Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

MODEL-BASED Q-LEARNING WITH MONOTONE POLICIES FOR PERSONALIZED MANAGEMENT OF HYPERTENSION

Wesley J. Marrero[1], and Lan Yi[1]

[1]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

ABSTRACT

Hypertension is a crucial controllable risk factor of atherosclerotic cardiovascular disease, a leading cause of death in the United States. While traditional analytic techniques may capture the complexities of hypertension treatment planning, they generally provide unintuitive treatment recommendations. This paper aims to advance the acceptance of analytic techniques in clinical practice by presenting a method to obtain interpretable treatment plans. To this end, we introduce the monotone Q-learning algorithm, which guarantees policies are nondecreasing on patients' health severity by limiting the exploration of treatment choices and solving simple integer programs. We represent a set of clinically representative patient profiles through Markov decision process models and compare the performance of our approximately optimal monotone policies with the optimal policy, optimal monotone policy, and current clinical guidelines. The approximately optimal monotone policies outperform the current clinical guidelines while displaying small losses in quality-adjusted life years compared to the optimal policy.

1 INTRODUCTION

Atherosclerotic cardiovascular disease (ASCVD), mainly manifested through heart attacks and strokes, is a leading cause of death in the United States (Kochanek et al. 2024). National statistics report that coronary heart disease, primarily expressed as heart attacks, and stroke account for 40.3% and 17.5% of deaths attributable to cardiovascular diseases, respectively (Martin et al. 2024). Hypertension or high blood pressure (BP) is a vital controllable risk factor of ASCVD, which affects 46.7% of adults in the United States (Martin et al. 2024). Effective hypertension management is key to reducing adverse ASCVD events.

Hypertension treatment guidelines play a critical role in BP management (Whelton et al. 2018). These guidelines are often developed based on the judgment of experts aiming to synthesize the latest research and clinical evidence. Nevertheless, expert-designed guidelines may not fully capture all the risks, benefits, and uncertainties inherent to treatment planning. Moreover, they may be considered subjective and controversial (Cohen and Townsend 2018; Solberg and Miller 2018; Ioannidis 2018). Conversely, analytic techniques may better capture the complexities of treatment planning and have outperformed clinical guidelines in simulation studies (Steimle et al. 2021; Bonifonte et al. 2022). Unfortunately, these techniques may generate complicated or unintuitive treatment recommendations (Lakkaraju and Rudin 2017). Despite their potential effectiveness in reducing ASCVD events, a lack of interpretability can limit the usability and acceptance of analytic techniques in clinical practice (Sethi et al. 2020; Wang et al. 2020).

In this paper, we aim to advance the acceptability of analytic techniques in clinical practice by presenting a method to obtain interpretable treatment plans. We consider the setting when patients' health dynamics are unknown or highly complex but can be estimated through simulation. This setting considers the problem of finding the best treatment plan among alternatives amenable to human intuition and cognition, based on patients' simulated health trajectories and treatment effects. Based on the input from our clinical collaborators, we define a hypertension treatment plan as interpretable if it is monotone (i.e., it does not decrease in intensity as a patient gets older or as their health worsens).

979-8-3315-3420-2/24/$31.00 ©2024

Due to their wide adoption in the healthcare literature, we focus on interpretable treatment plans obtained by solving Markov decision process (MDP) models (Puterman 2014). The literature on MDP models in healthcare applications is abundant. Recent examples in cardiovascular disease include research by Steimle et al. (2021), Marrero et al. (2021), Bonifonte et al. (2022), and Garcia et al. (2024). Most of these studies show that treatment plans attained through MDP models satisfying certain conditions are intrinsically interpretable. Unfortunately, the sufficient conditions that guarantee an interpretable optimal policy are often violated to some degree in clinical practice. Closer to our work, Garcia et al. (2024) use mixed integer programming to obtain monotone policies for hypertension treatment given an exact representation of patients' health evolution. We build upon this prior work by presenting a method to design monotone treatment plans within the model-based reinforcement learning field (Shakya et al. 2023).

Although the literature on interpretable reinforcement learning is vast (Glanois et al. 2024), there has been limited research on monotone policies as a definition for interpretability. Most work on monotone policies in reinforcement learning focuses on threshold learning. Within the context of healthcare applications, Hu et al. (2018) propose an approach to find health condition thresholds for mobile physical activity recommendations using Q-learning with function approximation (Murphy 2005). Bertsimas et al. (2022) present optimal policy trees for personalized threshold learning in mobile health behavioral interventions. Outside of healthcare applications, Roy et al. (2019) and Roy et al. (2022) consider the problem of structure-aware online learning and show the asymptotic optimality of their threshold policies. Liu and Mitra (2020) modify the standard Q-learning algorithm (Watkins and Dayan 1992) by estimating the optimal policy in a subset of the original problem given the knowledge that it has a threshold structure. In contrast to these studies, our proposed approach extends past learning a threshold among binary alternatives into monotonicity across an arbitrary finite number of treatment options.

Beyond thresholds among binary choices, Ngo and Krishnamurthy (2010) provide conditions that guarantee an optimal policy is a mixture of two monotonically increasing threshold policies. Djonin and Krishnamurthy (2007) and Fu and Van Der Schaar (2012) exploit the monotonicity exhibited by optimal solutions to derive learning algorithms that preserve this structural property. Furthermore, Roy et al. (2020) harness the known monotonicity of an optimal policy to enhance the convergence speed of a learning algorithm. While these studies focus on finding monotone policies given prior knowledge of the problem, we center on learning the best monotone policy despite no known structural properties.

1.1 Contributions

This research extends the existing work on monotone policies (Garcia et al. 2024) by exploring cases where patients' health progression is not fully known and can be estimated through simulation. Our approach builds upon the standard Q-learning algorithm (Watkins and Dayan 1992) and constrained optimization. The primary contributions of this research are summarized as follows:

1. **A new model-based Q-learning algorithm that guarantees monotone policies in each iteration, which we will refer to as monotone Q-learning.** This algorithm restricts the exploration of standard Q-learning (Watkins and Dayan 1992) by leveraging the requirement that the intensity of treatment must not decrease as a patient's health worsens. In addition, it ensures that the approximately optimal treatment plan obtained is monotone by solving a simple integer program.
2. **Application of our monotone Q-learning method to personalized hypertension treatment planning.** Using a set of clinically representative patient profiles, we provide interpretable decision support that is personalized to each patient's characteristics.
3. **Comparison of our monotone Q-learning to methods in the literature within the context of hypertension treatment planning.** We compare the performance of our proposed method to the optimal and optimal monotone policies (Garcia et al. 2024) as well as the most recent clinical guidelines (Whelton et al. 2018).

