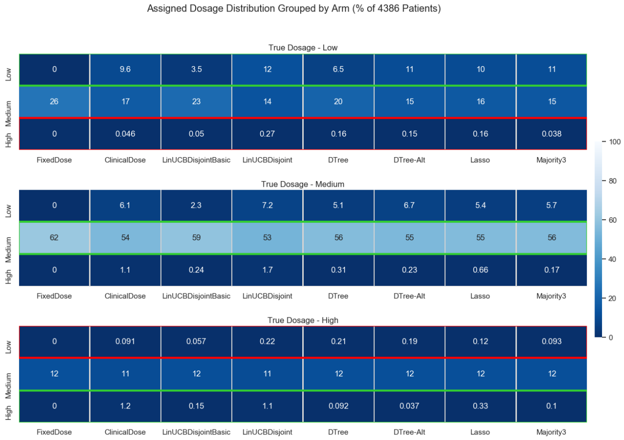

Figure 7. Assigned Dosage Distribution Grouped by Arm

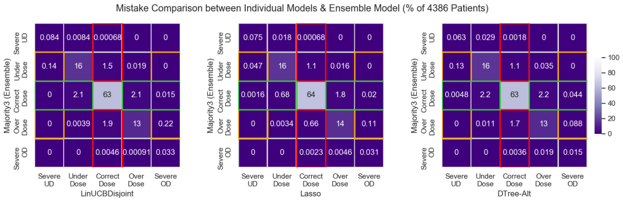

Figure 8. Assigned Dosage Distribution Grouped by Arm

highlight where the constituents are correct but the ensemble makes mistakes, thus, areas for future improvements on the ensemble. Furthermore, the cells circled in orange identify the *risk reduction* potential of our ensemble in which severe mistakes could be mitigated down to less severe mistakes by our ensemble. Overall, this helps explain how our ensemble has helped and highlight the areas for future improvements.

5.2.6. On Interpretability

One common criticism against applying artificial intelligence in general and especially to medicine is its lack of interpretability. Interpretability provides transparency for how AI models reach their conclusions and offers us confidence in the decisions they make and most importantly, peace of mind. For this reason, we explored decision tree based bandit algorithm. Figure 9 is an illustration of such interpretability. It shows the decision tree learned for sampling probability of selecting low dosage arm: if patient is Asian and age group is greater than 5 (60+ years) and weight is less than 63kg, then sample for the probability that *the correct dosage is LOW* from $Beta(215, 113)$.

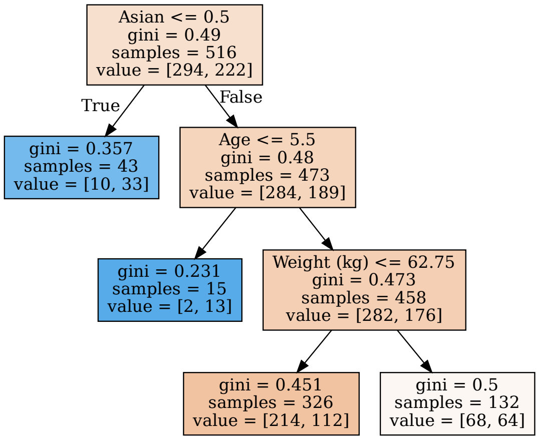

Figure 9. Decision Tree for sampling probability of low dosage arm

6. Conclusion

In this project, we applied *contextual bandit* algorithms to the warfarin dosing problem. We studied and compared linear UCB-based algorithm, LASSO-based algorithm, and the decision tree-based algorithm for better interpretability. We observed that feature set, data size, and the nature of the algorithms all contribute to the overall performance of the algorithms. Our best models outperform the two baselines we measure against. We provided detailed analysis on our experiment results; we also showed that ensemble method can both improve the overall performance and potentially reduce risk for making severe mistakes.

7. Future

Even though we have done extensive study on several *contextual bandit* algorithms, we barely scratched the surface of this research frontier. Considering that contextual features are crucial for these algorithms, we would suggest more studies to be done on better feature engineering and feature selection for *contextual bandit*. Another direction is to focus on *safe reinforcement learning* techniques in providing more safeguards in AI-enabled medical decision making. Finally, we think it would be worthwhile to investigate other ensemble techniques in the *contextual bandit* setting to gain better understanding on their associated benefits and challenges.