1.2 Organization of the Paper

The remainder of this paper is organized as follows. We begin by presenting our modeling framework and defining the monotone treatment planning problem in Section 2. In Section 3, we introduce our monotone Q-learning algorithm. Section 4 presents our numerical analysis applying the monotone Q-learning method for the management of hypertension. Lastly, Section 5 discusses conclusions and future research directions.

2 MODELING FRAMEWORK

In this section, we describe our MDP formulation for hypertension treatment planning. A patient's health is modeled through a finite number of states $\mathscr{S} = \{1, \ldots, S\}$ comprising their demographic information, clinical observations, and overall health condition to account for their history of cardiovascular events. At each year $t \in \mathscr{T} = \{0, 1, \ldots, T\}$, a decision-maker observes the state $s_t \in \mathscr{S}$ and prescribes a finite number of antihypertensive medications $a_t \in \mathscr{A} = \{1, \ldots, A\}$. Once treatment a_t is prescribed in state s_t at time t, the health of the patient evolves to a new state s_{t+1} according to a transition function $s_{t+1} = f_{t+1}(s_t, a_t, \omega_t)$, where $\omega_t \sim \mathscr{U}(0,1)$ is an independent and uniformly distributed random disturbance in $(0,1)$ representing the uncertainty in the state transition. The transition function is derived from patients' risk for ASCVD events, the benefit from treatment, ASCVD mortality, and non-ASCVD mortality. After transitioning to state s_{t+1}, the patient receives a finite reward $r_t(s_{t+1}, a_t, \omega_t)$ defined as the quality of life weight associated with state s_{t+1} minus the side effects from treatment a_t. The terminal reward $r_T(s_{T+1}, \omega_T)$ represents the patients' expected quality-adjusted life years (QALYs) after transitioning to a terminal health state s_{T+1}, a commonly used metric to quantify the quality and quantity of life a patient lives. Rewards are discounted annually by $\gamma \in (0,1)$. Given an initial state s, the decision-maker aims to design a treatment policy $\pi := (\pi_t(s_t) : t \in \mathscr{T} \setminus \{T\}, s_t \in \mathscr{S})$ that maximizes the expected total discounted QALYs over the horizon:

$$J^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=0}^{T-1} \gamma^t r_t(s_{t+1}, \pi_t(s_t), \omega_t) + \gamma^T r_T(s_{T+1}, \omega_T) \middle| s \right],$$

where $\mathbb{E}^\pi$ denotes the expectation following policy π with respect to the joint distribution of $\omega_1, \ldots, \omega_T$. The optimal policy is given by $\pi^* = \text{argmax}_{\pi \in \Pi} J^\pi(s)$, where Π denotes the set of all admissible policies. Splitting the problem into decision epochs, we obtain the set of dynamic programming equations:

$$Q_t(s_t, a_t) := \mathbb{E}^\pi \left[r_t(s_t, a_t, \omega_t) + \gamma v_{t+1}\left(f_{t+1}(s_t, a_t, \omega_t)\right) | s_t, a_t \right],$$

where $a_t = \pi_t(s_t)$, $v_t(s_t) := \max_{a_t \in \mathscr{A}} Q_t(s_t, a_t)$ and $v_T(s_T) := \mathbb{E}[r_T(s_t, \omega_t)|s_t]$. Starting from the terminal period T and proceeding backward until the initial year 0, we can find an optimal decision rule $\pi_t^*(s_t) \in \text{argmax}_{a_t \in \mathscr{A}} Q_t(s_t, a_t)$ at each year t to identify an optimal policy $\pi^* = (\pi_t^*(s_t) : t \in \mathscr{T} \setminus \{T\}, s_t \in \mathscr{S})$.

2.1 Monotone Policies

In this paper, we are interested in learning treatment plans that maximize the expected total discounted rewards over the planning horizon while leveraging the inherent interpretability of monotonicity. We assume $\mathscr{S}$ and $\mathscr{A}$ are ordered and focus on the set of all admissible monotone policies $\Pi^M \subset \Pi$ defined as:

Definition 1 Under ordered states and actions, a monotone policy is a function $\pi : \mathscr{S} \to \mathscr{A}$ such that $\pi(s_t) \leq \pi(s_t')$ for all $s_t, s_t' \in \mathscr{S}$ when $s_t \leq s_t'$.

Due to their ease of interpretation and implementation, monotone policies are typically appealing to practitioners. For example, clinicians may find treatment policies more interpretable if they follow a natural order, like increasing treatment intensity as patients' health worsens. Similar to Garcia et al. (2024), our goal is to determine the policy $\pi^M \in \Pi^M$ which achieves the greatest expected total discounted reward without requiring any conditions on the MDP data. Formally, our optimization problem is given by:

$$\pi^M = \underset{\pi \in \Pi^M}{\text{argmax}} J^\pi(s).$$

However, we do not presume perfect knowledge of the state transitions or rewards. We assume that these two components in MDP models can be estimated through simulation instead.

3 MODEL-BASED Q-LEARNING FOR MONOTONE POLICIES

We now present our approach to learning monotone policies based on simulated health trajectories in MDP models with ordered states and actions, referred to as monotone Q-learning. Our goal is to learn an approximately optimal monotone policy $\hat{\pi}^M$ from action-value function estimates $\hat{Q}_t(s_t,a_t)$ that are updated through the exploration of actions taken according to a behavior policy b. Although we present our algorithm within the context of hypertension treatment planning, the ideas extend beyond this application and outside model-based learning. Our procedure is included as Algorithm 1.

Algorithm 1: Monotone Q-Learning.

Input : Let b be an ε-greedy policy with $\varepsilon = 1$.
Initialize $N \in \mathbb{N}$, $\mathscr{A}_t^-(s_t) = A$, $\mathscr{A}_t^+(s_t) = 1$, and $\hat{Q}_t(s_t,a_t) = 0$ for all $s_t \in \mathscr{S} \setminus S$ and $a_t \in \mathscr{A}$.

for $n \leftarrow 1$ **to** N **do**
  Initialize s_0.
  for $t = 0,1,\ldots$ until s_t is terminal **do**
    if $s_t = 1$ **then**
      Set $\mathscr{A}_t(s_t) \leftarrow \{a_t \in \mathscr{A} \ : 1 \le a_t \le \mathscr{A}_t^-(s_t+1)\}$
    else if $s_t = S$ **then**
      Set $\mathscr{A}_t(s_t) \leftarrow \{a_t \in \mathscr{A} \ : \mathscr{A}_t^+(s_t-1) \le a_t \le A\}$
    else
      Set $\mathscr{A}_t(s_t) \leftarrow \{a_t \in \mathscr{A} \ : \mathscr{A}_t^+(s_t-1) \le a_t \le \mathscr{A}_t^-(s_t+1)\}$.
    end if
    Set $greedy \leftarrow \text{argmax}_{a_t \in \mathscr{A}(s_t)} \hat{Q}_t(s_t,a_t)$.
    Set $\varepsilon \leftarrow \frac{1}{n+1}$ and let $b_t(a_t|s_t) \leftarrow \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|\mathscr{A}_t(s_t)|} & \text{if } a_t = greedy \\ \frac{\varepsilon}{|\mathscr{A}_t(s_t)|} & \text{if } a_t \neq greedy,\ a_t \in \mathscr{A}_t(s_t) \\ 0 & \text{otherwise} \end{cases}$
    Choose $a_t \sim b_t(s_t)$, generate $s_{t+1} \leftarrow f_{t+1}(s_t,a_t,\omega_t)$, and observe $r_t(s_{t+1},a_t,\omega_t)$.
    Update $\hat{Q}_t(s_t,a_t)$ using equation (1) and set $s_t \leftarrow s_{t+1}$.
  end for
end for
Obtain $\{x_t(s_t,a_t) : s_t \in \mathscr{S}, a_t \in \mathscr{A}\}$ from formulation (2) with $\{\hat{Q}_t(s_t,a_t) : s_t \in \mathscr{S}, a_t \in \mathscr{A}\}$.
Set $\hat{\pi}_t^M(s_t) \leftarrow \text{argmax}_{a_t} x_t(s_t,a_t)$ for all $s_t \in \mathscr{S} \setminus S$.

Output: $\{\hat{Q}(s,a) : s \in \mathscr{S}, a \in \mathscr{A}\}, \hat{\pi}^M$.

Before the algorithm starts, we initialize the action value function estimates $\hat{Q}_t(s_t,a_t)$ for all t, s_t, and a_t. We then select a patient's initial state s_0 according to their characteristics prior to the decision-making process in every simulated health trajectory $n = 1,...,N$.

As no action has been taken before the first health trajectory $n = 1$, we initialize the set of permissible treatment choices $\mathscr{A}_t(s_t)$ to the complete action space by setting $\mathscr{A}_t^-(s_t) = A$ and $\mathscr{A}_t^+(s_t) = 1$. For each health trajectory $n > 1$ and state $1 < s_t < S$ at year $t = 0,1,...,T-1$, we update the set of permissible treatment choices as:

$$\mathscr{A}_t(s_t) \leftarrow \{a_t \in \mathscr{A} \ : \mathscr{A}_t^+(s_t-1) \le a_t \le \mathscr{A}_t^-(s_t+1)\},$$

where $\mathscr{A}_t^-(s_t+1) \coloneqq \min\{\bar{a}_t(s_t+1),\ldots,\bar{a}_t(S),A\}$, $\mathscr{A}_t^+(s_t-1) \coloneqq \max\{1,\bar{a}_t(1),\ldots,\bar{a}_t(s_t-1)\}$, and $\bar{a}_t(s_t)$ denotes the last action taken at state s_t, provided it has been observed. When $s_t = 1$ or $s_t = S$, we replace $\mathscr{A}_t^+(s_t-1)$ by 1 and $\mathscr{A}_t^-(s_t+1)$ by A, respectively. Through these sets, we restrict the exploration of the algorithm to actions that are monotone in the state order. Subsequently, we identify an action that maximizes the current estimate of the action-value function $\hat{Q}_t(s_t,a_t)$ at state s_t (i.e., a greedy action) and update the behavior policy b_t to only cover actions in $\mathscr{A}_t(s_t)$. We then select action $a_t \sim b_t(s_t)$ according to behavior policy b_t at state s_t and simulate a realization of $\omega_t \sim \mathscr{U}(0,1)$. Next, we use a_t and ω_t to generate the next state $s_{t+1} = f_{t+1}(s_t,a_t,\omega_t)$ and reward $r_t(s_{t+1},a_t,\omega_t)$. Based on this information, we update our estimate of the action value function in state s_t as:

$$\hat{Q}_t(s_t,a_t) \leftarrow \hat{Q}_t(s_t,a_t) + \alpha_n \left(\left[r_t(s_{t+1},a_t,\omega_t) + \gamma \max_{a_{t+1}\in\mathscr{A}} \hat{Q}(s_{t+1},a_{t+1}) \right] - \hat{Q}_t(s_t,a_t) \right) \tag{1}$$

where $\alpha_n \in (0,1)$ is the learning rate of the algorithm at health trajectory n. We then let $s_{t+1} \leftarrow s_t$ and the same procedure is repeated at time $t+1$ until $T-1$. Once the patient reaches this stage, their next state action-value function is replaced by the terminal reward $r_T(s_{T+1},\omega_T)$ and we proceed to the next health trajectory simulation $n+1$.

To ensure the policy $\hat{\pi}_t^M(s_t)$ obtained from the current action-value function estimates is monotone, we solve the following binary integer program:

$$\max_{\mathbf{x}} \quad \sum_{t\in\mathscr{T}\setminus\{T\}} \sum_{s_t\in\mathscr{S}} \sum_{a_t\in\mathscr{A}} \hat{Q}_t(s_t,a_t) x_t(s_t,a_t) \tag{2a}$$

$$\text{s.t.} \quad x_t(s_t,a_t) \le \sum_{a_t'\ge a_t} x_t(s_t+1,a_t') \text{ for all } t\in\mathscr{T}\setminus\{T\}, s_t\in\mathscr{S}\setminus S, a_t\in\mathscr{A}, \tag{2b}$$

$$x_t(s_t,a_t) \in \{0,1\} \text{ for all } t\in\mathscr{T}\setminus\{T\}, s_t\in\mathscr{S}, a_t\in\mathscr{A}. \tag{2c}$$

After formulation (2) is solved, a monotone policy may be obtained as $\hat{\pi}_t^M(s_t) = \text{argmax}_{a_t} x_t(s_t,a_t)$ for all $s_t \in \mathscr{S}\setminus S$.

3.1 Justification for the Monotone Q-learning Algorithm

This section presents our result on the convergence of the monotone Q-learning algorithm, assuming the standard Robbins–Monro conditions (Powell 2011) on the learning rate are satisfied. The proof of our claim can be found in Appendix A.

Theorem 1 The monotone policy obtained from the monotone Q-learning algorithm converges to an optimal monotone policy as $N \to \infty$.