Acknowledgements

We gratefully acknowledge Prof Emma Brunskill and her CS234 course assistants (CA's) at Stanford University for their devotion to optimizing our learning experience and their inspiring passion in reinforcement learning research. Our project has also benefited from valuable advice from Andrea Zanette, our project mentor, and Ramtin Keramati, the head CA. They have been instrumental in guiding us to improve our work.

References

Agrawal., R. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

AH., W. Use of genetic and nongenetic factors in warfarin dosing algorithms. *Pharmacogenomics*, 8:851–861, 2007.

Anderson JL, Horne BD, S. S. e. a. Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. *Circulation*, 116: 2563–2570, 2007.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003.

Bastani, H. and Bayati, M. Online decision-making with high-dimensional covariates, 2015.

Berry, D. A. and Fristedt, B. (eds.). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

Budnitz, DS, D. P. K. W. A. M. T. S. J. A. National surveillance of emergency department visits for outpatient adverse drug events. *Journal of the American Medical Association*, 296:1858–1866, 2006.

Bussey HI, Wittkowsky AK, H. E. W. M. Genetic testing for warfarin dosing? not yet ready for prime time. *Pharmacotherapy*, 28:141–3, 2008.

Caldwell MD, Awad T, J. J. e. a. Cyp4f2 genetic variant alters required warfarin dose. *Blood*, 111:4106–4112, 2008.

Chu, Wei, L. L. L. R. R. S. Contextual bandits with linear payoff functions. volume 15, pp. 208214, 2011.

Consortium, I. W. P. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

Elmachtoub, A. N., McNellis, R., Oh, S., and Petrik, M. A practical method for solving contextual bandit problems using decision trees. *CoRR*, abs/1706.04687, 2017. URL http://arxiv.org/abs/1706.04687.

Fraj, M. B. Indepth: Parameter tuning for decision tree, 2017. URL https://medium.com/@mohtedibf/indepth-parameter-tuning-for-decision-tree-6753118a03c3.

J.Langford and T.Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *NIPS'07 Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 817–824, 2007.

Lai, T. L. and Robbins., H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. 2010. doi: 10.1145/1772690.1772758.

Millican EA, Lenzini PA, M. P. e. a. Genetic-based dosing in orthopedic patients beginning warfarin therapy. *Blood*, 110:1511–1515, 2007.

P. Auer, N. C.-B. and Fischer., P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3): 235–256, 2002.

Robbins., H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Sconce EA, Khan TI, W. H. e. a. The impact of cyp2c9 and vkorc1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood*, 106:2329–2333, 2005.

Thompson., W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Wysowski DK, Nourjah P, S. L. Bleeding complications with warfarin use: a prevalent adverse effect resulting in regulatory action. *Arch Intern Med*, 167:1414–1419, 2007.

A. Appendix

Here we include the full size version of the figures as supplemental materials for better legibility.