4 PERSONALIZED HYPERTENSION TREATMENT NUMERICAL ANALYSIS

We now present our simulation model to evaluate approximately optimal monotone hypertension treatment plans developed with the proposed monotone Q-learning algorithm. For comparison purposes, we also evaluate the optimal monotone and optimal treatment plans as described in Garcia et al. (2024) and the 2017 Hypertension Clinical Practice Guidelines (Whelton et al. 2018). These clinical guidelines suggest pharmacological treatment for patients with stage 1 hypertension (i.e., systolic BP of 130-139 mm Hg or diastolic BP of 80-89 mm Hg) if their 10-year risk for ASCVD exceeds 10%. For patients with stage 2 hypertension (i.e., systolic BP of at least 140 mm Hg or a diastolic BP of at least 90 mm Hg), the guidelines recommend treatment until they reach controlled BP levels below stage 1 hypertension.

The trajectory of a single patient in our simulation framework is summarized in Figure 1. We first calculate the risk for ASCVD events each year (Yadlowsky et al. 2018). Subsequently, we estimate transition probabilities and transition functions. We then execute $N = 100,000$ episodes (i.e., patients' health trajectories) of the monotone Q-learning method and use formulation (2) to find an approximately

optimal monotone policy. Lastly, we determine the optimal and optimal monotone treatment strategies along with the clinical guidelines for comparison purposes.

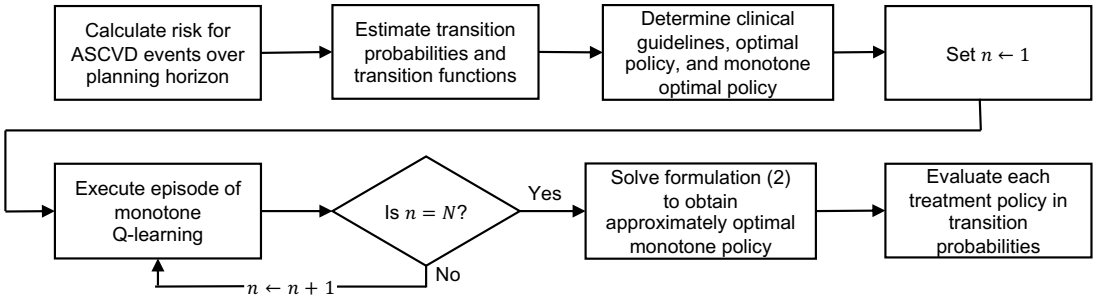

Figure 1: Summary of simulation framework for a single patient. The index n represents the episode in monotone Q-learning.

We now describe our patient profiles, MDP formulation, and patients' outcomes following the recommendations obtained with each treatment strategy.

4.1 Patient Profiles

Based on conversations with clinical collaborators, we identify a set of patient profiles that are representative of a population with a high prevalence of ASCVD that can benefit considerably from hypertension treatment. We first consider a patient profile with no major clinical risk factors for ASCVD. This base patient profile has the following characteristics: 45-year-old, non-diabetic, non-smoker, normal BP, and normal cholesterol levels. We then modify the BP levels of the patient profile to elevated BP, stage 1, and stage 2 hypertension, as defined by the 2017 Hypertension Clinical Practice Guidelines (Whelton et al. 2018).

We use data from the National Health and Nutrition Examination Survey (NHANES) to parameterize the health evolution of our patient profiles. Our population is composed of adult Caucasian or African-American patients from 40 to 60 years old with no history of ASCVD. Assuming that smoking and diabetes status remain constant, we linearly regress systolic BP, diastolic BP, high-density lipoprotein, and total cholesterol on age, age squared, sex, race, smoking status, and diabetes status. This regression model allows us to estimate the progression of patients' risk factors over the planning horizon. We then use these estimates as inputs into the revised pooled cohort equations to calculate each patient's ASCVD risk (Yadlowsky et al. 2018), which is adjusted if the patient experiences an adverse event. Death from non-ASCVD causes is modeled as an independent process and not considered in the risk factor progression.

4.2 Markov Decision Process Formulation

In this study, we adopt the MDP simulation model in Marrero et al. (2021). However, the objective of the MDP simulation model in our study is to determine the treatment strategy that maximizes the expected total discounted QALYs instead of the life years before an adverse event. The adjusted model has the following elements:

- $\mathscr{T}$: 10-year planning horizon with decisions made at the beginning of each year $t \in \mathscr{T} \setminus \{10\}$. We use $T = 10$ to represent the effects of treatment on patients' lifetime. This planning horizon is selected based on the major high BP management guidelines (Whelton et al. 2018).
- $\mathscr{S}$: state space composed of patients' age, sex, race, smoking status (i.e., demographic information), BP, diabetes status, cholesterol readings (i.e., clinical observations), and health condition h_t to account for their history of cardiovascular events. We categorize patients' health condition into the following mutually-exclusive groups: healthy ($h_t = 1$), history of heart attack but no adverse event in the

current year ($h_t = 2$), history of stroke but no adverse event in the current year ($h_t = 3$), history of heart attack and stroke but no adverse event in the current year ($h_t = 4$), survival of a heart attack ($h_t = 5$), survival of a stroke ($h_t = 6$), death from a non-ASCVD related cause ($h_t = 7$), death from a heart attack ($h_t = 8$), death from stroke ($h_t = 9$), and dead ($h_t = 10$). We use s_t to denote a patient's state or $s_t(h_t)$ when a specific health condition must be emphasized.

- $\mathscr{A}$: action space comprising from 0 to 5 antihypertensive medications at a half and standard dosage, totaling $A = 21$ treatment choices. In contrast to Garcia et al. (2024), this paper focuses only on the number of medications since research suggests that the benefit from treatment is determined by the BP reduction achieved, with little effect attributable to specific drugs (Sundström et al. 2014).
- $f_{t+1}(s_t, a_t, \omega_t)$: transition function derived from a patient's risk for ASCVD events (Yadlowsky et al. 2018), benefit from treatment (Sundström et al. 2014; Sussman et al. 2013), ASCVD event mortality (NCHS 2017), and non-ASCVD mortality (Arias and Xu 2019). Similar to previous studies (Garcia et al. 2024; Marrero et al. 2021), we assume independence among heart attacks and strokes. Furthermore, heart attacks account for 70% of the ASCVD risk and stroke events account for the remaining 30%. In addition, we assume that patients are more likely to have additional heart attacks or strokes if they have a history of such ASCVD events. We incorporate this assumption by adjusting patients' heart attack and stroke odds if they have a history of either ASCVD event (Brønnnum-Hansen et al. 2001; Burn et al. 1994).
- $r_t(s_{t+1}, a_t, \omega_t)$: reward given by the quality of life weight associated with patients' health condition h_t (Kohli-Lynch et al. 2019) minus the disutility from medication a_t (Sussman et al. 2013).
- $r_T(s_{T+1}, \omega_T)$: terminal reward representing each patient's total QALYs after transitioning to state s_{T+1} computed as the product of their expected lifetime (Arias and Xu 2019), a mortality factor that accounts for the effect of ASCVD events on future mortality (Pandya et al. 2015), and a terminal quality of life weight (Kohli-Lynch et al. 2019).
- γ: 3% discount on future quality-adjusted life-year gains as recommended in the medical literature (Neumann et al. 2016); $\gamma = 0.97$.

Table 1 lists the parameters used in our analyses. Please refer to Appendix B in Marrero et al. (2021) for a description of the calibration and validation of our MDP simulation model.

4.2.1 State and Action Ordering

Our state ordering is based on patients' associated risk for ASCVD events. At each decision epoch t, each patient's state has the same demographic information and estimated clinical observations. Differences in the risk for ASCVD events between each patient's states are driven by their health condition h_t. Consequently, each patient's state ordering is determined by the severity of their health condition h_t. Excluding health conditions related to death, this observation leads to the following order of states for each patient at a fixed decision epoch: $s_t(h_t = 1) < s_t(h_t = 2) < s_t(h_t = 5) < s_t(h_t = 3) < s_t(h_t = 6) < s_t(h_t = 4)$.

We order actions in terms of their associated number of medications. This order is equivalent to sorting medications according to their expected systolic BP or risk reductions (Sundström et al. 2014).

4.3 Numerical Results

In this subsection, we evaluate and offer insights into the implications of approximately optimal monotone treatment. We frame our results in terms of the price of interpretability of monotone policies. This quantity is defined by Garcia et al. (2024) as the difference between the expected total discounted QALYs between an optimal and a monotone policy:

$$\mathrm{PI}(\Pi^M) := J^{\pi^*}(s) - \max_{\pi \in \Pi^M} J^{\pi}(s).$$

Table 1: Base case parameters.

Parameter	Value	Source
BP reduction: standard dose (half dose)		
Systolic BP	5.5 (3.7) mm Hg	(Sundström et al. 2014; Sussman et al. 2013)
Diastolic BP	3.3 (2.2) mm Hg	(Sundström et al. 2014; Sussman et al. 2013)
Risk for ASCVD events	Varies by patient	(Yadlowsky et al. 2018)
ASCVD risk reduction: standard dose (half dose)		
Heart attack	13% (7%)	(Sundström et al. 2014; Sussman et al. 2013)
Stroke	21% (14%)	(Sundström et al. 2014; Sussman et al. 2013)
ASCVD risk due to heart attack	70%	(Martin et al. 2024)
Mortality from ASCVD events		
Heart attack	Varies by patient	(NCHS 2017)
Stroke	Varies by patient	(NCHS 2017)
Scaling factor to account for history of ASCVD events		
Heart attack ($h_t = 2,5$)	3	(Brønnnum-Hansen et al. 2001)
Stroke ($h_t = 3,6$)	2,3	(Burn et al. 1994)
Quality of life weight		(Kohli-Lynch et al. 2019)
Healthy ($h_t = 1$)	1	
ASCVD events ($h_t = 2,\ldots,6$)	Varies by patient	
Dead ($h_t = 7,\ldots,10$)	0	
Treatment-related disutility		
Half dose	0.001	(Marrero et al. 2021; Sussman et al. 2013)
Full dose	0.002	(Marrero et al. 2021; Sussman et al. 2013)
Life expectancy	Varies by patient	(Arias and Xu 2019)
Non-ASCVD mortality	Varies by patient	(Arias and Xu 2019)

4.3.1 Insights from Interpretable Treatment

To understand the implications of approximately optimal interpretable treatment plans, we examine the effect of patients' characteristics on the policies obtained with monotone Q-learning. For comparison purposes, we also determine the optimal monotone and optimal policies along with the recommendations from the clinical guidelines as described in Garcia et al. (2024).

Figure 2 illustrates each of the policies over the health conditions of the patient profiles with stage 1 and stage 2 hypertension in the last year of our study. The policies are less aggressive in earlier years because of our monotonicity restrictions on the actions over time. All policies recommend no treatment (not shown) in the base patient profile. If the profile's BP is increased to elevated levels or stage 1 hypertension, the policies prescribe either two medications at half dose or one medication at half dose and another at standard dose in the healthy state. However, the optimal policy decreases the intensity to no treatment if the patient's profile experiences a stroke. While optimal in terms of QALYs gained, this policy is not intuitive for physicians or their patients. Contrastingly, the interpretable policies maintain treatment intensity if the patient were to experience more severe states. There is no considerable consequence for providing interpretability in these patient profiles as the difference between their price of interpretability is at most 0.001 QALYs. When the patient profile's BP rises to stage 2 hypertension, the optimal monotone and optimal policies recommend three medications at half dose while the approximately optimal monotone policy prescribes two medications at half dose. Similarly to the profiles with elevated BP or stage 1 hypertension, the optimal policy decreases intensity while the optimal monotone maintains aggressiveness. Nevertheless, the approximately optimal monotone policy increases intensity to two medications at standard dose as the patient's health worsens. We note that three medications at half dose provide the same benefit as two medications at standard dose. The differences among the policies' price of interpretability reach a maximum of 0.007 QALYs between the optimal policy and the approximately optimal monotone policy in stage 2 hypertension.