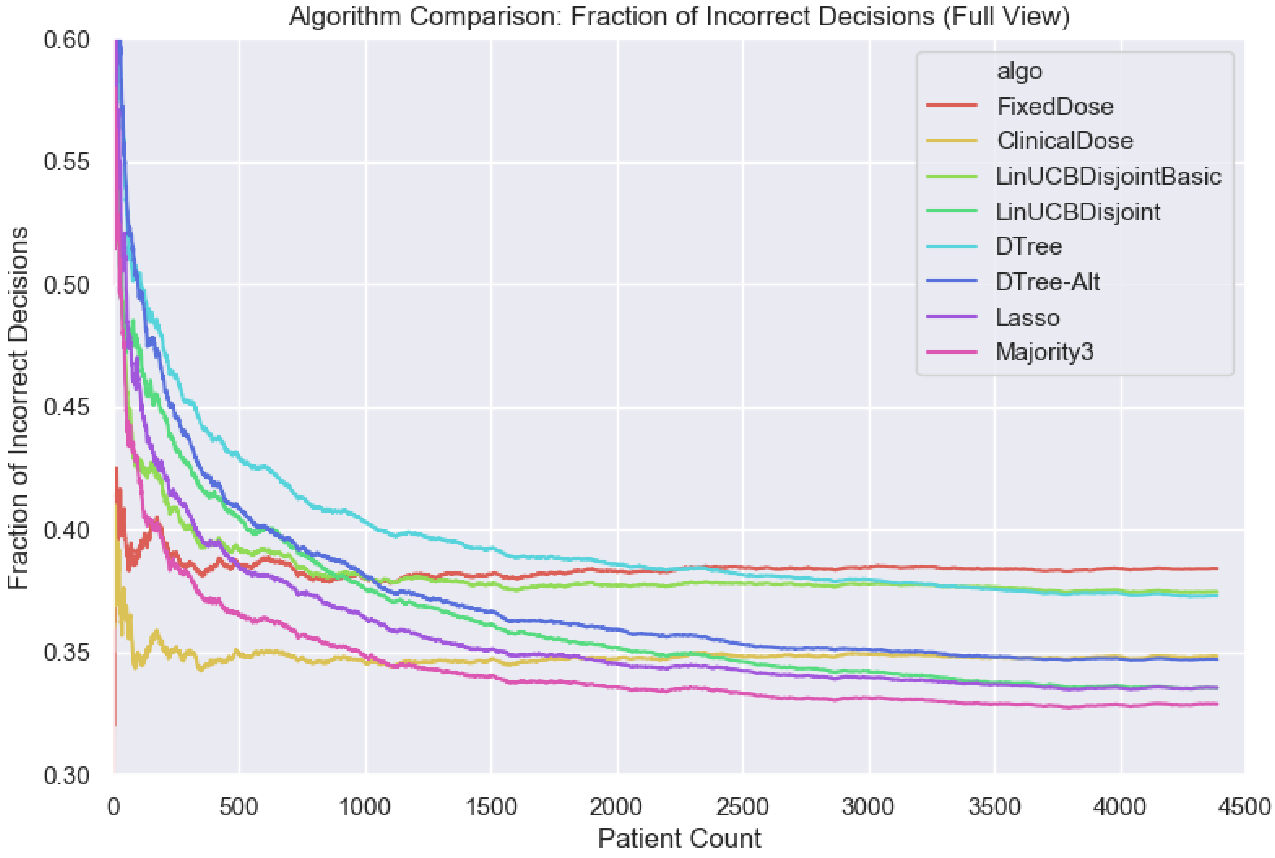

Figure 10. Average Fraction of Incorrect Dosing Decisions

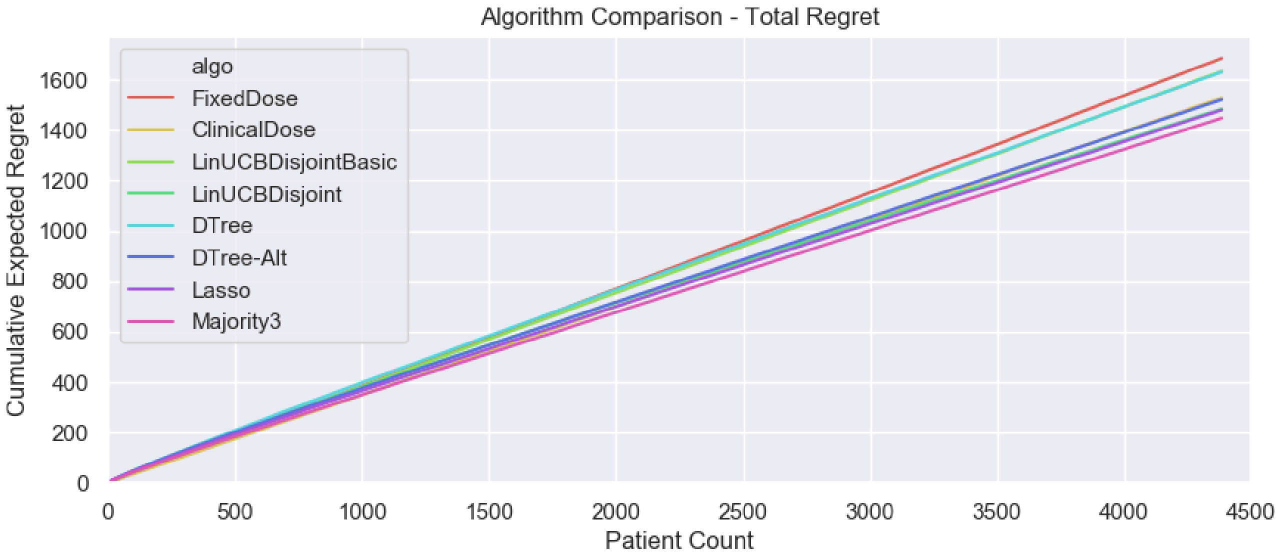

Figure 11. Total Regret

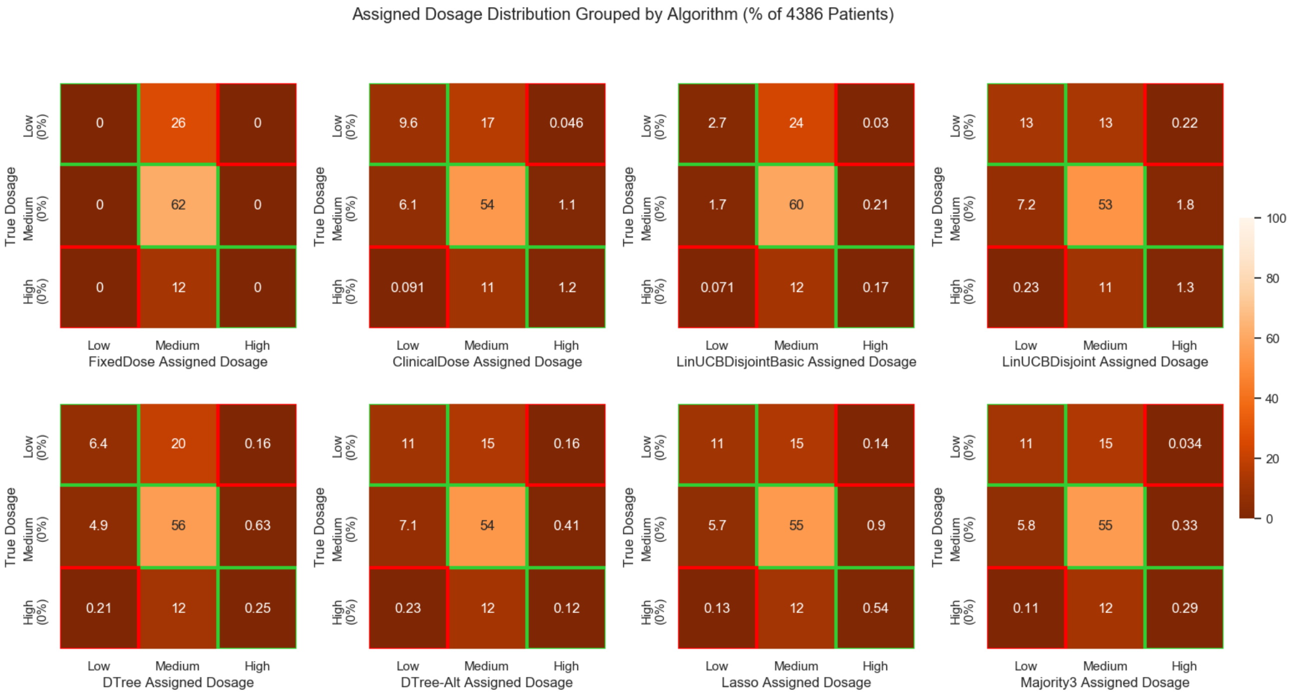

Figure 12. Assigned Dosage Distribution Grouped by Algorithm

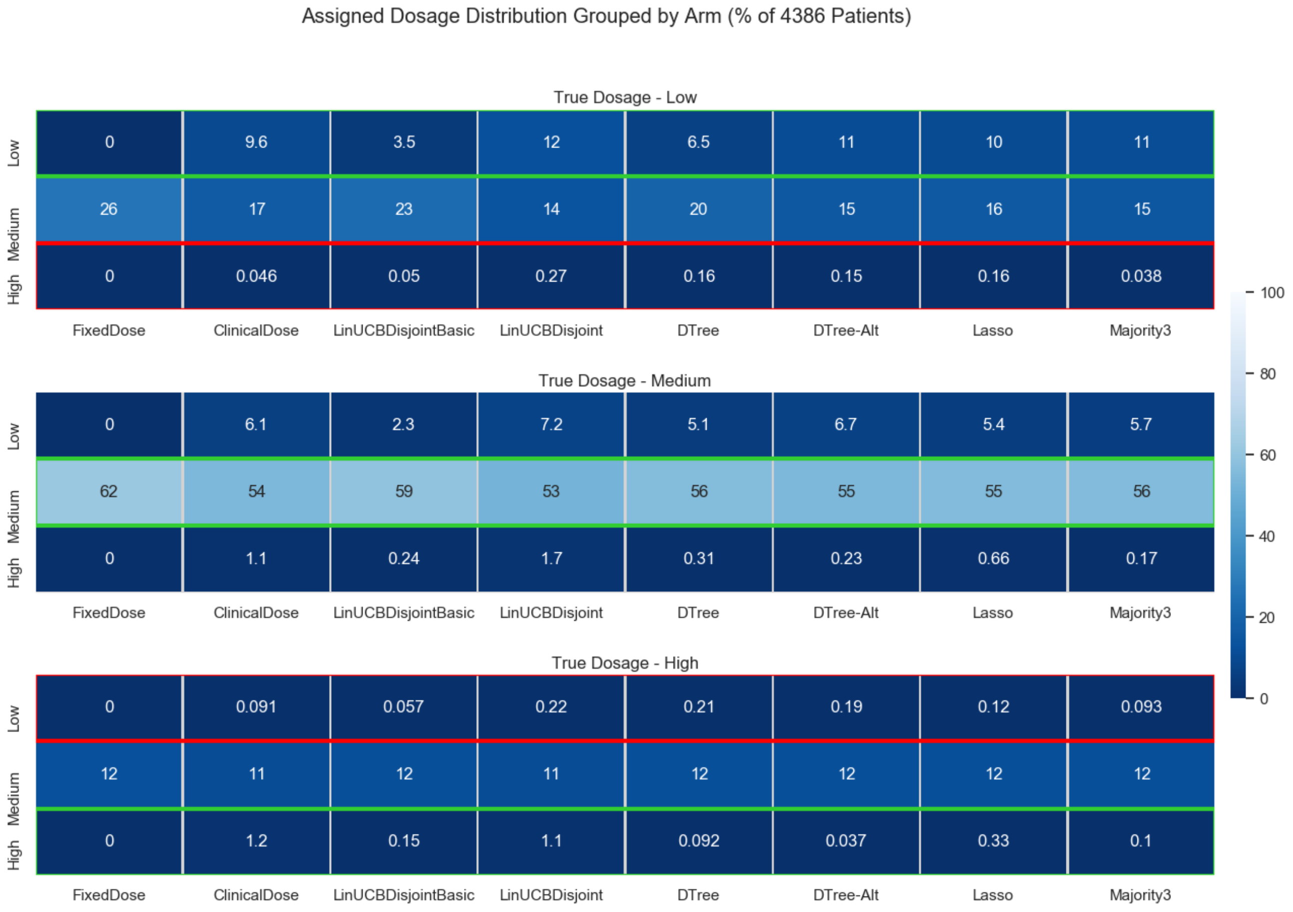