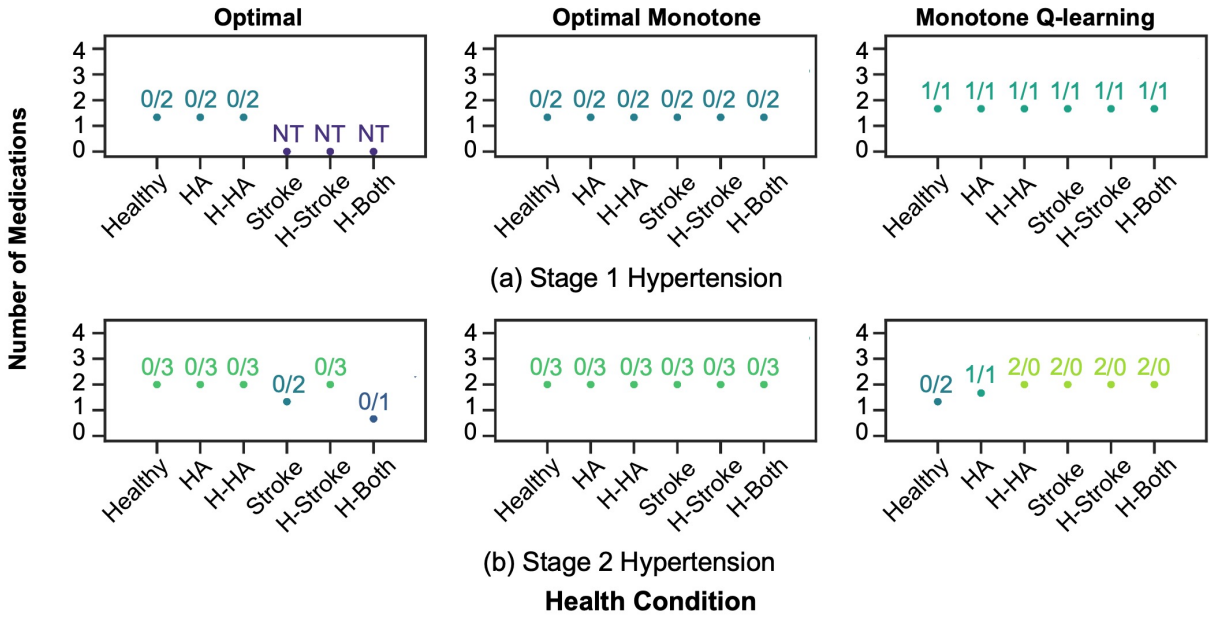

Figure 2: Treatment policies over the health conditions of profiles with (a) stage 1 and (b) stage 2 hypertension. H-: History of; HA: heart attack; NT: No treatment; Number before and after forward slash represent medications at standard and half dose, respectively.

We exclude the clinical guidelines from Figure 2 to facilitate the comparison between the approximately optimal monotone policy with the optimal and optimal monotone policies. The clinical guidelines prescribe no treatment for normal and elevated BP levels. The profile with stage 1 hypertension is recommended one medication at standard dose across all the states. In stage 2 hypertension, the guidelines recommend one medication at standard dose for the healthy state and two medications at standard dose at the ASCVD-related states. Following the clinical guidelines may have implications on patients' health as the price of interpretability paid for each patient profile is higher than in the interpretable treatment strategies.

Our patient-level results lead to three critical observations. First, similar to the findings in Garcia et al. (2024), the optimal policy typically suggests less treatment as patients' health severity increases. This behavior does not reflect physicians' intuition in practice. A potential reason for this behavior is that the policy aims to maximize the expected discounted QALYs and not to minimize the total number of ASCVD events. Second, we also find the optimal monotone policy typically prescribes a constant treatment across all health conditions in each year of the planning horizon. Lastly, the policies obtained with monotone Q-learning tend to recommend less aggressive treatment than the optimal monotone policies in healthier states and coincide with the optimal monotone policies in more severe states. While the three strategies may lead to different treatment recommendations, they all result in prices of interpretability within 0.007 QALYs which may not be practically meaningful. The patterns displayed by the approximately optimal monotone and optimal monotone policies align with the intuition of clinicians in practice. They both provide more intuitive strategies than the optimal policy with only a small loss in QALYs.

5 CONCLUSIONS

In this paper, we introduced a new approach to obtain approximately optimal monotone policies in MDP simulation models. We presented a modification of the standard Q-learning algorithm, which we called monotone Q-learning. Our algorithm restricts the exploration in standard Q-learning acknowledging that actions must be monotone on an order of states specified by a decision maker. Furthermore, it guarantees

that the approximately optimal policies obtained are monotone by solving a simple binary integer program. Although presented when a simulator is available, our approach holds for model-free situations where a transition function is not accessible.

Our numerical analysis studied the implications of approximately optimal monotone antihypertensive treatment plans. Two principal conclusions can be made from our analysis. First, the optimal monotone policy typically prescribes a constant treatment across all health conditions in each year of the planning horizon, while the approximately optimal monotone policies tend to recommend slightly less intense treatment in healthier states and become marginally more aggressive as patients' health worsen. Second, the performance losses of the optimal monotone and approximately optimal monotone policies with respect to the optimal policy are comparable. The losses in QALYs associated with the interpretability of these policies are likely practically negligible.

Our work on approximately optimal monotone antihypertensive treatment plans can be extended in multiple ways. From a technical point of view, the ideas from our algorithm can be adapted to ensure monotone policies in other temporal-difference reinforcement learning methods. Our algorithm inherits some of the limitations of the standard Q-learning in large state and action spaces. Future work may focus on overcoming these challenges. From a clinical perspective, this research can be expanded by incorporating other conditions, such as high cholesterol or diabetes. Moreover, future work may also consider a population-level analysis or examining the effects of varying model parameters.

The monotone Q-learning algorithm continues a line of work aiming to increase the acceptance of MDP models in practice by providing policies that harness the natural interpretability of monotonicity. Strategies that follow an intuitive order may offer decision-makers support that is amenable to their cognition. We showed that approximately optimal monotone policies are capable of leading to minor performance losses when compared to optimal policies, even in complex settings like personalized treatment planning. Interpretable policies have great potential to enable the implementation of MDP-guided recommendations into practice within and beyond healthcare applications.

A PROOF OF THEOREM 1

The proof of Theorem 1 depends on the following lemma:

Lemma 1 Every action $a_t \in \mathscr{A}$ is included infinitely often in $\mathscr{A}_t(s_t) \subseteq \mathscr{A}$ as $N \to \infty$.

Proof. Rewrite $\mathscr{A}_t(s_t) = \mathscr{A} \setminus \big([1, \mathscr{A}_t^+(s_t-1)) \cup (\mathscr{A}_t^-(s_t+1), A]\big)\big)$. We first show that $\mathscr{A}_t^+(s_t-1)$ will change with positive probability for an arbitrary state s_t visited at time t and episode n. When state s_t is visited again at time t in some episode $n' > n$, $\mathscr{A}_t^+(s_t-1)$ has shifted with probability $\sum_{s'=1}^{s_t-1} p_t\big(s'|s,\bar{a}(s)\big) \sum_{a'=\mathscr{A}_t^+(s'-1)}^{\bar{a}(s_t)-1} b_t(a'|s') > 0$ for each state $s \in \mathscr{S}$ visited during episodes between n and n' as long as $p_t\big(s'|s,\bar{a}(s)\big) > 0$ for some $s' = 1,...,s_t-1$ and $\bar{a}(s') \neq \mathscr{A}_t^+(s_t-1)$ for all $s' = 1,...,s_t-1$. In here, $p_t\big(s'|s,\bar{a}(s)\big) \coloneqq \mathbb{E}[\mathbb{1}\{f_{t+1}(s,\bar{a}(s),\omega_t) = s'\}|s,\bar{a}(s)]$ and $\mathbb{1}\{\cdot\}$ represents an indicator function.

We now show that $|\mathscr{A}_t(s_t)| > 1$ for some episode $m > n'$ if $\bar{a}(s') = \mathscr{A}_t^+(s_t-1)$ for all $s' = 1,...,s_t-1$ between n and n'. Let $\mathscr{S}^R \coloneqq \{s \in \mathscr{S} : |\mathscr{A}_t(s)| = 1\}$ denote the set of states with only one permissible action (i.e., restricted states), $\mathscr{S}^U \coloneqq \mathscr{S} \setminus \mathscr{S}^R$ the set of unrestricted states, and $\mathscr{S}^B \coloneqq \{s \in \mathscr{S}^R : (s-1 \in \mathscr{S}^U) \vee (s+1 \in \mathscr{S}^U)\}$ the set of boundary states. Under an ε-greedy behavior policy, any state $s \in \mathscr{S}^B$ can become unrestricted as an action $\hat{a} \neq \bar{a}(s)$ will be taken with positive probability when state $s-1 \in \mathscr{S}^U$ or $s+1 \in \mathscr{S}^U$ is visited. Moreover, any state $s \in \mathscr{S}^R$ can become a boundary state. If $s-1 \in \mathscr{S}^B$ or $s+1 \in \mathscr{S}^B$, the behavior policy guarantees the boundary state will become unrestricted with positive probability, which implies s will then become a boundary state. If $s \in \mathscr{S}^R$ has multiple restricted predecessors and successors, s will inductively become a boundary state and eventually an unrestricted state.

As a result, $\mathscr{A}_t^-(s'+1)$ for $s' \in \{1,...,s_t-1\}$ is expanded with at least one action which implies that $\mathscr{A}_t^+(s_t-1)$ will shift with positive probability in a later episode. Symmetrically, the probability that $\mathscr{A}_t^-(s_t+1)$ shifts is positive for every s_t. This result suggests that there exist episodes when $\mathscr{A}_t^+(s_t-1) = 1$

or $\mathscr{A}_t^-(s_t+1)=A$. Since $\mathscr{A}_t(s_t)$ is a consecutive sequence of actions from 1 to A, this observation implies that every action is included in each subset $\mathscr{A}_t(s_t)$ infinitely often as $N\to\infty$. □

Proof of Theorem 1. Lemma 1 implies that every state-action pair will be visited infinitely often. Thus, the monotone Q-learning algorithm finds decision rules in the same spaces as the traditional Q-learning algorithm. By Watkins and Dayan (1992), it follows that $\hat{Q}_t(s_t,a_t)\to Q_t(s_t,a_t)$ with probability 1 as $N\to\infty$.

Formulation (2) guarantees the algorithm searches within the same space as formulation (5) in Garcia et al. (2024). Thus, our monotone Q-learning algorithm outputs an approximately optimal monotone policy $\hat{\pi}^M$ that converges to an optimal monotone policy π^M as $N\to\infty$. Moreover, $\hat{\pi}^M\to\pi^M$ when the monotone optimal policy is unique. □

REFERENCES

Arias, E. and J. Xu. 2019. "United States Life Tables, 2017". *National Vital Statistics Reports* 68(7).

Bertsimas, D., P. Klasnja, S. Murphy, and L. Na. 2022. "Data-driven Interpretable Policy Construction for Personalized Mobile Health". In *2022 IEEE International Conference on Digital Health (ICDH)*, 13–22. Barcelona, Spain: IEEE https://doi.org/10.1109/ICDH55609.2022.00010.

Bonifonte, A., T. Ayer, and B. Haaland. 2022. "An analytics approach to guide randomized controlled trials in hypertension management". *Management Science* 68(9):6634–6647 https://doi.org/10.1287/mnsc.2021.4226.

Brønnnum-Hansen, H., T. Jørgensen, M. Davidsen, M. Madsen, M. Osler, L. U. Gerdes *et al*. 2001. "Survival and cause of death after myocardial infarction: the Danish MONICA study". *Journal of Clinical Epidemiology* 54(12):1244–1250 https://doi.org/10.1016/S0895-4356(01)00405-X.

Burn, J., M. Dennis, J. Bamford, P. Sandercock, D. Wade and C. Warlow. 1994. "Long-term risk of recurrent stroke after a first-ever stroke. The Oxfordshire Community Stroke Project". *Stroke* 25(2):333–7 https://doi.org/http://dx.doi.org/10.1161/01.STR.25.2.333.

Cohen, J. B. and R. R. Townsend. 2018. "The ACC/AHA 2017 Hypertension Guidelines: Both Too Much and Not Enough of a Good Thing?". *Annals of Internal Medicine* 168(4):287 https://doi.org/10.7326/M17-3103.

Djonin, D. V. and V. Krishnamurthy. 2007. "Q-Learning Algorithms for Constrained Markov Decision Processes With Randomized Monotone Policies: Application to MIMO Transmission Control". *IEEE Transactions on Signal Processing* 55(5):2170–2181 https://doi.org/10.1109/TSP.2007.893228.

Fu, F. and M. Van Der Schaar. 2012. "Structure-Aware Stochastic Control for Transmission Scheduling". *IEEE Transactions on Vehicular Technology* 61(9):3931–3945 https://doi.org/10.1109/TVT.2012.2213850.

Garcia, G.-G. P., L. N. Steimle, W. J. Marrero, and J. B. Sussman. 2024. "Interpretable Policies and the Price of Interpretability in Hypertension Treatment Planning". 26(1):80–94 https://doi.org/10.1287/msom.2021.0373.

Glanois, C., P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao *et al*. 2024. "A survey on interpretable reinforcement learning". *Machine Learning*:1–44 https://doi.org/https://doi.org/10.1007/s10994-024-06543-w. Publisher: Springer.

Hu, X., P.-Y. S. Hsueh, C.-H. Chen, K. M. Diaz, F. E. Parsons, I. Ensari, *et al*. 2018. "An interpretable health behavioral intervention policy for mobile device users". *IBM Journal of Research and Development* 62(1):4:1–4:6 https://doi.org/10.1147/JRD.2017.2769320.

Ioannidis, J. P. 2018. "Diagnosis and treatment of hypertension in the 2017 ACC/AHA guidelines and in the real world". *JAMA - Journal of the American Medical Association* 319(2):115–116 https://doi.org/10.1001/jama.2017.19672.

Kochanek, K. D., S. L. Murphy, J. Xu, and E. Arias. 2024. *Mortality in the United States, 2022*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Kohli-Lynch, C. N., B. K. Bellows, G. Thanassoulis, Y. Zhang, M. J. Pletcher, E. Vittinghoff *et al*. 2019. "Cost-effectiveness of Low-density Lipoprotein Cholesterol Level–Guided Statin Treatment in Patients With Borderline Cardiovascular Risk". *JAMA Cardiology* 4(10):969–977 https://doi.org/10.1001/jamacardio.2019.2851.