Figure 13. Assigned Dosage Distribution Grouped by Arm

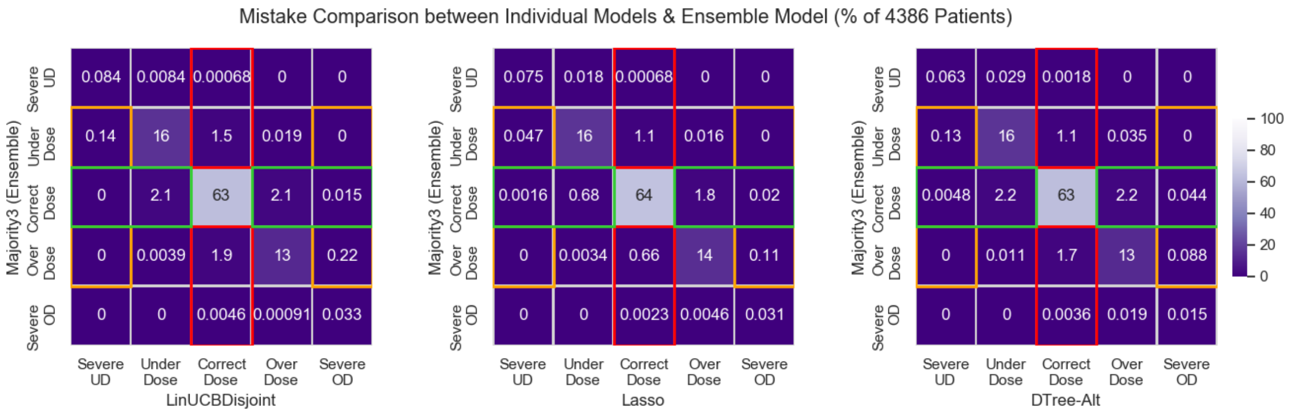

Figure 14. Assigned Dosage Distribution Grouped by Arm