Lakkaraju, H. and C. Rudin. 2017. "Learning cost-effective and interpretable treatment regimes". *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017* 54.

Liu, L. and U. Mitra. 2020. "On Sampled Reinforcement Learning in Wireless Networks: Exploitation of Policy Structures". *IEEE Transactions on Communications* 68(5):2823–2837 https://doi.org/10.1109/TCOMM.2020.2974216.

Marrero, W., M. S. Lavieri, A. Tewari, J. Sussman and R. A. Hayward. 2021. "Data-Driven Ranges of Near-Optimal Actions for Finite Markov Decision Processes". *Optimization Online*.

Marrero, W. J., M. S. Lavieri, and J. B. Sussman. 2021. "Optimal cholesterol treatment plans and genetic testing strategies for cardiovascular diseases". *Health Care Management Science* https://doi.org/10.1007/s10729-020-09537-x.

Martin, S. S., A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery *et al*. 2024. "2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association". *Circulation* 149(8):e347–e913.

Murphy, S. A. 2005. "A Generalization Error for Q-Learning". *Journal of Machine Learning Research* 6:1073–1097.

NCHS 2017. "Health, United States, 2016: with chartbook on long-term trends in health". *Center for Disease Control*:314–317.

Neumann, P., G. Sanders, L. Russell, and J. Siegel. 2016. *Cost-effectiveness in health and medicine*. Oxford University Press.

Ngo, M. H. and V. Krishnamurthy. 2010. "Monotonicity of Constrained Optimal Transmission Policies in Correlated Fading Channels With ARQ". *IEEE Transactions on Signal Processing* 58(1):438–451 https://doi.org/10.1109/TSP.2009.2027735.

Pandya, A., S. Sy, S. Cho, M. C. Weinstein and T. A. Gaziano. 2015. "Cost-Effectiveness of 10-Year Risk Thresholds for Initiation of Statin Therapy for Primary Prevention of Cardiovascular Disease". *Journal of the American Medical Association* 314(2):142–150 https://doi.org/10.1001/jama.2015.6822.

Powell, W. B. 2011. *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Roy, A., V. Borkar, P. Chaporkar, and A. Karandikar. 2020. "Low Complexity Online Radio Access Technology Selection Algorithm in LTE-WiFi HetNet". *IEEE Transactions on Mobile Computing* 19(2):376–389 https://doi.org/10.1109/TMC.2019.2892983.

Roy, A., V. Borkar, A. Karandikar, and P. Chaporkar. 2019. "A Structure-aware Online Learning Algorithm for Markov Decision Processes". In *Proceedings of the 12th EAI International Conference on Performance Evaluation Methodologies and Tools*, 71–78. Palma Spain: ACM https://doi.org/10.1145/3306309.3306321.

Roy, A., V. Borkar, A. Karandikar, and P. Chaporkar. 2022. "Online Reinforcement Learning of Optimal Threshold Policies for Markov Decision Processes". *IEEE Transactions on Automatic Control* 67(7):3722–3729 https://doi.org/10.1109/TAC.2021.3108121.

Sethi, T., A. Kalia, A. Sharma, and A. Nagori. 2020. *Interpretable artificial intelligence: Closing the adoption gap in healthcare*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-817133-2.00001-x.

Shakya, A. K., G. Pillai, and S. Chakrabarty. 2023. "Reinforcement learning algorithms: A brief survey". *Expert Systems with Applications* 231:120495 https://doi.org/10.1016/j.eswa.2023.120495.

Solberg, L. I. and W. L. Miller. 2018. "The new hypertension guideline: logical but unwise". *Family Practice* 35(5):528–530 https://doi.org/10.1093/fampra/cmy026.

Steimle, L. N., D. L. Kaufman, and B. T. Denton. 2021. "Multi-model Markov decision processes". *IISE Transactions* 53(10):1124–1139 https://doi.org/https://doi.org/10.1080/24725854.2021.1895454.

Sundström, J., H. Arima, M. Woodward, R. Jackson, K. Karmali, D. Lloyd-Jones, *et al*. 2014. "Blood pressure-lowering treatment based on cardiovascular risk: A meta-analysis of individual patient data". *The Lancet* 384(9943):591–598 https://doi.org/10.1016/S0140-6736(14)61212-5.

Sussman, J., S. Vijan, and R. Hayward. 2013. "Using benefit-based tailored treatment to improve the use of antihypertensive medications". *Circulation* 128(21):2309–2317.

Wang, F., R. Kaushal, and D. Khullar. 2020. "Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine?". *Annals of Internal Medicine* 172:59 https://doi.org/10.7326/M19-2548.

Watkins, C. J. C. H. and P. Dayan. 1992. "Q-learning". *Machine Learning* 8(3):279–292 https://doi.org/10.1007/BF00992698.

Whelton, P. K., R. M. Carey, W. S. Aronow, D. E. Casey, K. J. Collins, C. Dennison Himmelfarb, *et al*. 2018. "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults". *Journal of the American College of Cardiology* 71(19):e127–e248 https://doi.org/10.1016/j.jacc.2017.11.006.

Yadlowsky, S., R. A. Hayward, J. B. Sussman, R. L. McClelland, Y.-I. Min and S. Basu. 2018. "Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk". *Annals of Internal Medicine* 169(1):20 https://doi.org/10.7326/M17-3011.

AUTHOR BIOGRAPHIES

WESLEY J. MARRERO is an Assistant Professor of Engineering in Thayer School at Dartmouth College. Before joining Dartmouth, he was a postdoctoral research fellow in the Massachusetts General Hospital Institute for Technology Assessment at Harvard Medical School. His research interests lies on overcoming the challenges associated with the implementation of decision-support tools in practice, such as lack of interpretability, inequity, irrational behavior, and need for flexibility. To this end, he designs and applies techniques from operations research and statistics, with an emphasis on simulation and optimization. His current work addresses various application areas, including cardiovascular disease, substance use disorder, mental health, and organ transplantation. His e-mail address is wesley.marrero@dartmouth.edu. His website is https://engineering.dartmouth.edu/community/faculty/wesley-marrero.

LAN YI is a Ph.D. student at Thayer School of Engineering at Dartmouth College. Her current research interest is developing interpretable reinforcement learning algorithms in healthcare. She holds a master's degree of mathematics from Tufts University with a focus on statistical machine learning and convex optimization. Her previous research experience is related to cancer survival analysis in bioinformation and machine learning. Her email address is mina.yi.th@dartmouth.edu.